BEYOND CONFIDENCE: ACTIVE FALSIFICATION FOR EMBODIED QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied question answering (EQA) requires an agent to explore an unfamiliar environment and gather visual evidence before answering. However, question-relevant observations may fail to distinguish competing answers, allowing an early, confident error to persist. We propose Active Falsification for Embodied Question Answering (AFE-EQA), a framework that treats the current answer as a revisable hypothesis and actively seeks observations capable of refuting it. Three-state grounding distinguishes decidable, partial, and not-visible observations to regulate belief updates and evidence retention. A structured counter-evidence memory links testable propositions to their source hypotheses and visual assessments, while falsification-aware imagination guides exploration toward viewpoints that distinguish competing answers. At termination, an offline-trained resolver integrates accumulated visual and trajectory evidence to determine whether the current answer should be revised. These components operate along a single exploration trajectory. Experiments on HM-EQA show improved final answer accuracy under matched conditions, while component ablations and paired analyses characterize the contributions of exploration and terminal revision, including both corrected errors and harmful switches. These findings support active hypothesis testing as a design principle for coupling embodied exploration with evidence-based answer revision.

1 INTRODUCTION

Navigation enables embodied agents to actively acquire visual information for downstream tasks (Bajcsy, 1988). Advances in generalist navigation models have extended navigation capabilities across tasks and embodiments (Zheng et al., 2024; Zhang et al., 2026a). In embodied question answering (EQA), navigation serves to acquire the visual evidence needed to answer a question about an unfamiliar environment (Das et al., 2018a; Majumdar et al., 2024). Unlike visual question answering over a supplied image (Antol et al., 2015), EQA requires the agent to select the observations on which its answer will depend. This interdependence between perception, memory, and action (Zhang et al., 2026b) makes evidence acquisition part of the reasoning process: the agent must determine which observations can resolve uncertainty about the answer.

An observation can be relevant to a question without distinguishing its possible answers. Consider a question about a round blue cushion. A kitchen view may cause a model to confidently choose "on the kitchen chair," even though the cushion's location is unresolved. Further views of the kitchen can repeat this prediction without testing it. A more useful action seeks a view that distinguishes the chair from the living-room floor. In Figure 1, a later observation reveals the cushion on the floor, challenging the initial chair hypothesis. Retaining this visual evidence allows the agent to reconsider its answer even after the cushion leaves view.

Existing approaches provide several ingredients for this process. Explore until Confident combines semantic exploration with calibrated stopping (Ren et al., 2024). Memory-based methods retain observations for later reasoning (Yang et al., 2025), while imagination and predictive planning identify useful destinations (Zhao et al., 2025; Yuan et al., 2026). Belief-guided exploratory inference also revises hypotheses under partial observability (Bae et al., 2026). These advances motivate a specific question: *when an answer already appears plausible, what should the agent observe to establish*

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.

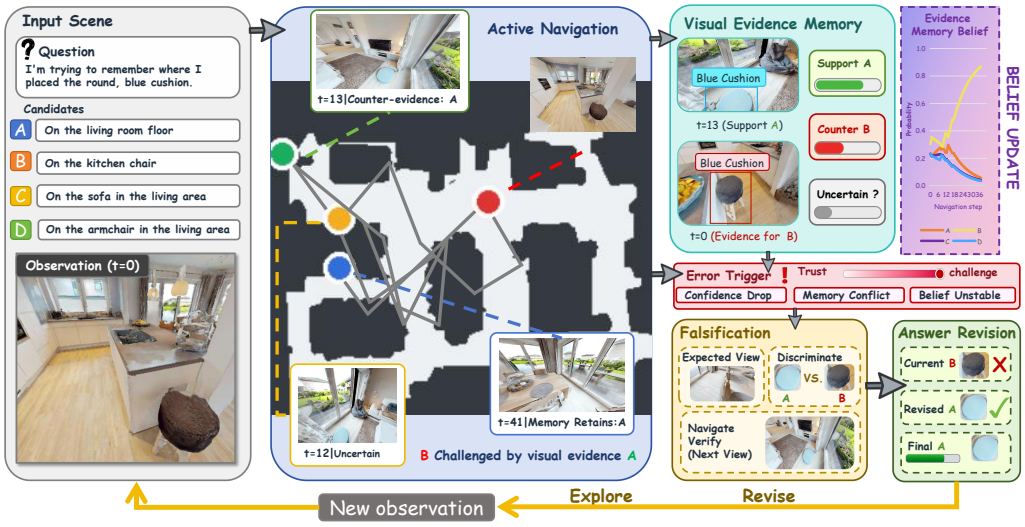

Figure 1: Active falsification couples evidence acquisition with answer revision. The initial image is the context of an erroneous prediction, not verified support for B. A later view supports A, and memory can retain this evidence after the target leaves view. This cushion-location schematic illustrates the decision process. Appendix E analyzes a separate recorded trajectory and distinguishes instantaneous predictions, persistent beliefs, and the terminal answer.

that it is wrong? A concentrated output distribution can reflect model priors or ambiguous input rather than an observed distinction between candidates (Guo et al., 2017; Li et al., 2023b).

We propose Active Falsification for Embodied Question Answering (AFE-EQA), a framework that treats the current answer as a revisable hypothesis and actively seeks observations capable of refuting it. Given a leading answer, the agent constructs testable counter-propositions that specify what visual evidence would contradict it or support an alternative. These propositions guide both where the agent looks next and how it interprets the acquired evidence. Motivated by confirmation bias in human reasoning (Nickerson, 1998), this design targets the tendency of an erroneous prediction to steer subsequent exploration toward preserving that error.

As illustrated in Figure 1, AFE-EQA connects exploration, evidence memory, and answer revision along a single trajectory. Three-state grounding assesses whether each observation is decidable, partial, or not-visible, controlling its influence on the answer belief and its retention for later review. A structured counter-evidence memory links each proposition to its source hypothesis and visual witnesses, preserving supported, refuted, and inconclusive assessments. Falsification-aware imagination translates unresolved propositions and distinctions between competing answers into observable search cues that guide viewpoint selection. At termination, error-aware verification reexamines retained images without conditioning on the previous answer. The full evaluated configuration then uses a resolver learned offline from scene-disjoint training data to compare candidate answers through visual and trajectory evidence and decide whether to revise the primary answer.

Our contributions include: ❶ formulating answer-conditioned active falsification for embodied exploration, using testable counter-propositions to identify observations that distinguish competing answers; ❷ developing AFE-EQA, which integrates evidence sufficiency, structured counter-evidence memory, and imagination to connect viewpoint selection with evidence-based answer revision; ❸ evaluating overall accuracy and the effects of answer switching through component ablations and paired corrections and regressions. On the 500-question evaluation, the Prism-based system improves from 58.0% to 62.0% accuracy, while a separate Qwen3-VL configuration achieves 64.0%.

2 Related Work

EQA and evidence acquisition. Early EQA combines navigation with answering (Das et al., 2018a), hierarchical control (Das et al., 2018b), interactive inspection (Gordon et al., 2018), and reasoning about multiple targets (Yu et al., 2019). Question-only baselines reveal that accuracy can also exploit dataset regularities without acquiring the intended visual evidence (Anand et al., 2018). OpenEQA broadens evaluation to open-vocabulary questions in episodic-memory and active settings (Majumdar et al., 2024). Explore until Confident introduces HM-EQA and couples semantic exploration with conformal stopping (Ren et al., 2024). Fine-EQA and EXPRESS-Bench emphasize exploration-aware evidence evaluation (Jiang et al., 2025). EfficientEQA studies open-vocabulary exploration and retrieval (Cheng et al., 2025). ToolEQA extends multi-step reasoning through tool use (Zhai et al., 2025). These tasks differ in answer spaces, episode subsets, and scoring rules, so their reported scores require protocol-specific interpretation.

Spatial memory and predictive exploration. Semantic exploration uses goal-conditioned maps to guide navigation (Chaplot et al., 2020). VLMaps, CLIP-Fields, and ConceptGraphs connect language to spatial representations (Huang et al., 2023; Shafiullah et al., 2023; Gu et al., 2024), while L3MVN uses language-model knowledge for navigation (Yu et al., 2023). For embodied reasoning, 3D-Mem retains compact visual memories (Yang et al., 2025), and MrSteve organizes what–where–when information across interactions (Park et al., 2025). ImagineNav, Navigation World Models, and GenEx use imagined or predicted observations for navigation and exploration (Zhao et al., 2025; Bar et al., 2025; Lu et al., 2025). Pred-EQA is particularly close in coupling predictive planning with specialized memories (Yuan et al., 2026). DIVRR refines ambiguous views and controls memory admission in static and dynamic environments (Lu et al., 2026). Our emphasis is the hypothesis-specific content of a search: a proposition records which answer it challenges, which alternative it supports, and which image was used to assess it.

Evidence-guided reasoning and self-correction. ReAct interleaves reasoning with actions that obtain external information (Yao et al., 2023). Reflexion and Self-Refine use feedback to revise language-agent behavior or outputs (Shinn et al., 2023; Madaan et al., 2023), whereas CRITIC and Self-RAG incorporate tool or retrieval evidence (Gou et al., 2024; Asai et al., 2024). Intrinsic correction does not consistently repair reasoning errors (Huang et al., 2024), and multi-turn interactions can expose failures absent from single-turn evaluations (Laban et al., 2026). Recent work examines belief-guided embodied inference, efficient information gathering, and belief deviation during agent training (Bae et al., 2026; Arumugam & Griffiths, 2026; Zou et al., 2026). Calibration and selective prediction distinguish confidence from reliable decisions (Guo et al., 2017; Geifman & El-Yaniv, 2019), while embodied abstention benchmarks expose ambiguous and unanswerable questions (Wu et al., 2026). AFE-EQA connects these concerns to physical evidence acquisition: revising a hypothesis depends on inspecting a discriminative view, rather than on producing another reasoning trace alone. Appendix F develops the connections to visual grounding, foundation models, and evaluation methodology.

