

NoisyEQA: benchmarking Embodied Question Answering with imperfect queries from non-expert users



Tao Wu¹, Chuhao Zhou¹, Haozhi Cao¹, Yenheng Wong¹, Lin Gu^{2,3} and Jianfei Yang^{1,*}

¹ Mars Lab, Nanyang Technological University, Singapore

² RIKEN Center for Advanced Intelligence Project, Wako, Japan

³ Tohoku University, Sendai, Japan

* Correspondence author; E-mail: jianfei.yang@ntu.edu.sg.

Highlights:

- Introduces NoisyEQA, a benchmark dataset with diverse noise types for EQA robustness.
- Proposes an LLM-powered framework to generate scalable and controllable noisy questions.
- Develops a Self-Correction mechanism to detect and correct noisy user queries.
- Presents an evaluation framework assessing EQA agents under realistic noisy conditions.

Abstract: Embodied Question Answering (EQA) enables robots to explore the environment and answer human questions, which is important for human-robot interaction and has been significantly enhanced by the recent advancement of Vision-Language Models (VLMs). However, EQA in real-world scenarios remains challenging, as human-posed questions often contain noise that can interfere with an embodied agent's reasoning, bringing challenges for language beginners and non-expert users. To address this, we introduce a NoisyEQA benchmark designed to evaluate the ability of the robot to identify and correct noisy questions. NoisyEQA is inspired by three common types of noise observed in real-world applications: Memory Noise, Perception Noise, and Semantic Noise, generated through an automated dataset creation framework. Additionally, we propose a 'Self-Correction' prompting mechanism to enhance EQA against noise robustness and a novel evaluation metric to measure both noise detection capability and answer quality. Our comprehensive study reveals that current embodied agents often struggle to detect noise in questions, leading to responses that frequently contain incorrect information. Through our self-correct prompting mechanism, we can effectively improve the accuracy of agent answers.

Keywords: Embodied Question Answering; navigation; embodied LLM; active agents

1. Introduction

Embodied Question Answering (EQA) is a task in which an embodied agent of the robot actively explores a scene, builds a comprehensive understanding of the environment, and answers questions about the scene



Copyright©2026 by the authors. Published by ELSP. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.

using natural language [1–3]. Unlike Visual Question Answering (VQA) [4,5] based on a fixed view of visual content without exploration, EQA requires agents to actively explore in the real world environment to collect images to answer questions [1].

Although existing benchmarks assume that the messages posed by humans are clear and accurate, this assumption fails to align with the complex and error-prone nature of interactions between humans and robots. In social psychology, a suggestive question is defined as one that surreptitiously inserts a false presupposition as if it were established truth [6], and human studies have shown that both question wording and subtle cues within questions can influence how people recall events and bias their responses [7,8]. Non-expert users, who lack specialized communication skills, often find it difficult to convey their thoughts precisely. This is particularly true for the elderly and Alzheimer’s patients, who may struggle to pose precise questions to robots. In an aging society, as seen in Japan and Singapore [9], elderly robots play a crucial role in interacting with diverse users, especially non-expert users such as the elderly. This human-robot interaction is of great concern in EQA systems. Since EQA systems engage with a wide range of clients and situations in real-world scenarios, incorrect responses can not only frustrate users but also potentially cause harm. For instance, when a robot helps an Alzheimer’s patient find a frequently misplaced item like glasses, the incorrect assumptions from the user can cause the robot to repeatedly search an unrelated area, resulting in inefficiency or serious safety hazards (e.g., collision with obstacles or patients).

To understand existing EQA systems’ limitations in handling imperfect queries in the real world from non-expert users, we examine the largest EQA datasets, Explore-EQA [10] and OpenEQA [3]. As in Figure 1a and the pink path in Figure 1c, when faced with a misguided question from a drunk person, the agent may search the same spot repeatedly and give incorrect answers. The quantitative results in Figure 1d further demonstrate our advocacy. We define such inaccuracies and ambiguities in user queries as ‘noise’, which can impede accurate interpretation and response. The results clearly expose a major flaw in current EQA systems when they are tasked with handling noisy questions.