3 Problem Formulation and Failure Mode

3.1 Embodied Question Answering under Partial Observability

We study EQA with a finite candidate set $\mathcal{Y}(q)$ of at least two valid answers, following the setting of Ren et al. (2024). An agent receives a question q about an unseen environment and acquires observations through navigation. The correct answer $y^\star \in \mathcal{Y}(q)$ is unavailable to the agent. Let o_t contain the visual and pose information available at step t, and let $h_t = (o_0, a_0, o_1, \ldots, a_{t-1}, o_t)$ denote the interaction history. A policy $\pi(a_t \mid q, h_t)$ selects an exploratory action or stops. The terminal step T satisfies $0 \le T \le B$, where B is the exploration-step budget and stopping is forced at B. The agent returns $\hat{y}_{\text{out}} = d(q, h_T)$ using an answer rule d that may revisit accumulated evidence.

At each observed step, the VLM scores the candidate answers using the current image, question, and available memory context. Writing $\ell_t(y) \in \mathbb{R}$ for the candidate scoring loss and $\eta > 0$ for its

temperature, we define

$$p_t(y) = \frac{\exp[-\ell_t(y)/\eta]}{\sum_{z\in\mathcal{Y}(q)} \exp[-\ell_t(z)/\eta]}, \qquad \hat{y}_t = \arg\max_{y\in\mathcal{Y}(q)} p_t(y). \tag{1}$$

Ties use a fixed candidate order. The confidence $c_t = p_t(\hat{y}_t)$ measures concentration of the normalized candidate scores and need not be calibrated to answer correctness. The instantaneous hypothesis $\hat{y}_t$ and the final output $\hat{y}_{\text{out}}$ are distinct, allowing evidence aggregation and verification to revise the answer.

3.2 Early High-Confidence Errors

For an early-step cutoff $k \in \{0, \ldots, B\}$ and confidence threshold $\tau \in (0, 1)$, define the set of early high-confidence error steps as

$$\mathcal{E}_{k,\tau} = \big\{t \in \{0, \ldots, \min(k, T)\} : c_t \geq \tau \text{ and } \hat{y}_t \neq y^\star\big\}. \tag{2}$$

An episode contains an early high-confidence error when $\mathcal{E}_{k,\tau} \neq \varnothing$. This diagnostic uses ground truth, whereas online decisions use observable evidence and model outputs. Repeated question-relevant views may leave the distinction between competing answers unresolved, allowing such an error to persist. The criterion distinguishes early-error recovery from ordinary terminal corrections. Measuring its recovery rate requires fixed k and τ and complete trajectories, as defined in Appendix G.

3.3 Evidence Sufficiency and Hypothesis Testing

We distinguish confidence from the sufficiency of the current visual evidence. A grounding query assigns normalized scores $\rho_t(s)$ to three states:

$$\mathcal{S} = \{\text{decidable}, \text{partial}, \text{not-visible}\}, \qquad g_t = \arg\max_{s\in\mathcal{S}} \rho_t(s). \tag{3}$$

where $\rho_t(s) \geq 0$ and $\sum_{s\in\mathcal{S}} \rho_t(s) = 1$, with fixed tie-breaking. A decidable view exposes the targets and attributes needed to answer. A partial view reveals relevant content but leaves the requested property unresolved. A not-visible view lacks a required target or anchor. These are model assessments of the current view, so $g_t = \text{decidable}$ does not guarantee that $\hat{y}_t$ is correct.

Evidence sufficiency also differs from evidence polarity. For an observable proposition associated with an answer hypothesis, an assessed observation can be supporting, refuting, or inconclusive. For example, seeing a mug's handle may leave its body color unresolved. Moving to reveal a red body can refute the hypothesis that the body is blue. Failing to see the mug provides no such refutation. With multiple alternatives, refuting the leading answer does not by itself identify the correct replacement. This motivates seeking distinguishing observations and grounding answer revision in their content. Section 4 operationalizes this hypothesis-testing perspective through memory, exploration, and verification.

4 Active Falsification for Embodied Question Answering

4.1 Overview of the Inference Policy

Figure 1 presents the module-level data flow, while Algorithm 1 specifies the execution order. During one trajectory, AFE-EQA maintains a spatial map, an answer belief b_t, selected images, and a ledger of hypothesis-conditioned propositions. The leading persistent hypothesis is $\bar{y}_t = \arg\max_y b_t(y)$. The instantaneous prediction $\hat{y}_t$ from Section 3 supports early-error diagnosis.

4.2 Three-State Visual Grounding

The grounding query in Eq. (3) asks whether the current image exposes the targets and properties required by the question. It instructs the VLM to assess direct visual evidence rather than substitute

Algorithm 1 Active falsification inference

Require: Question q, candidates $\mathcal{Y}(q)$, exploration budget B
Ensure: Final answer $\hat{y}_{\text{out}}$

1: **Initialize:** Initialize the spatial map, uniform belief b_{-1}, image buffers, and proposition ledger, and set $t = 0$.
2: **Observe and ground:** Observe the scene and compute answer scores p_t and grounding scores ρ_t.
3: **Update memory:** Update the belief, refresh scheduled search cues, and retain selected visual evidence.
4: **Seek counter-evidence:** Generate and assess counter-evidence propositions, and merge their cues with imagination.
5: **Explore:** If exploration continues, update the search map, navigate to a feasible viewpoint, increment t, and return to line 2.
6: **Resolve:** Form the primary answer, re-score retained images, extract candidate-evidence features, and apply the configured terminal rule in Eqs. (9)–(10).

scene priors or interpret an unseen object as absent. We use the resulting scores differently for belief updates and image retention:

$$r_t = \rho_t(\text{decidable}) + \alpha\rho_t(\text{partial}), \qquad u_t = \rho_t(\text{decidable}) + \beta\rho_t(\text{partial}), \tag{4}$$

where $0 \le \alpha \le \beta \le 1$. Hence $0 \le r_t \le u_t \le 1$. The evidence weight r_t controls how strongly a view can influence the answer. The retention score u_t permits a partial view to remain available for later review, even when it is too weak to justify a strong belief update.

Belief updates are gated by grounding, prediction confidence, and the gap between the two highest answer scores. An insufficiently grounded view receives zero update weight. For an initialized belief, the update has the form

$$b_t = (1 - \lambda_t)\widetilde{b}_{t-1} + \lambda_t p_t, \qquad 0 \le \lambda_t \le 1, \tag{5}$$

where $\widetilde{b}_{t-1}$ is the normalized previous belief after any contradiction-aware reweighting. A sufficiently grounded, confident disagreement can downweight the old leading answer before the blend. Repeated agreement under weak grounding reduces the update gate. These operations limit reinforcement from redundant predictions while preserving the ability of a new observation to change the hypothesis. Appendix A.2 gives the update details.

Images passing the retention threshold enter a bounded review buffer with their pose, step index, answer scores, and grounding scores. Similar views are merged using pose and visual-signature thresholds, keeping the higher-scoring representative. The buffer complements the main evidence memory by preserving partial observations for subsequent inspection.

4.3 Structured Counter-Evidence Memory

For a leading answer, the agent generates inspectable propositions that would contradict it or support a different candidate. A ledger entry contains the proposition text, source hypothesis, optional target alternative, object/attribute/relation cues, a quality score, its assessment state, and the indices and scores of witnessing observations. Generation is bounded by an active-proposition limit and a refresh interval. Duplicate or question-irrelevant entries are rejected. Retaining the source hypothesis prevents a later answer change from reversing an entry's meaning.

Each new proposition starts as *unseen*. When an eligible image is available, a separate query scores the labels supported, refuted, and inconclusive for that proposition. Let $v_{j,t}(s)$ be these normalized scores for proposition j, and let $s^\star_{j,t}$ be their fixed-tie-breaking maximizer. Its recorded state becomes

$$\sigma_{j,t} = \begin{cases} s^\star_{j,t}, & v_{j,t}(s^\star_{j,t}) \ge \tau_{\text{prop}}, \\ \text{inconclusive}, & \text{otherwise}, \end{cases} \tag{6}$$

with $\tau_{\text{prop}} \in (0, 1)$. When no observation is eligible, the previous state is unchanged. Each assessment is appended to the entry's witness history.

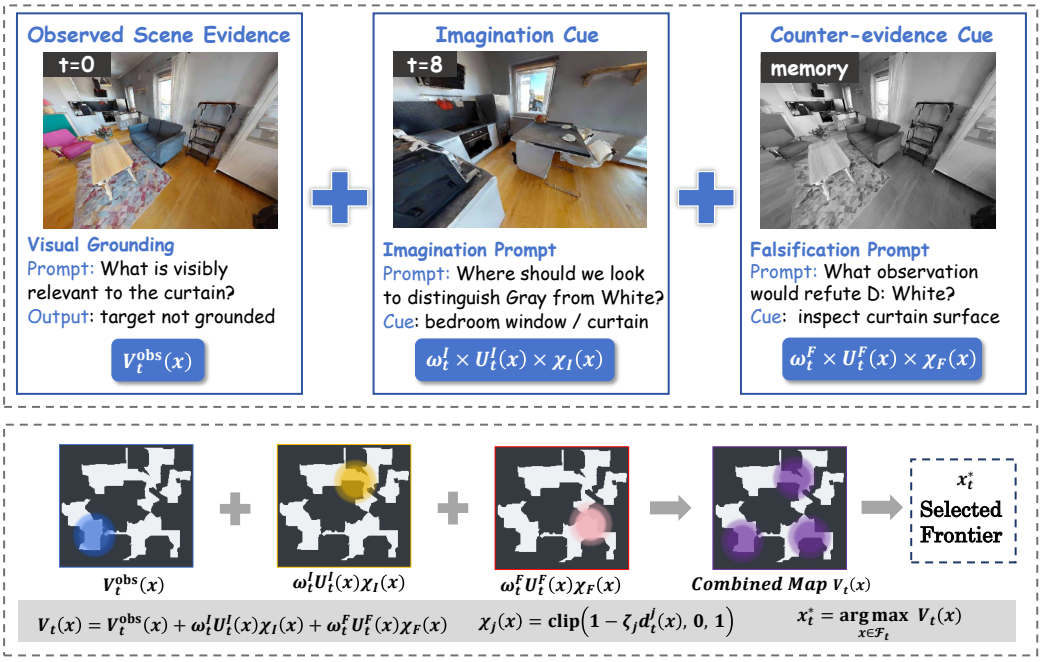

Figure 2: Schematic value decomposition using observations from example 322. At $t = 8$, the illustrated inputs are accumulated observed value, an imagination cue from the current view, and a counter-evidence cue from retained $t = 0$ memory. The heatmaps illustrate the terms in Eq. (7).

The labels refer to the proposition, not directly to the source answer. If a proposition states that the mug's body is red while the source answer is blue, *supported* records potential counter-evidence against blue. Conversely, refuting red does not establish blue when other colors remain possible. The ledger preserves this relationship, including inconclusive observations, rather than reducing all evidence to a single confidence value. Generated propositions remain search targets until assessed against images. Their text can inform subsequent prompts and exploration, while retained witness images can be revisited during visual verification.

4.4 Falsification-Aware Imagination and Viewpoint Selection

After an initial warm-up, the imagination prompt generates observable cues that distinguish the leading answer hypotheses. The grounding state directs the agent to locate an unseen target, improve a partial view, or verify an apparently decidable observation. Unresolved counter-evidence propositions provide additional cues, while previously attempted cues are excluded to reduce repetitive search.

The cues are converted into directional scores and projected onto the spatial map as virtual search targets. These locations encode where useful evidence might be found. They are not inserted into the observed scene representation as established facts. At a map location x, the semantic value supplied to the frontier planner is

$$V_t(x) = V_t^{\mathrm{obs}}(x) + \omega_t^I U_t^I(x)\chi_I(x) + \omega_t^F U_t^F(x)\chi_F(x), \tag{7}$$

where U_t^I and U_t^F are the smoothed imagination and counter-evidence cue maps. The weights $\omega_t^I, \omega_t^F \geq 0$ control their influence, and $\chi_j(x) = \mathrm{clip}(1 - \zeta_j d_t^j(x), 0, 1)$ attenuates uncertain predictions, with $\zeta_j \geq 0$ and uncertainty map d_t^j. The counter-evidence weight depends on the outstanding proposition ledger. Separate maps preserve the two cue contributions when enabled. A shared-map implementation uses the same spatial representation for both sources. Figure 2 illustrates these terms using observations from example 322.

The inherited frontier planner combines this semantic value with unexplored-space, traversability, and distance terms (Ren et al., 2024). It selects a feasible frontier and a viewing direction into

the unexplored region. This is a heuristic for acquiring discriminative observations, rather than an estimate of expected information gain. Newly acquired images are evaluated by the grounding and proposition queries before contributing evidence about the answer.

4.5 Error-Aware Verification and Answer Revision

At termination, the trajectory answer selector produces a primary distribution p^{pri} and answer y^{pri}. This interface accommodates the inherited grounding-weighted trajectory and memory summaries and is held fixed in comparisons that isolate verification. We retrieve a bounded set of images from the evidence and review buffers, remove duplicate step indices, and re-score each image using a prompt that omits the previous answer and memory conclusions.

For n retained images, let $v_i(y)$ be the native re-scored value for candidate y, $s_i \geq 0$ its retention score, and $r_i \in [0, 1]$ its evidence weight. Set $w_i = \max\{\epsilon, s_i(1 + r_i)/2\}$ and $\bar{w}_i = w_i / \sum_j w_j$, where $0 < \epsilon < 1$ is a numerical floor. The verifier pools these scores and normalizes over valid answers as a weighted geometric mean:

$$v(y) = \frac{\exp[\sum_{i=1}^{n} \bar{w}_i \log \max\{\epsilon, v_i(y)\}]}{\sum_{z \in \mathcal{Y}(q)} \exp[\sum_{i=1}^{n} \bar{w}_i \log \max\{\epsilon, v_i(z)\}]}. \tag{8}$$

This pools visual scores without assuming that the frames are statistically independent. Verification is unavailable when fewer than the required number of valid images are retained.

Guarded visual verification. Let $y^v = \arg\max_y v(y)$, $c_v = v(y^v)$, and Δ_v be the top-two verifier margin. The visual-verification branch accepts a revision when there are sufficient images, an observable risk signal, adequate confidence and margin, and no veto from the existing-evidence guard. Risk signals include review events, summary disagreement, and an early prediction that remains confident and stable. Appendix A.6 specifies its indicator A_v. This branch returns

$$\hat{y}_{\mathrm{vis}} = \begin{cases} y^v, & A_v = 1, \\ y^{\mathrm{pri}}, & A_v = 0. \end{cases} \tag{9}$$

Learned error-set resolution. The full evaluated configuration uses a learned resolver to decide among available answer candidates using their visual and trajectory evidence. This is a distinct terminal selection rule. Let $\mathcal{C}(q, h_T) \subseteq \mathcal{Y}(q)$ contain the retained candidates, including y^{pri}. An offline-trained ranker assigns $f_\theta(\phi(q, h_T, y))$ to each candidate's observable feature vector ϕ. Write y^e for the highest-scoring candidate and $d_e = f_\theta(\phi(q, h_T, y^e)) - f_\theta(\phi(q, h_T, y^{\mathrm{pri}}))$. For a training-selected threshold $\delta_e \geq 0$, the output is

$$\hat{y}_{\mathrm{out}} = \begin{cases} y^e, & y^e \neq y^{\mathrm{pri}} \text{ and } d_e > \delta_e, \\ y^{\mathrm{pri}}, & \text{otherwise.} \end{cases} \tag{10}$$

The feature vector includes candidate scores, temporal evidence summaries, and visual-comparison outputs. The ranker uses scene-disjoint training and threshold selection. Appendix D details the nested scene-disjoint training and out-of-fold evaluation protocol used for this resolver. Visual evidence informs the learned score, but passing its margin does not imply that the deterministic guard A_v also passes. Neither score is a calibrated guarantee of correctness. Both branches process evidence from one exploration trajectory. The full evaluated configuration uses the learned rule in Eq. (10).

5 Experiments

We evaluate final answer success and exploration cost, then examine the contribution of terminal revision. The 100-question and 500-question results are presented separately. The tables distinguish answer formats and answering backbones.

5.1 Experimental Setup

Task and observations. The main evaluation uses the 500 multiple-choice HM-EQA questions (Ren et al., 2024), covering identification, counting, existence, state, and location. The environment

Table 1: Exact-match results on 100 HM-EQA questions under open-vocabulary and multiple-choice protocols. Both settings report mean navigation steps.

Method	Answering backbone	Accuracy (%) ↑	Mean steps ↓
Open-vocabulary answers: matched 100-question evaluation			
Explore-EQA	Prismatic	42	38.87
Explore-EQA, black-box	Black-box VLM	39	38.87
EfficientEQA	GPT-4o	54	**30.16**
Multiple-choice answers: matched 100-question evaluation			
Base exploration	Prism 7B	60	37.87
+ Falsification	Prism 7B	61	46.93
AFE-EQA	Prism 7B	**65**	46.93
AFE-EQA	Qwen3-VL 32B	62	45.78

uses HM3D and Habitat-Sim (Ramakrishnan et al., 2021; Savva et al., 2019). The evaluation manifest fixes question identifiers 0–499, initial poses, and seed 42. Internal exploration variants share episode initialization, sensing interface, and budget rules. Their observations can differ because their navigation decisions differ. Terminal-selector comparisons on a fixed trajectory reuse the same recorded observations. Baseline results in Tables 1 and 2 are obtained from our reproduced runs.

RGB-D images have resolution 640×480, a 120° horizontal field of view, camera height 1.5 m, and downward tilt 30°. Depth and pose update a TSDF map with a 0.1 m grid. For navigable scene area S_i, the episode budget is

$$B_i = \left\lfloor 4.5\sqrt{S_i} \right\rfloor . \tag{11}$$

The resolved budget is recorded per episode. We distinguish a navigation action from an observation: a trajectory indexed $0, \ldots, T_i$ contains $T_i + 1$ observations and T_i intervening actions.

Models and terminal decisions. The Prism configuration uses Prism DINO–SigLIP+7B (Karamcheti et al., 2024), CLIP-based semantic matching (Radford et al., 2021), and Qwen3-VL-32B-Instruct for multiview evidence analysis (Bai et al., 2025). The second configuration replaces the navigation-time answering backbone with Qwen3-VL-32B-Instruct. The 58.0%, 59.0%, and 62.0% stages share the Prism backbone. The 64.0% result is a separate backbone comparison. Full AFE-EQA includes the terminal error-set resolver. Appendix B reports its paired answer transitions, and Appendix D details the scene-disjoint training and out-of-fold evaluation protocol used to obtain the final result.