Hence, we introduce a novel multimodal benchmark, NoisyEQA, to evaluate embodied agents’ abilities in handling noisy questions. After analyzing typical noise sources in human-posed questions, we categorize them into three main types including Memory Noise, Perception Noise, and Semantic Noise in Figure 1b. Each noise source is divided into objects and five attributes—position, color, material, count, and shape. Memory noise, related to memory errors, involves incorrect recall of objects or their attributes. Perception noise comes from inaccurate object descriptions due to visual disturbances or occlusions. Semantic noise, stemming from misinterpretations of the environment, occurs when original objects are replaced by semantically related but incorrect ones. Leveraging these definitions, we meticulously design a systematic framework powered by a Large Language Model (LLM) to automatically generate 500 noisy questions, thereby forming our proposed NoisyEQA benchmark.

Furthermore, we take the first step in enhancing EQA agents’ robustness towards noisy questions by proposing a simple yet effective Self-Correction mechanism. This mechanism, acting as plug-ins, can be seamlessly incorporated into any EQA system. Its core is to prompt agents to inspect the noise by comparing their visual observations with human questions. Once noise is detected, agents are asked to provide a clear explanation of the noise source, based on which noise-free responses can be generated.

Notably, the Self-Correction mechanism can be easily extended to a human-in-the-loop model. Here, agents' corrected responses are verified by humans before execution. This ensures that our method fully meets the AI Act's transparency requirements [11], which are absolutely vital for embodied systems.

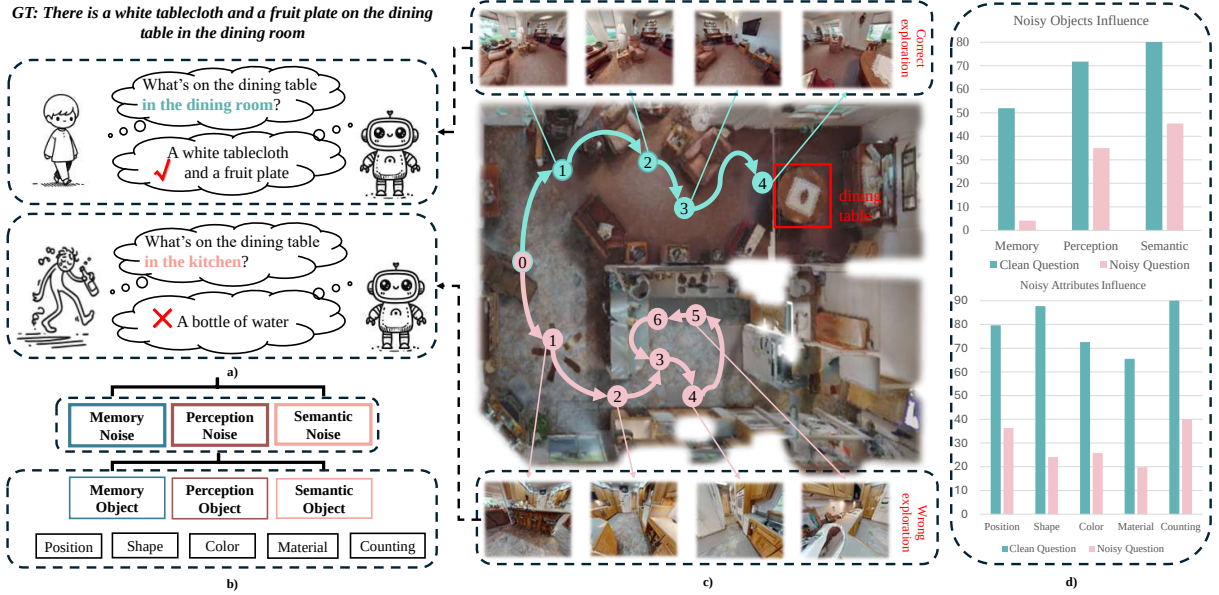


Figure 1. Overview of noisy human questions and their effects on embodied question answering. **(a)** A dining table is located in the dining room; a drunk person misremembers its location, causing the agent to produce a fabricated answer; **(b)** A systematic taxonomy of noisy human questions; **(c)** Comparison of the agent's exploration paths under an accurate question (green) and a noisy question (pink); the latter may get trapped in a loop, posing potential safety risks; **(d)** The accuracy of Explore-EQA decreases significantly on noisy questions compared to clean questions.

Finally, we argue that existing metrics, such as accuracy, cannot adequately evaluate agents' performance on noisy questions. Thus, we propose a novel evaluation framework tailored to noisy questions. Specifically, we design an evaluation scale from 1 to 5 to assess agents' capabilities in noise detection, correction, and accurate response generation. To thoroughly examine agents' fine-grained noise-handling abilities, two noise-related metrics, Detection Rate (DR) and Correction Rate (CR), are introduced to determine whether agents can detect and correct noise. These metrics provide a comprehensive and highly practical framework for evaluating agents' performance in complex real-world EQA tasks.