Metrics. For multiple-choice questions, Accuracy (also reported as Success) is the percentage of final options matching the reference answer. With executed navigation-step count t_i and navigable scene area S_i, we report the mean scene-area-normalized cost:

$$\text{Area-normalized steps} = \frac{1}{N} \sum_{i=1}^{N} \frac{t_i}{S_i}. \tag{12}$$

Mean navigation steps and area-normalized steps are reported in their respective table columns. The normalization denominator is scene area S_i, distinct from the exploration budget B_i. Corrections and regressions are paired answer transitions. Their definitions and the scope of uncertainty analysis appear in Appendix G.

5.2 Results and Diagnostic Analysis

Table 1 separates exact-match results under the open-vocabulary protocol from those under the multiple-choice protocol. Table 2 reports the matched 500-question multiple-choice results, grouped by answering backbone.

With the Prism backbone, falsification-aware exploration improves Success from 58.0% to 59.0%, a net gain of five correct answers. Adding the terminal resolver reaches 62.0%, a further 15 correct answers without changing the navigation steps. This result aggregates held-out predictions from the five outer folds of the nested scene-disjoint evaluation. The complete system gains 20 correct

Table 2: Matched multiple-choice results on 500 HM-EQA questions. Component gains are computed within the Prism group. The Qwen3-VL row changes the backbone and is not a component ablation. Normalized steps follow Eq. (12).

Method	Answering backbone	Correct / 500	Success (%) ↑	Norm. steps ↓
Reproduced baselines under the matched evaluation				
Explore-EQA	Prismatic	260	52.0	**0.23**
3D-Mem	GPT-4o	252	50.4	0.33
Fine-EQA	GPT-4o-mini	280	56.0	0.26
DIVRR	Qwen2.5-VL 7B	299	59.8	0.36
Prism backbone: component comparison				
Base exploration	Prism 7B	290	58.0	0.24
+ Falsification	Prism 7B	295	59.0	0.35
AFE-EQA	Prism 7B	310	62.0	0.35
Qwen3-VL backbone: separate configuration				
AFE-EQA	Qwen3-VL 32B	320	**64.0**	0.35

answers, or 4.0 percentage points, over base exploration. These net changes do not determine how many individual errors were corrected: corrections can be offset by regressions. The Qwen3-VL configuration achieves 64.0% with 0.35 area-normalized steps, compared with 62.0% and 0.35 for the complete Prism configuration. This comparison changes model capacity and therefore does not isolate a falsification component.

Appendix B provides paired answer-revision statistics, scene-disjoint fold results, and question-type breakdowns. Appendix Table 7 compares the full system with three variants: removing the terminal resolver, removing imagination, and jointly removing memory and imagination. The adjacent five-panel curves show category-wise Success over trajectory progress for the 500-question evaluation.

The trajectory in Appendix E illustrates why answer scores, persistent belief, and the terminal answer must be distinguished. The initial prediction incorrectly places the banjo in the living area with confidence 0.9983. At step 39, a bedroom view reveals the banjo on the wall, making B the leading instantaneous answer and persistent belief. At the final observation the target is outside the camera view and both step-level quantities favor D, while the weighted trajectory evidence retains the correct terminal answer B. This selected trajectory shows how a discriminative observation can revise an early high-confidence error and remain available to the terminal decision.

6 Conclusion

We introduced AFE-EQA, a framework that treats the current answer as a hypothesis to be challenged through observation. Three-state grounding assesses evidence sufficiency, a counter-evidence ledger records testable alternatives and their witnesses, and contrastive imagination guides exploration toward distinguishing views. A terminal resolver uses the accumulated evidence to select an answer. On 500 HM-EQA questions, the Prism configuration improves success by 4.0 percentage points over base exploration, while the Qwen3-VL configuration reaches 64.0%. Matched ablations show lower success after removing the terminal resolver or imagination, with the largest decrease when memory and imagination are removed together. Paired analysis also identifies harmful revisions, showing that evidence acquisition and the decision to switch require separate evaluation. The results support hypothesis-directed exploration and evidence-based revision, while leaving their total computational cost to be quantified.

AI Use Statement

Generative AI tools were used for language editing, including improving grammar, clarity, and readability. All revisions were reviewed and approved by the authors, who remain responsible for the final manuscript, claims, citations, and results

Reproducibility Statement

Section 4 specifies the inference schedule and distinguishes visual verification from learned terminal resolution. Appendices A and C provide the update rules, visual-buffer operations, prompts, and structured records. Appendix B reports the staged results, paired answer transitions, and component ablations. Appendices D and G document the resolver protocol and conditional scene-bootstrap procedure. Appendix E presents a logged trajectory with selected observations and complete score curves. Section 5.1 summarizes the evaluation settings and metrics. Detailed pseudocode for reproducing the complete inference procedure is provided in the supplementary material.

References

Ankesh Anand, Eugene Belilovsky, Kyle Kastner, Hugo Larochelle, and Aaron Courville. Blindfold baselines for embodied qa. *arXiv preprint arXiv:1811.05013*, 2018.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Dilip Arumugam and Thomas L Griffiths. Toward efficient exploration by large language model agents. In *International Conference on Learning Representations*, volume 2026, pp. 40734–40772, 2026.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pp. 9112–9141, 2024.

Seohui Bae, Jeonghye Kim, Youngchul Sung, and Woohyung Lim. Align while search: Belief-guided exploratory inference for world-grounded embodied agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29642–29651, 2026.

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.

Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15791–15801. IEEE, 2025.

Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *2017 International Conference on 3D Vision (3DV)*, 2017.

Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in neural information processing systems*, 33:4247–4258, 2020.

Kai Cheng, Zhengyuan Li, Xingpeng Sun, Byung-Cheol Min, Amrit Singh Bedi, and Aniket Bera. Efficienteqa: An efficient approach to open-vocabulary embodied question answering. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 20357–20364. IEEE, 2025.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2432–2443. IEEE, 2017.

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.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pp. 2151–2159. PMLR, 2019.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

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.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Nan Duan, Weizhu Chen, et al. Critic: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, volume 2024, pp. 57734–57811, 2024.

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.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10608–10615. IEEE, 2023.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International conference on learning representations*, volume 2024, pp. 32808–32824, 2024.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

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.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: investigating the design space of visually-conditioned language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 23123–23144. PMLR, 2024.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. In *International Conference on Learning Representations*, volume 2026, pp. 54738–54778, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in neural information processing systems*, volume 36, pp. 34892–34916, 2023.

Taiming Lu, Tianmin Shu, Alan Yuille, Daniel Khashabi, and Jieneng Chen. Genex: Generating an explorable world. In *International Conference on Learning Representations*, volume 2025, pp. 52310–52335, 2025.

Xin Lu, Rui Li, Xun Huang, Weixin Li, Chuanqing Zhuang, Jiayuan Li, Zhengda Lu, Jun Xiao, and Yunhong Wang. Memory-guided view refinement for dynamic human-in-the-loop eqa. *CoRR*, March 2026.

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.

Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Junyeong Park, Junmo Cho, and Sungjin Ahn. Mrsteve: Instruction-following agents in minecraft with what-where-when memory. In *International Conference on Learning Representations*, volume 2025, pp. 16922–16960, 2025.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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. In *Proceedings of Robotics: Science and Systems*, 2024.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 4035–4045, 2018.

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.

Nur Muhammad (Mahi) Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. CLIP-Fields: Weakly Supervised Semantic Fields for Robotic Memory. In *Proceedings of Robotics: Science and Systems*, 2023.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1):91, 2006.

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

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

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.

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, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Bangguo Yu, Hamidreza Kasaei, and Ming Cao. L3mvn: Leveraging large language models for visual target navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3554–3560. IEEE, 2023.

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, 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*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Jiazhao Zhang, Anqi Li, Yunpeng Qi, Minghan Li, Jiahang Liu, Shaoan Wang, Haoran Liu, Gengze Zhou, Yuze Wu, Xingxing Li, et al. Embodied Navigation Foundation Model. In *International Conference on Learning Representations*, volume 2026, pp. 127293–127322, 2026a.

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

Xinxin Zhao, Wenzhe Cai, Likun Tang, and Teng Wang. Imaginenav: Prompting vision-language models as embodied navigator through scene imagination. In *International Conference on Learning Representations*, volume 2025, pp. 94387–94401, 2025.

Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13624–13634. IEEE, 2024.

Deyu Zou, Yongqiang Chen, Jianxiang Wang, Garry Yang, Mufei Li, Qing Da, James Cheng, Pan Li, and Yu Gong. Reducing belief deviation in reinforcement learning for active reasoning of llm agents. In *International Conference on Learning Representations*, volume 2026, pp. 21511–21545, 2026.

A Implementation Details and Evaluation Protocol

A.1 Observations and Candidate Scoring

The agent receives RGB observations together with depth and pose information used by the spatial mapper. Question answering and proposition assessment use the RGB image. The question prompt lists the valid answer candidates and requests their option letter. Available memory context can be included in the exploration-time answer prompt. The terminal per-image verification prompt uses the question and candidates without the previously selected answer.

For a candidate continuation with token sequence $(z_1, \dots, z_L)$, the Qwen3-VL scoring backend computes the mean continuation negative log-likelihood,

$$\ell_t(y) = -\frac{1}{L} \sum_{\ell=1}^{L} \log P_\theta(z_\ell \mid I_t, P_t, z_{<\ell}) \,, \tag{13}$$

where I_t is the image and P_t is the rendered prompt. Prompt tokens are excluded from the loss, and every scored continuation must contain at least one token. For numerical stability, the maximum of the negative scaled losses is subtracted before exponentiation. This common shift cancels in the normalization. For multi-token continuations, the length-normalized score need not equal the probability of the complete answer string.

The runner may score a fixed set of option slots containing invalid placeholders. Let $\widetilde{p}_t$ denote this native distribution. Masking invalid slots and renormalizing yields $p_t(y) = \widetilde{p}_t(y) / \sum_{z \in \mathcal{Y}(q)} \widetilde{p}_t(z)$ for valid y, as in Eq. (1). Belief updates use the valid-answer distribution. The inherited trajectory summary instead pools native scores before masking, as specified below. These conventions must be kept distinct: confidence values in Eq. (2) use the valid-answer distribution and must be reconstructed accordingly from native-score logs.

A.2 Grounding-Gated Belief Updates

Let $C(x) = \min\{1, \max\{0, x\}\}$ and Δ_t denote the gap between the two largest instantaneous answer scores. The gate thresholds lie in $[0, 1)$. The base update gate combines evidence weight, confidence, and margin:

$$\gamma_t = C\left(0.60C\left(\frac{r_t - \theta_r}{\max\{\epsilon, 1 - \theta_r\}}\right) + 0.25C\left(\frac{c_t - \theta_c}{\max\{\epsilon, 1 - \theta_c\}}\right) + 0.15C\left(\frac{\Delta_t - \theta_\Delta}{\max\{\epsilon, 1 - \theta_\Delta\}}\right)\right). \tag{14}$$

After computing Eq. (14), the implementation sets the gate to zero if the grounding requirement fails or both the evidence and confidence terms are zero. Repeated predictions with low grounding can multiply the gate by a decay factor. Review-related suppression can further adjust it. Denote the resulting gate by $\bar{\gamma}_t \in [0, 1]$. The episode starts with b_{-1} uniform over valid candidates, and the first observation is processed by the same gated update as later observations.

Let y^- be the previous belief's leading answer and let $y^+ = \hat{y}_t$. The contradiction check requires $y^+ \neq y^-$, $\Delta_t \geq \theta_{\text{ctr}}$, $c_t \geq \max\{\theta_c, b_{t-1}(y^-) - 0.15\}$, and $\bar{\gamma}_t \geq 0.4$. When it passes, the strength $d_t = C(0.6\bar{\gamma}_t + 0.4(\Delta_t - \theta_{\text{ctr}}) / \max\{\epsilon, 1 - \theta_{\text{ctr}}\})$ rescales the prior:

$$\widetilde{b}_{t-1}(y) = \frac{b_{t-1}(y)\, \xi_t(y)}{\sum_z b_{t-1}(z)\, \xi_t(z)}, \qquad \xi_t(y) = \begin{cases} \max\{0.05, 1 - \kappa d_t\}, & y = y^-, \\ 1 + \kappa d_t, & y = y^+, \\ 1, & \text{otherwise}, \end{cases} \tag{15}$$

where $\kappa \geq 0$. Otherwise, $\widetilde{b}_{t-1} = b_{t-1}$. With base rate $\lambda_{\text{base}} \in [0, 1]$ and boost coefficient $\lambda_{\text{boost}} \geq 0$, the effective blend weight is $\lambda_t = \min\{0.85, \lambda_{\text{base}}\bar{\gamma}_t + \lambda_{\text{boost}} d_t\}$ when the contradiction check passes and $\lambda_t = \lambda_{\text{base}}\bar{\gamma}_t$ otherwise. The positive floor in Eq. (15) preserves the old hypothesis's support. These expressions describe the update in exact arithmetic. The implementation also floors valid scores before normalization for numerical stability. These are score-update rules rather than Bayesian likelihood updates.

A.3 Visual Buffers and Proposition Records

The review buffer stores the image, pose, observation signature, step index, candidate distribution, grounding distribution, evidence weight r_t, and retention score u_t. A new entry is eligible when u_t exceeds the retention threshold. If its pose and observation signature are both sufficiently close to an existing entry, the higher-scoring entry is retained. Entries are sorted by retention score and truncated to the buffer capacity. This bounds the stored review images while allowing incomplete but informative views to survive.

The proposition ledger stores the text, source-answer index and label, target-alternative index and label when available, object, attribute, relation, room cue, quality score, assessment state, creation and verification steps, and witness score records. A proposition is active with respect to its source hypothesis. Generation prioritizes inspectable facts relevant to the requested attribute or relation. Checks are limited to a bounded number of active propositions per step and require an eligible image. Witness histories retain earlier assessments when the latest state changes, so the source of conflicting judgments remains inspectable. Grounding states assess whether the view can answer the question, whereas proposition states assess a particular proposed visual fact.

A.4 Outstanding Cues and Spatial Search

The planner assigns a nonnegative priority to the outstanding ledger. For active entries $\mathcal{J}_t$ with quality scores $q_j \in [0, 1]$, its form is

$$R_t = \sum_{j \in \mathcal{J}_t} q_j \, a(\sigma_{j,t}), \qquad \omega_t^F = \omega_{\max}^F C(R_t / s_R), \tag{16}$$

where $s_R > 0$ and the nonnegative state weights $a(\cdot)$ determine how unresolved and assessed propositions contribute. Quality filtering, a minimum priority, and question-specific applicability checks can disable this contribution, in which case $\omega_t^F = 0$. R_t is a search-priority score, not a probability that the answer is wrong. The cue maps are smoothed before the uncertainty gate and additive fusion in Eq. (7). They influence semantic frontier values while the geometric map controls feasible movement.

A.5 Primary Answer Summaries

The inherited trajectory interface includes a per-candidate maximum of evidence-weighted scores and the distribution from the most strongly grounded view. With $N(x) = x / \sum_y x(y)$ for a nonnegative vector of positive mass, these are

$$p^{\mathrm{w}} = N\left(\left[\max_{0 \le t \le T} r_t \widetilde{p}_t(y) \right]_{y \in \mathcal{Y}(q)} \right), \qquad p^{\mathrm{g}} = N\left(\left[r_{t^\star} \widetilde{p}_{t^\star}(y) \right]_{y \in \mathcal{Y}(q)} \right), \quad t^\star = \arg\max_t r_t. \tag{17}$$

Zero-mass inputs use the implementation's uniform fallback over valid candidates. The weighted trajectory summary is a maximum over time, not an arithmetic average. Pooling native scores before masking preserves the runner's operation order. Renormalizing each view first can change the result because the maximum is taken separately for each candidate. Memory evidence supplies an additional candidate distribution and support statistics. The configured primary selector chooses among these summaries using agreement, support, confidence, and margin guards. Its output is p^{pri} in Section 4.5. The selected source is recorded per episode. This selection uses evidence from one trajectory without combining predictions from separately run policies.

A.6 Terminal Verification Guards

Let J_{review} indicate a review event such as a confident trap, low-grounding lock, or stalled low-relevance exploration. Let J_{disagree} indicate disagreement among the available trajectory and evidence-summary leaders, and let J_{lock} indicate sufficiently stable and confident early predictions. The enabled trigger conditions form J_{risk} by disjunction. The inherited lock statistic uses the modal-answer fraction and mean confidence of native scores $\widetilde{p}_t$ in an initial observation window. It is an online trigger, distinct from the valid-answer, ground-truth diagnostic in Eq. (2).

Let Δ_{pri} be the primary distribution's top-two gap. The primary-answer guard permits an override if a review event has occurred or $\Delta_{\mathrm{pri}} \le \delta_{\mathrm{pri}}$. Separately, let G_{evidence} be a veto that protects a currently selected answer when the existing evidence has sufficient support count, confidence, and gap and agrees with that answer. The veto is released when the verifier agrees with the evidence leader, or when its confidence and gap pass the configured stronger-verifier thresholds. With verification enabled and $\mathbf{1}[\cdot]$ denoting a Boolean indicator, the acceptance rule is

$$A_v = \mathbf{1}[n \ge n_{\min}]\,\mathbf{1}[J_{\mathrm{risk}}]\,\mathbf{1}[c_v \ge \tau_v]\,\mathbf{1}[\Delta_v \ge \delta_v] \cdot \mathbf{1}[J_{\mathrm{review}} \vee \Delta_{\mathrm{pri}} \le \delta_{\mathrm{pri}}]\,\mathbf{1}[\neg G_{\mathrm{evidence}}]. \tag{18}$$

If scores or sufficient valid frames are unavailable, $A_v = 0$ before any confidence or margin is evaluated. When the verifier and primary leader coincide, the output is unchanged regardless of A_v. This guard defines the visual-verification branch in Eq. (9). It is distinct from the learned resolver in Eq. (10).

A.7 Execution Schedule and Resource Accounting

Answer and grounding queries operate on each usable observation. Imagination and proposition generation use refresh intervals, while proposition checking is capped per step. Image retrieval and terminal verification occur after navigation ends. The visual-only branch returns Eq. (9), whereas the full resolver configuration uses Eq. (10). Selector identity is part of the experimental configuration.

The persistent image buffers bound the number of frames re-read at termination. If M is the per-step proposition-check limit and K_v is the terminal image limit, proposition checking uses at most $(T{+}1)M$ assessments and per-image verification at most K_v image-level assessments. These counts describe semantic scoring calls. The number of model forward passes depends on candidate batching and tokenization. Generation calls, navigation distance, and latency therefore require separate accounting in the experimental results.

B Additional Development Results

B.1 Staged Performance

We report the staged comparison on a frozen 500-question HM-EQA development set. All configurations use multiple-choice answers and the same Prism navigation-time backbone. The base exploration policy answers 290 questions correctly. Adding falsification-aware exploration reaches 295 correct answers, and the complete policy with the error-set resolver reaches 310 correct answers.

Table 3: Staged results on the frozen 500-question development set. Gains are measured relative to base exploration.