In summary, our contributions lie in 4 folds:

- (1) We propose the NoisyEQAbenchmark, studying how various types of noisy questions affect the EQA of robots, which is important for the elderly and non-expert users.
- (2) We develop a novel LLM-powered framework to systematically generate noisy questions across various types, enabling controlled and scalable data generation.
- (3) We introduce a Self-Correction mechanism, a method that detects and corrects noisy questions to generate noise-free responses for transparent human-robot interaction.
- (4) We propose an innovative evaluation framework that specifically assesses agents' ability to handle noisy questions, comprehensively evaluating the agents' performance in real-world EQA tasks.

2. Related work

2.1. Embodied Question Answering

The EQA task aims to navigate a robot agent to an appropriate location in real-world scenarios and enable question-answering through physical interaction with humans. Existing methods for EQA can be divided into two categories: Traditional Neural Networks and those utilizing LLMs/Vision-Language Models (VLMs) [12].

Traditional Neural Network: Das *et al.* [1] first defined the EQA as a task requiring agents to navigate and answer questions in a real environment. To address this task, most early works employ traditional neural networks with relatively small scales. For example, Luo *et al.* [13] concentrated on label noise within the VQA and navigation modules of EQA to enhance system robustness. Conversely, our work is centered on addressing noisy questions in real-world scenarios.

LLMs and VLMs: Leveraging the strong reasoning and generalization capabilities of large-scale VLMs, EQA has witnessed significant advancements, particularly in complex reasoning, memory integration, and adaptive exploration capabilities [3,10]. Based on VLMs, more complex EQA tasks that are similar to real applications can be achieved. For instance, Open-EQA [3] focuses on EQA tasks with episodic memory (EM-EQA) and active exploration (A-EQA). To explore the real environment more efficiently, Ren *et al.* [10] leveraged an external semantic map to highlight regions worth investigating, optimizing the agents' exploration process by minimizing unnecessary actions.

2.2. Information inconsistency

Information inconsistency is a tricky problem faced by existing VLM. Hallucinations, where the generated language output does not align with the visual input [14,15], is one of the typical forms of information inconsistency. Numerous works are proposed to alleviate hallucinations in VLMs [16,17]. However, existing works predominantly focus on hallucinations generated by agents and overlook one of the most common types of information inconsistency: misalignment between user instructions and visual cues. Some researchers have tried to formulate this issue [18–21]. For example, Taioli *et al.* [19] introduced a novel benchmark for Vision-Language Navigation in Continuous Environments (VLN-CE) that simulates realistic instruction errors. Followed by a cross-modal transformer to detect these errors, the robustness against inaccurate human instructions is enhanced. However, a framework for systematically analyzing noise types in human input is absent. Our work fills this gap by defining a comprehensive taxonomy of noise types within human questions, comprehensively modeling, and solving the information inconsistency in EQA.

3. EQA with imperfect queries

3.1. Overview

When non-expert users, like the elderly, query about agents, they often unconsciously introduce various presumptions and implicit biases, which can impede the agents from generating accurate responses. To model these presumptions and implicit biases in practical scenarios, we define them as different types of “Noise” and taxonomize them into three main categories: Memory Noise, Perception Noise, and Semantic

Noise, which are illustrated in Figure 2. For each category, more detailed noisy questions regarding the “Attributes” of objects are further explored, including Position, Counting, Shape, Color, and Material. In total, we generate 500 noisy questions corresponding to three noise categories, forming a comprehensive and fine-grained benchmark.