Variant	Correct / 500	Success (%)	Gain (points)
Base exploration	290	58.0	0.0
+ Falsification-aware exploration	295	59.0	+1.0
Full AFE-EQA	310	**62.0**	+4.0

The full configuration produces 20 additional correct answers over base exploration. Its error-set resolver is evaluated with scene-disjoint out-of-fold predictions on the frozen development set. The staged comparison records the contribution of falsification-aware exploration followed by terminal answer revision under a fixed development protocol.

B.2 Answer-Revision Accounting

The terminal resolver operates on the falsification-aware trajectories, which answer 295 questions correctly before revision. It changes 113 answers, correcting 64 and regressing 49. The resulting net gain of 15 raises Success from 59.0% to 62.0%. Among changed answers, 56.6% are corrections and 43.4% are regressions. Conditional recovery is $64/205 = 31.2\%$, while conditional damage is $49/295 = 16.6\%$.

Table 4: Paired terminal-revision summary on the frozen development evaluation. Corrected and regressed are measured relative to falsification-aware exploration.

Comparison	Before	Final	Corrected	Regressed	Changed	Net
Falsification → Full AFE-EQA	295	310	64	49	113	+15

B.3 Variation across Scene-Disjoint Folds

Table 5 reports the paired result within each held-out scene fold. The fold sizes differ because all questions from one scene remain in the same partition. Four folds gain correct answers, while one fold has a negative net change.

Table 5: Scene-disjoint fold results for terminal revision on the frozen development set.

Fold	N	Before	Final	Corrected	Regressed	Net
0	78	53	48	5	10	−5
1	100	60	67	15	8	+7
2	118	77	82	20	15	+5
3	112	60	65	13	8	+5
4	92	45	48	11	8	+3
All	500	295	310	64	49	+15

B.4 Results by Question Type

Table 6 breaks down the same paired predictions by HM-EQA question type. Location and identification questions receive the largest positive changes, while existence questions show more regressions than corrections.

Table 6: Results by question type on the frozen development set. Before and Final are Success percentages.

Type	N	Before	Final	Corrected	Regressed
Count	60	53.33	53.33	8	8
Existence	123	69.11	60.98	13	23
Identification	40	70.00	82.50	5	0
Location	165	54.55	65.45	23	5
State	112	53.57	55.36	15	13
All	500	59.00	62.00	64	49

B.5 Matched Component Ablations

Table 7 pairs the matched removal comparisons with category-wise navigation diagnostics. The ablations share the 500-question manifest, answering backbone, initialization, and navigation budget. Candidate construction and resolver training are held fixed where the corresponding components are retained. The memory removal is evaluated jointly with imagination removal. The adjacent curves track Success on the same 500 questions as each trajectory advances from its initial observation to its endpoint. This trajectory-progress axis is distinct from the area-normalized steps cost in Eq. (12).

(a) Component ablations

Variant	Correct / 500	Success (%)
Full AFE-EQA	310	62.0
w/o terminal resolver	295	59.0
w/o imagination	275	55.0
w/o memory and imagination	258	51.6

(b) Success over trajectory progress

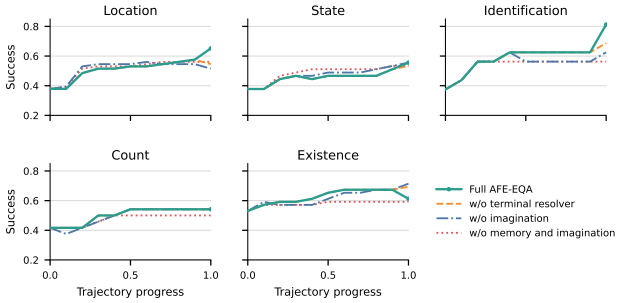

Table 7: Matched component ablations and category-wise navigation diagnostics on 500 questions. (a) Component ablations on the frozen development set. (b) Category Success over trajectory progress on the same set. Full AFE-EQA follows the falsification trajectory and applies the error-set resolver at the endpoint.

C Prompts and Structured Outputs

The blocks below specify the principal instruction text used by the corresponding modules. Bracketed fields are substituted at runtime. The VLM's model-specific chat template wraps these instructions and the input image. Question-specific guidance and memory context are inserted only in the fields indicated below.

C.1 Three-State Grounding Prompt

The grounding query evaluates a single view. Its output candidates are A, B, and C, mapped to decidable, partial, and not visible.

Three-state grounding — one image

```
Judge only whether this single image contains the direct visual
evidence needed to answer the question.
Question: [question]
Task-specific guidance: [optional guidance]
A. Decidable: every referenced target or anchor needed for the
answer is visible, and the requested state, relation, color, count,
or location can be judged directly.
B. Partial: a referenced target or anchor is visible, but the
requested property cannot yet be judged reliably.
C. Not visible: the required target or anchor is not visible in this
image.
Do not use common sense, scene priors, or absence from the image as
evidence for a negative answer. Answer with A, B, or C only.
```

The candidate scores are normalized before applying Eq. (4). The highest-scoring candidate supplies the descriptive state label, and the continuous scores supply the update and retention weights.

C.2 Counter-Evidence Generation and Assessment

The generation query requests observable propositions about the question's target. Each proposition is associated with the answer it challenges and, when identified, the alternative it supports.

Generate counter-evidence — JSON output

```
You generate counter-evidence for embodied multiple-choice question
answering.
Question: [question]
Choices: [candidate letters and texts]
Current answer hypothesis: [source hypothesis]
```

Generate counter-evidence — JSON output (continued)

```
If the current hypothesis is wrong, list observable visual facts that
would disprove it or support a competing answer.  Use facts an
embodied agent could check from RGB observations.  Avoid vague
feelings.
Every item must contradict the current answer hypothesis and support
a different answer option.  Do not include facts that would support
the current answer hypothesis.
Every item must be directly about the object, attribute, relation,
count, or location asked in the question.  Do not mention unrelated
objects or room details just because they are visible.
Return JSON only:  an array of objects with keys text, object,
attribute, relation, room, target_answer, quality_score.  Set
target_answer to the competing option letter that the fact would
support.  quality_score should be between 0 and 1.
```

Parsing validates the available fields, maps the target to a valid candidate, removes duplicates, and applies question-relevance checks. The quality score prioritizes proposed checks without establishing that a proposition holds. A separate image-conditioned query assesses each eligible proposition:

Assess a proposition — one image

```
You verify a visual proposition for embodied question answering.
Question:  [question]
Current answer hypothesis:  [source hypothesis]
Proposition to verify:  [proposition text]
Using only the current image, decide whether the image supports this
proposition, refutes this proposition, or is inconclusive.
A. Supported
B. Refuted
C. Inconclusive
Answer with one letter.
```

The resulting three candidate scores update the assessment state using Eq. (6). The witness record includes the step, proposition, state, scores, source hypothesis, and target alternative.

C.3 Contrastive Imagination Prompt

The contrastive prompt requests discriminative observations rather than a new answer:

Contrastive imagination — search cues

```
You are assisting embodied question answering by actively testing two
competing answers.
Question:  [question].
Candidate answer hypotheses:  [leading two hypotheses].
[grounding-specific guidance]
[memory context]
List 3 to 5 concrete visual observations that would distinguish the
top two hypotheses.  Include direct evidence for or against both, and
avoid restating either answer as a conclusion.  Prefer inspectable
objects, relations, states, colors, counts, or room locations.  Use a
short comma-separated list without explanation.
```

For a not-visible state, guidance prioritizes likely rooms, doorway transitions, and stable anchors leading to the missing target. For a partial state, it requests a closer or different viewpoint exposing the requested property. For a decidable state, it asks for a distinct viewpoint that could confirm or falsify the observed answer. An optional novelty instruction lists previously attempted cues and requests different cues or locations. Active ledger propositions are merged with the generated search text before spatial projection.

C.4 Terminal Visual Verification Prompt

Each retained image is re-scored separately with the following instruction:

Terminal visual scoring — one image

```
This image is one informative view collected while exploring the same
scene.  Answer the question using the visual evidence in this single
view.  [question and choices]
Answer with the option letter only.
```

This prompt supplies no primary-answer label or generated counter-evidence conclusion. Candidate scores from the retained images are pooled by Eq. (8). The visual-only branch uses Eq. (18). The full learned resolver uses evidence features and the separate rule in Eq. (10).

C.5 Diverse-View Answer Candidate

The following interface processes several views from one scene. Its output is mapped to an available answer candidate before use by the structured resolver.

Diverse-view candidate — answer text

```
These images are diverse views collected from the same scene.  Use
only visible evidence and identify the exact target in the question.
Compare the valid hypotheses by their meaning, reconcile
complementary views, and do not infer absence from a partial view.

Question:  [question]
[valid answer choices]

Answer with the complete answer text only.  Return exactly one line
in this format:  ANSWER: <complete answer text>
```

C.6 Pairwise Falsification

This prompt compares two answer meanings. The keep/switch labels are VLM evidence outputs supplied to the tree-based resolver, which makes the final decision.

Pairwise falsification — keep or switch

```
These images are diverse views collected from the same scene.  Use
only visible evidence and identify the exact target and relation
named in the question.  Actively look for evidence that contradicts
each hypothesis.  Do not favor the current hypothesis merely because
it was selected first.  For a negative or absence claim, require a
clear view of the relevant location.

Question:  [question]
Current hypothesis:  [current answer text]
Alternative hypothesis:  [alternative answer text]

keep means the current hypothesis has stronger direct visual support;
switch means the alternative hypothesis has stronger direct visual
support.  Answer keep or switch only.
```

C.7 Grouped Retrieved-Evidence Comparison

Images retrieved for different candidates are kept in explicit groups so that retrieval membership is not confused with proof of a candidate.