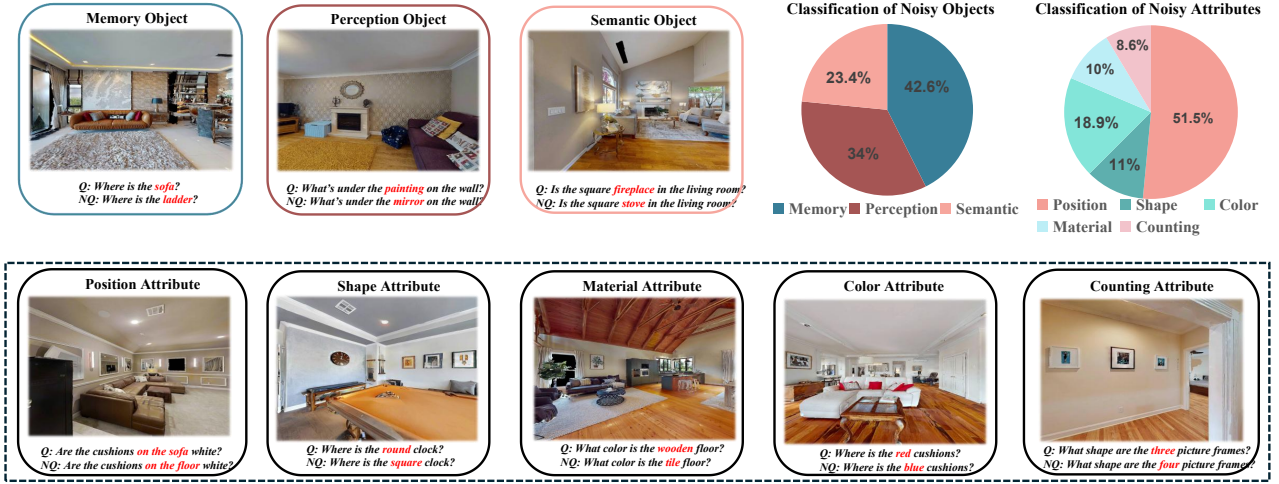


Figure 2. Noise types in NoisyEQA benchmark: Memory Noise, Perception Noise, and Semantic Noise. Attributes Noise is further subdivided into Position, Shape, Color, Counting, and Material.

3.2. Definition of noise types

In this subsection, drawing on the human error presented in [22,23], we elaborate on the definition of different types of noise in the NoisyEQA dataset.

Memory Noise arises from errors in recalling the existence or attributes of objects. For example, someone remembers seeing a red book on the table, but there was no book on the table. This involves both a misremembered object and its wrongly attributed color, blurring the line between memory and reality.

Perceptual Noise originates from inaccuracies in the perception of the environment by humans, simulating visual ambiguities during observation. For instance, a nearsighted person without glasses may misidentify the mirror as a picture in the scene. In this case, a question containing perceptual noise could be “What is the shape of the picture?”.

Semantic Noise is caused by incorrect substitutions of objects/attributes in a scene when humans attempt to describe the environment. Unlike perceptual noise, which involves the visual misidentification of objects, semantic noise contains logically feasible substitutions with semantically related but incorrect objects. Intuitively, in a living room, substituting “fireplace” with a semantically related object (e.g., “stove”) could introduce such noise.

3.3. Paradigms of question generation

Based on our definitions of different types of noise, we present a systematic framework powered by structured templates and LLMs to automatically generate noisy questions. Inspired by OpenEQA [3], we focus on generating open-ended question-answer pairs to better simulate the challenges of real-world

environments. We use both Wh questions and yes/no questions templates, such as “Where is the <OBJ>?” and “Is the <OBJ> {Attributes}?”, to generate noisy questions related to objects themselves (<OBJ>) or their attributes ({Attributes}). For noisy questions from the three sources of noise, the perturbations regarding attributes are similar, with the original correct attributes being replaced by incorrect ones.

For Memory Noise, we use non-existent objects <OBJ> to induce hallucinated responses from the model. Given a scene, we first establish a list of all the present objects and then introduce a non-existent object (with a random attribute) to form a noisy question. For example, if a living room contains a table, chair, and book, one of the questions related to memory noise about objects and their attributes could be “Where is the vase in the living room?” or “Is the vase in the living room white?” Although this question seems reasonable, it introduces a phantom object (vase), increasing the likelihood of fabricated responses.

To generate Perceptual Noise, we add Gaussian noise to the original scene image to simulate perceptual confusion. VLM is subsequently utilized to generate a description of the disturbed image, where some objects are incorrectly described due to distortion. We then construct questions with perceptual noise based on these misidentified objects and their attributes. For example, the sofa near the television is recognized as a bed, and a noisy question is formulated as “What is next to the bed?” or “Is the bed near the television made of wood?”.

Finally, we generate Semantic Noise by replacing objects in the scene with semantically related but incorrect ones. Specifically, objects are collected from OpenEQA [3] and Habitat-Matterport Embodied Question Answering(HM-EQA) [10] datasets. Bidirectional Encoder Representations from Transformers(BERT) is used to embed all candidate objects, and a pairwise similarity matrix is created to show semantic correlations. For each original object in a question, we pick the top 10 most similar objects as candidates. Then, a logically related object is selected from the subset by humans based on two principles: category overlap (e.g., “sofa” and “armchair”, which are both “furniture”) or functional association (e.g., “fireplace” and “stove”, which both provide heat). We substitute the original object with the logically related one to introduce semantic noise. For example, in a living room scene, a question with semantic noise related to the object and its attributes could be “Where is the stove?” or “Is the stove in the living room silver?” when the “stove” should be “fireplace” in reality.

After generating the dataset, two volunteers are recruited to perform a thorough validation. Volunteers meticulously examine each entry for inconsistencies, errors, and potential biases. In addition, they verify that the generated data accurately contains the intended noise types in various scenarios. Consequently, the two volunteers will annotate the ground-truth response for each noisy question.

4. Self-Correction mechanism

Once the presumptions and implicit biases are incorporated into noisy human questions, the accuracy of the responses generated by agents will significantly degrade. To address this issue, we propose a Self-Correction mechanism to enhance the robustness of agents against noisy questions. As shown in Figure 3a, for a scene with multiple views, the noisy question is marked by blue. Instead of directly answering the question, we first design an additional “confidence question” in yes/no format (which is marked by green) for the agent to answer. Taking the probability regarding the “yes” token as a

confidence score, the agent could identify the most reliable view of the scene. Then, the agent equipped with our Self-Correction mechanism will answer the noisy question based on the selected view. The Self-Correction mechanism has two following formats: Noise Aware Prompt (NAP) and Noise Aware Chain of Thought (NACoT) (Prompt details in Table A2).

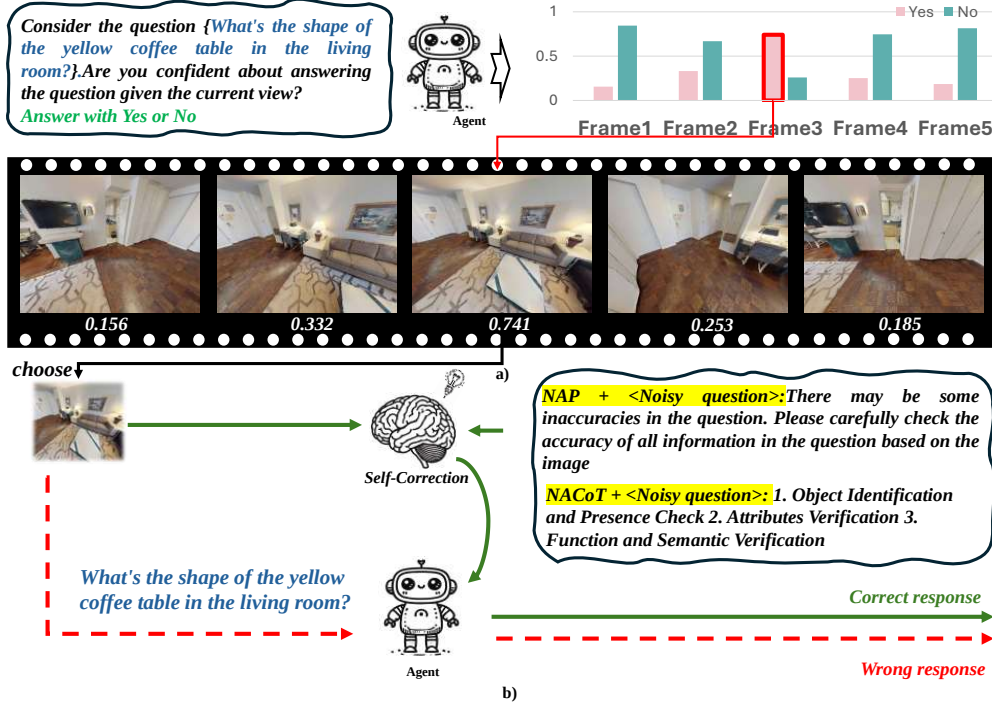


Figure 3. Overview of the reliable-view selection and Self-Correction mechanism. **(a)** Illustration of the reliable-view selection strategy for multi-frame visual inputs, where confidence scores are used to identify the most informative frame for answering the query; **(b)** Overview of the proposed Self-Correction mechanism, which leverages NAP and NACoT to verify object presence, attributes, and semantic consistency, reducing errors caused by noisy or ambiguous inputs.