Grouped evidence — independent retrieval sets

```
Images 1-[N] are trajectory views retrieved for the current
hypothesis.  Images [N+1]-[N+M] are trajectory views retrieved
independently for the alternative hypothesis.  Retrieval is not
proof.  Inspect the pixels, reject similar but incorrect objects, and
actively look for evidence that falsifies each hypothesis.  Use only
direct visual evidence.  For a negative or absence claim, require a
clear view of the relevant location.

Question:  [question]
Current hypothesis:  [current answer text]
Alternative hypothesis:  [alternative answer text]

keep means the current hypothesis has stronger evidence; switch means
the alternative hypothesis has stronger evidence.  Answer keep or
switch only.
```

C.8 Structured Records and Parsing

Table 8 separates generated fields from observation-derived records. A generated quality score is a priority estimate, not visual support. Invalid option labels and malformed outputs must not become new answer candidates.

Table 8: Principal fields in the counter-evidence interface. Runtime identifiers connect records to the hypothesis and observation that produced them.

Field	Meaning
`text`	An inspectable visual proposition
`object`, `attribute`, `relation`	Question-relevant visual search cues
`room`	Optional search-location cue
`target_answer`	Valid competing option supported if the proposition holds
`quality_score`	Generated priority in $[0, 1]$
`source_hypothesis`	Answer the proposition was generated to challenge
`assessment_state`	Unseen, supported, refuted, or inconclusive
`witness_record`	Observation index, assessment scores, and source/target identifiers

D Scene-Disjoint Error-Set Resolver Protocol

This appendix documents the scene-disjoint training and out-of-fold evaluation protocol for the terminal error-set resolver. The fold construction and held-out sizes are specified below, and the corresponding results are reported in Table 5.

D.1 Candidate and Feature Construction

The resolver ranks answer candidates using structured evidence obtained from one exploration trajectory. The candidate table contains 1,079 rows for 500 questions. Candidate-level inputs include stepwise VLM scores, margins, relevance-weighted scores, memory statistics, temporal support, proposition assessment events, candidate source, confidence, agreement, and answer-type features. Additional frozen inputs include candidate embeddings, direct-generation candidates, pairwise visual comparisons, and high-resolution multiview outputs. A generated candidate must be mapped to a valid answer before selection. Multiple candidate records associated with one question must remain in the same data partition.

The visual-language interfaces produce evidence features. The final resolver is an Extra Trees learner (Geurts et al., 2006), implemented with scikit-learn (Pedregosa et al., 2011). It is not a prompt-based classifier. The model's score gap in Eq. (10) is distinct from the VLM's answer probability and the deterministic visual guard. Ground truth supplies candidate-correctness labels inside training partitions. Neither the reference answer nor a correctness indicator is an inference feature.

D.2 Nested Scene-Disjoint Evaluation

Complete scenes are assigned to five outer folds. Since all questions from the same scene remain in one partition, the held-out folds contain 78, 100, 118, 112, and 92 questions, respectively. For outer fold k, the remaining $500 - N_k$ questions form the outer training partition. This partition is further divided into four scene-disjoint inner folds, which are used to select the ranker configuration, cluster count, and switching threshold. After selection, the chosen procedure is refitted on the complete outer training partition and evaluated once on the corresponding held-out scenes. This nested protocol separates hyperparameter selection from outer-fold evaluation and reduces the selection bias that can arise when the same cross-validation estimates are used for both purposes (Varma & Simon, 2006; Cawley & Talbot, 2010).

Within each inner split, all learned transformations of the candidate features, including cluster fitting, are estimated from the inner training folds and applied to the inner validation fold. After hyperparameter selection, these transformations are refitted using the complete outer training partition before evaluation on the held-out outer fold. Frozen pretrained embeddings may be computed for all questions because their parameters are not estimated from answer labels. Prompts, candidate-generation rules, and embedding-processing definitions are fixed before outer-fold evaluation and are not selected using held-out answer outcomes. The evaluation manifest records the fold assignments, input and code hashes, selected configurations, and scene-overlap checks, providing an auditable record of the evaluation procedure.

Table 9: Search space and execution settings for the nested scene-disjoint resolver evaluation. Hyperparameters are selected within each outer training partition.

Setting	Value or candidate set
Outer / inner folds	5 / 4, grouped by scene
Outer-fold sizes	78, 100, 118, 112, 92
Seed	42
Ranker	Extra Trees, 200 trees
Maximum depth	4, 8, unlimited
Minimum leaf size	2, 5, 10
Feature subsampling	Square root, 0.5
Candidate cluster counts	3, 4, 5, 6, 8
Switching threshold	0, 0.025, 0.050, 0.075, 0.100, 0.150, 0.200, 0.250, 0.300
Workers	4
Recorded software	Python 3.9.18; NumPy 1.26.4; SciPy 1.12.0; scikit-learn 1.4.1.post1

D.3 Switching Semantics and Audit Fields

The highest-scoring candidate is compared with the primary candidate from the same question. Switching requires a *strictly* positive margin above the fold-selected threshold. In particular, a zero threshold does not accept tied scores. Recomputing this condition from the 500 recorded score gaps reproduces all 500 archived trigger indicators. Fixed candidate ordering resolves score ties, and including the primary candidate guarantees a nonempty selection set.

Each held-out record contains its question identifier, scene identifier, outer fold, question type, assigned cluster, reference answer, base and revised predictions, correctness indicators, trigger flags, and candidate/base scores. Correctness indicators are recorded for evaluation after prediction. They are not inputs to the deployed ranker. The archive also retains fallback counts: six questions in outer fold 0 and five in fold 2 use the recorded fallback path. Reproducing the full learner requires the hashed training implementation, including that fallback and feature preprocessing, in addition to the settings table.

D.4 Scope of the Learned Component

Scene-disjoint out-of-fold evaluation tests prediction on scenes excluded from each fitted resolver. It is different from training one resolver on all 500 questions and reporting its in-sample accuracy.

It is also different from a new external test set collected after development. The online exploration framework can use a frozen resolver at deployment, but the overall method is not training-free when this learned terminal component is enabled. Using one trajectory avoids combining separately explored policies, while supervised fitting remains part of the answer selector.

E A Complete Error-Recovery Trajectory

E.1 Episode and Selection Criterion

This appendix examines episode 112 from scene `00176-wcJYziD5pmF`, floor 0. It is selected because an early high-confidence wrong prediction is followed by discriminative visual evidence and a correct terminal answer. Selection is qualitative, not random. The question asks where the banjo is kept. The four options are A: resting in the living area, B: on the bedroom wall, C: hanging in the hallway, and D: leaning against the kitchen cabinets. The reference and recorded final answer are B.

The trace contains 68 observations, indexed 0–67. Figure 3 presents the source images in chronological order. The image at step 0 and the stored "wrong-hypothesis evidence" image are the same view, so they are shown once. Their association with A records the model's erroneous initial interpretation of the banjo's location.



(a) Step 0: wrong prediction A.



(b) Step 29: the target remains unseen.



(c) Step 39: bedroom-wall evidence.



(d) Step 67: target out of view.

Figure 3: Four original observations from the selected trajectory. At step 39 the bedroom-wall view makes B the instantaneous prediction and the leading persistent belief. By the final observation the target is outside the camera view and both step-level quantities favor D, while the recorded terminal answer remains B. Images are displayed without content editing.

Table 10: Recorded values at key observations. p_t and b_t are different distributions. Values are rounded to four decimal places from the full-precision trace.

Step	$\hat{y}_t$	$p_t(A)$	$p_t(B)$	$b_t(A)$	$b_t(B)$	$\arg\max b_t$
0	A	0.9983	0.0006	0.3037	0.2321	A
29	A	0.6744	0.2189	0.4786	0.1669	A
39	B	0.0000	0.9999	0.1753	0.6183	B
67	D	0.0022	0.0004	0.0970	0.1313	D



Figure 4: Recorded top-down route for episode 112. The route and key observations support a spatial reading of the case. The map is a logged visualization, not evidence that every relevant surface was inspected.

E.2 Prediction, Belief, and Evidence Are Different Quantities

Table 10 separates instantaneous answer scores from persistent belief. At step 0, the instantaneous A score is 0.9983, while the persistent belief in A is 0.3037. At step 29, A remains the leading instantaneous answer and persistent belief, although its instantaneous score has fallen to 0.6744. At step 39, the view of the banjo on the bedroom wall raises the instantaneous B score to 0.9999 and the persistent belief in B to 0.6183.

Figure 4 situates the key observations along the recorded route, while Figure 5 shows how the instantaneous scores and persistent belief evolve over the trajectory. At step 67, the target is no longer visible. The instantaneous prediction and persistent belief favor D, whereas the weighted trajectory summary, retained evidence, and recorded terminal answer favor B. The retained views in Figure 6 show the visual context available for revisiting the answer. The example illustrates why a terminal answer need not equal the prediction from the final camera view or the final step-level belief. It does not isolate which module caused the recovery, nor does it establish the accuracy of the three-state classifier: this archived trace stores binary relevance scores rather than three-state grounding annotations.

F Extended Context and Design Boundaries

F.1 From Visual Answering to Controlled Evidence Acquisition

VQA evaluates answering from an input image (Antol et al., 2015). CLEVR and GQA further probe compositional reasoning and the relation between question semantics and visual content (Johnson et al., 2017; Hudson & Manning, 2019). These settings help explain why evidence requirements are question-dependent: identifying an object, comparing two targets, and establishing absence need

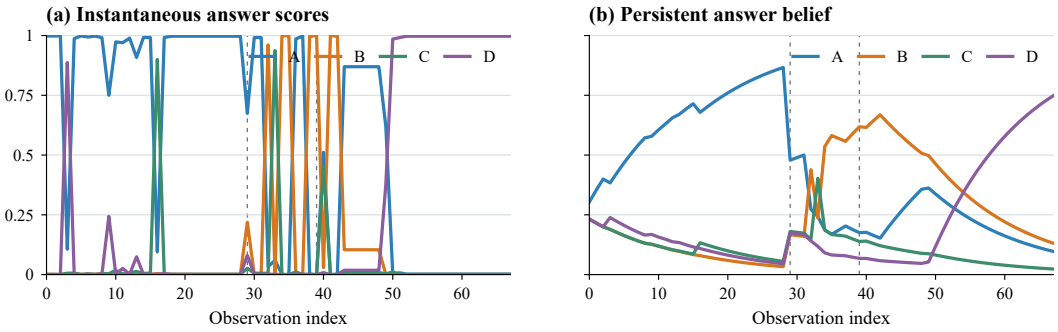

Figure 5: Instantaneous answer scores (left) and persistent belief (right), redrawn from the complete 68-observation log. Vertical guides identify observations 29 and 39. Both step-level quantities differ from the recorded terminal answer at the final observation.

different observations. EQA adds control over the observations themselves. A missing view is an information-acquisition problem, while an incorrect inference from a sufficient view is a reasoning problem. Accuracy alone can conflate the two, especially when question priors support plausible guesses (Anand et al., 2018).

Real indoor datasets provide different visual and spatial conditions. Matterport3D contributes building-scale RGB-D reconstructions (Chang et al., 2017). ScanNet supplies annotated indoor RGB-D sequences and reconstructions (Dai et al., 2017), and HM3D scales the collection of navigable reconstructed environments (Ramakrishnan et al., 2021). Habitat supports simulation over such scene assets (Savva et al., 2019). These resources explain the environment assumptions, but evaluating in one collection does not by itself establish transfer to another. An unseen scene in a resolver split is also not necessarily unseen during foundation-model pretraining.

F.2 Foundation Models and Visual Grounding

Contrastive image–text learning provides reusable semantic alignment in CLIP (Radford et al., 2021). DINOv2 supplies self-supervised visual representations (Oquab et al., 2024), and SigLIP uses a sigmoid image–text objective (Zhai et al., 2023). Prismatic investigates how visual representations, language backbones, and training choices affect VLM behavior (Karamcheti et al., 2024). BLIP-2 bridges frozen image and language models through a learned querying interface (Li et al., 2023a). Visual instruction tuning adapts models to multimodal instructions (Liu et al., 2023). Qwen3-VL provides the multimodal model family used in the additional configuration (Bai et al., 2025). These works motivate explicit backbone reporting rather than attributing a cross-model performance difference to an exploration module.

Visual competence does not guarantee evidence-faithful output. Object hallucination has been studied in captioning (Rohrbach et al., 2018) and large VLMs (Li et al., 2023b). Our grounding query assesses whether a view exposes the requested distinction, while a proposition query assesses the polarity of one specific fact. Neither assessment is external ground truth. Repeated calls to the same model can share the same perceptual error, and multiple frames can be redundant. Consequently, a geometric pool of their scores is a decision statistic rather than an independent-observation likelihood.

F.3 Reasoning Traces, External Feedback, and Physical Tests

Chain-of-thought prompting elicits intermediate reasoning (Wei et al., 2022), while self-consistency aggregates multiple sampled reasoning paths (Wang et al., 2023). Self-Refine and Reflexion use generated feedback or experience to revise outputs and behavior (Madaan et al., 2023; Shinn et al., 2023). CRITIC and Self-RAG add external tool or retrieval evidence (Gou et al., 2024; Asai et al., 2024). These mechanisms differ from executing a physical action that reveals an unobserved attribute. A second explanation from the same image cannot supply pixels that were never acquired.

The sequential structure is naturally related to decision-making under partial observability (Kaelbling et al., 1998). Our persistent belief is a heuristic score state, not a posterior derived from a



Figure 6: Eight retained visual-memory views supplied with the case. The views preserve spatial context after the camera points elsewhere. Their relevance to an answer must still be assessed from their content.

specified transition and observation model. Likewise, the semantic cue map is not an expected information-gain calculation. Language-agent exploration and belief-deviation studies provide complementary perspectives on how uncertainty affects action selection (Arumugam & Griffiths, 2026; Zou et al., 2026). The specific object of falsification here is a candidate answer and the visible proposition that distinguishes it from an alternative.

F.4 Confidence, Abstention, and Decision Scope

Calibration measures whether predicted probabilities correspond to observed correctness (Guo et al., 2017). Language models can express useful information about what they know while remaining imperfectly calibrated (Kadavath et al., 2022). Conformal prediction provides set-valued coverage under its assumptions (Angelopoulos & Bates, 2023), and Explore until Confident applies calibration

to sequential EQA stopping (Ren et al., 2024). SelectiveNet instead learns a rejection mechanism together with prediction (Geifman & El-Yaniv, 2019). Our acceptance thresholds define heuristic operating points without conformal coverage guarantees.

Embodied abstention benchmarks address cases where reliable answering may be impossible (Wu et al., 2026). The present multiple-choice formulation assumes a reference answer among the candidates and does not add an abstention output. "Not visible" therefore controls evidence use, rather than serving as another answer option. Negative and absence claims are particularly sensitive to this distinction: an occluded or unvisited region cannot certify that a target is absent.

F.5 Limitations of the Current Evidence

The framework combines several fallible judgments: target recognition, grounding, proposition assessment, cue generation, and final ranking. An incorrect early target association can affect several of them simultaneously. Navigation may also fail to reach a decisive viewpoint within budget. The selected trajectory illustrates one recovery but cannot establish aggregate suppression of confirmation bias. The component ablations measure changes in final Success, while paired revision counts expose both corrections and harmful switches. These quantities do not measure recovery specifically from early high-confidence errors, which requires a pre-specified cohort under Eq. (2). The joint removal of memory and imagination also leaves the independent effect of memory unresolved. Finally, navigation steps capture only part of the resource cost because multiview verification and candidate generation add model calls.

G Evaluation Definitions and Resource Accounting

G.1 Paired Corrections and Regressions

For episode i, let z_i^0 and z_i^1 be correctness indicators for the base and revised answers. Define

$$C = \sum_i (1 - z_i^0) z_i^1, \qquad R = \sum_i z_i^0 (1 - z_i^1), \qquad \mathrm{Acc}_1 - \mathrm{Acc}_0 = \frac{C - R}{N}. \tag{19}$$

The equality follows by expanding $z_i^1 - z_i^0 = (1 - z_i^0) z_i^1 - z_i^0 (1 - z_i^1)$ for each episode. Thus a net gain of 20 on 500 questions is 4.0 percentage points, but does not establish that exactly 20 errors were repaired. A wrong-to-wrong switch contributes neither a correction nor a regression.

Let S be the number of changed answers. Correction precision is C/S, while the harmful-switch fraction is R/S. Conditional recovery and damage rates are $C/\sum_i (1 - z_i^0)$ and $R/\sum_i z_i^0$, respectively. They use different denominators and should not share one "recovery rate" label. A rate with a zero denominator is undefined and is reported as a dash. Early-high-confidence recovery further restricts episodes to those satisfying Eq. (2). This cohort must be specified before inspecting final outcomes. Selecting only episodes that the method later corrects would bias the estimate.

G.2 Scene Bootstrap and Its Interpretation

The archived analysis uses a paired scene bootstrap (Efron, 1979). Each replicate samples 266 scene identifiers with replacement and includes all questions of each sampled scene with that multiplicity. Accuracy is recomputed as total correct divided by total sampled questions. Both selectors use the same resampled scenes. We use 10,000 replicates, seed 42, and the 2.5th and 97.5th percentiles of each paired difference.

This calculation quantifies variation across sampled scenes conditional on the available out-of-fold predictions. It does not refit the rankers or repeat hyperparameter selection, so it excludes variability from the full training-and-selection procedure. Nested selection and unbiased evaluation remain distinct issues (Varma & Simon, 2006; Cawley & Talbot, 2010). Moreover, different question categories and cross-validation folds of one benchmark are not independent datasets on which to apply a multiple-dataset rank comparison (Demšar, 2006). This archived bootstrap procedure does not establish statistical significance for the final 62.0% and 64.0% configurations.

G.3 Budget and Resource Accounting

Comparisons should align initialization, sensors, actions, map resolution, and budget rules. Different policies acquire different images. Replaying one trajectory holds the acquired evidence fixed when comparing selectors. Area-normalized steps follow Eq. (12) and averages the per-episode ratio t_i/S_i. This requires consistent action counts and scene-area measurements. With a fixed trajectory and denominator, terminal answer revision alone cannot change this navigation cost.

Inference accounting separates scoring, grounding, proposition checks, imagination, retrieval, and verification. Query count, model forward passes, tokens, and elapsed time measure different costs: fewer navigation steps can coexist with more computation per observation. Backbone, candidate pool, and training data are further controls for paired comparisons.

G.4 Mathematical Invariants

The three-state scores sum to one and $0 \leq \alpha \leq \beta \leq 1$, giving $0 \leq r_t \leq u_t \leq 1$. The positive factors in Eq. (15) preserve positive total mass. Normalization therefore produces a valid $\widetilde{b}_{t-1}$, and the convex blend in Eq. (5) remains a distribution for $\lambda_t \in [0, 1]$.

For Eq. (8), positive floored weights give $\sum_i \bar{w}_i = 1$. Floored candidate scores make every logarithm finite. Exponentiation gives positive values, and normalization over the finite valid-answer set gives a distribution. These facts establish numerical well-definedness, not calibration or correctness. The empty-image case is handled before evaluating that equation. Finally, a supported counter-proposition can challenge its source hypothesis. Refuting one alternative does not establish the source answer when other alternatives remain possible.