4.1. Correcting prompt

NAP: NAP adopts a cognition prompt to enhance agents' sensitivity to noisy human questions. As shown in Figure 3b, a simple yet effective prompt is appended to the noisy questions, reminding the agent to carefully review the information related to the questions before responding. Instead of altering the reasoning process, NAP encourages the agent to validate the consistency between human questions and visual clues. The agent can therefore detect the noisy question more accurately, resulting in more reliable responses.

NACoT: Instead of simply prompting the agent to recognize the noise in human questions, we explore a more fine-grained mechanism, called the NACoT, to guide the agent in detecting and addressing noise step by step. Particularly, NACoT deconstructs a specific question into key components and systematically prompts the agent to verify whether each component is consistent with the visual content in the scene. As shown in Figure 3b, this process starts with object identification, ensuring that all objects mentioned in the questions are presented in the scene. Subsequently, the agent performs attribute verification to ensure the object attribute consistency between the questions and the real scene. Finally, the function and semantics verification is utilized to detect any semantic noise. Once any discrepancies are detected, the agent will

provide a clear explanation of the detected noise, and the expected responses that are not contaminated by noises.

4.2. Decision criteria and correction strategy

During evaluation, noise is considered detected if the generated response explicitly reports an inconsistency in any of the three verification stages. Specifically, detection is triggered when the model states that:

- an object mentioned in the question is not present in the selected view (object-level inconsistency);
- an attribute contradicts the visual evidence (attribute-level inconsistency); or
- the question contains a semantically related but incorrect substitution (semantic inconsistency).

If such inconsistency is identified, the model is instructed—within the same response—to:

- (1) explain the detected issue,
- (2) reformulate the corrected interpretation of the question, and
- (3) provide the final answer grounded in the corrected interpretation.

Otherwise, the agent directly answers the original question.

This design preserves computational efficiency through single-pass inference while introducing explicit verification signals in the output, thereby improving interpretability and enabling consistent measurement of detection and correction behaviors.

Overall, the complete inference workflow proceeds as follows: the embodied agent first selects the most informative view via confidence-based scoring, and then performs noise-aware reasoning over the selected observation using either NAP or NACoT. The agent verifies object presence, attribute consistency, and semantic correctness before committing to a final answer. If inconsistency is detected, it provides an explicit explanation and a corrected interpretation; otherwise, it directly responds to the original query. Notably, the entire procedure operates within a single forward pass and does not modify the underlying perception or navigation modules.

5. Experiments

In this section, we thoroughly evaluate the performance of different agents in addressing noisy questions.

5.1. Implementation details

Baseline Agents: Following Prismatic-VLMs, we adopt the mixed self-distillation with no labels (DINO)+SigLIP as vision encoder and replace the LLM with Llama-2 and GPT-4o to formulate our agents, respectively. In particular, the probability of the predicted tokens is necessary for our method to select the most reliable view from the scene. Therefore, LLMs whose probabilities are inaccessible are excluded in our experiments, such as Gemini, Claude, *etc.* Moreover, we follow the experimental setup of Explore-EQA, where the frame with the highest confidence is selected from multiple frames to answer questions.

Dataset and Evaluation Metric: We conduct all experiments on our NoisyEQA dataset, which contains 500 noisy questions. For evaluation, three metrics: Accuracy, Detection Rate, and Correction Rate, are adopted as described in Section 5.2. In addition, two volunteers are recruited to provide human responses to noisy questions, which enables the comparison between agents and human beings.

5.2. Evaluation criteria

As discussed in Section 1, existing metrics for question-answering systems are insufficient to evaluate the agents' performance in addressing noisy questions. Therefore, we propose a novel evaluation framework that considers both the quality of responses and the robustness against noises. The quality of the responses is evaluated on the basis of three aspects: noise detection, noise correction, and accurate generation. Specifically, responses are rated on a 1–5 scale, with detailed scoring criteria provided in the Appendix A3. Human assessment of these aspects is time-consuming and can introduce biases among different evaluators. Therefore, we design an evaluation scale based on all three aspects and employ a LLM for automatic evaluation. Based on the evaluation scale, for each noisy question q_i , let a_i^* and a_i represent the human-annotated ground truths and agent-generated responses, respectively. We then ask the LLM to score a_i based on the three aspects by comparing it with a_i^* . Following OpenEQA [3], the LLM-Match accuracy is adopted to evaluate the overall quality of all responses generated by an agent, which is formulated as follows:

$$C = \frac{1}{N} \sum_i^N \frac{\sigma_i - 1}{4} \times 100\%, \quad (1)$$

where N represents the number of noisy questions in NoisyEQA, and σ_i denotes the score for response a_i . In addition to the overall quality, two novel noise-related evaluation metrics—DR and CR—are introduced to evaluate the ability of an agent to handle noisy questions. Specifically, DR and CR are designed to quantitatively measure whether the agent can detect and correct different types of noise in questions, respectively. Formally, given the responses to all noisy questions $\mathbf{A} = \{a_i\}_{i=1}^N$, we denote the responses from the agent that successfully detect and correct the noise as $\mathbf{A}^d = \{a_i | \sigma_i \geq 3\}$ and $\mathbf{A}^c = \{a_i | \sigma_i \geq 4\}$, respectively. DR and CR are then formulated as:

$$\text{DR} = \frac{|\mathbf{A}^d|}{|\mathbf{A}|} \times 100\%, \quad \text{CR} = \frac{|\mathbf{A}^c|}{|\mathbf{A}|} \times 100\%. \quad (2)$$

Considering all these metrics, we comprehensively assess an agent's ability to generate accurate responses and handle the questions with significant noise and presumptions. Agents scoring high on all three metrics are more likely to achieve satisfactory performance in real-world EQA tasks.

5.3. Effectiveness of LLM-match

We follow the human alignment experiment in OpenEQA to validate the effectiveness of adopting LLM-Match approach for evaluation. Specifically, 150 questions are sampled from the NoisyEQA, and we let LLama2, GPT-4o, and real humans respectively answer 50 questions. Then, both the LLM and four human evaluators are asked to score these 150 responses using the evaluation scale in Section A3. without knowing their sources. The Spearman correlation coefficient between human and LLM-based evaluation scores was 0.87, validating the effectiveness of the LLM-Match method.

5.4. Evaluation results

The evaluation results for different types of noisy questions on LLaMA2, Qwen2.5-VL-7B and GPT-4o are summarized in Table 1, Table 2, and Table 3, respectively.

Table 1. Quantitative evaluation of LLaMA2 variants on the NoisyEQA-matched benchmark. NAP and NACoT significantly improve robustness to noisy questions, with NACoT achieving the best overall performance.

Method	Accuracy								Detection Rate	Correction Rate
	Memory Object	Perception Object	Semantic Object	Attributes						
				Position	Color	Shape	Counting	Material		
LLama2	6.08	12.86	22.45	16.00	15.91	18.55	12.07	20.00	3.0	2.8
LLama2+NAP	33.41	29.64	35.71	29.83	32.73	34.68	38.79	17.00	47.4	13.8
LLama2+NACoT	40.02	26.79	28.06	36.47	39.55	37.90	44.83	61.00	52.0	23.6
Humans	91.67	85.71	93.87	98.50	97.27	95.16	97.41	99.00	97.6	93.4

Table 2. Quantitative evaluation of Qwen2.5-VL-7B variants on the NoisyEQA-matched benchmark. NAP and NACoT significantly enhance robustness to noisy questions, with NACoT achieving the strongest overall performance among Qwen variants.

Method	Accuracy								Detection Rate	Correction Rate
	Memory Object	Perception Object	Semantic Object	Attributes						
				Position	Color	Shape	Counting	Material		
Qwen2.5-VL-7B	32.84	21.96	24.73	20.41	19.88	18.36	21.72	26.00	28.6	9.4
Qwen2.5-VL-7B+NAP	68.75	38.92	42.36	41.60	37.85	39.74	48.28	54.00	65.3	28.7
Qwen2.5-VL-7B+NACoT	74.30	41.40	47.85	43.15	41.92	44.63	52.41	58.00	69.5	31.4
Humans	91.67	85.71	93.87	98.50	97.27	95.16	97.41	99.00	97.6	93.4

Table 3. Quantitative evaluation of GPT-4o variants on the NoisyEQA-matched benchmark using the same metrics as Table 1. NACoT consistently improves semantic and attribute accuracy under noisy conditions.

Method	Accuracy								Detection Rate	Correction Rate
	Memory Object	Perception Object	Semantic Object	Attributes						
				Position	Color	Shape	Counting	Material		
GPT-4o	51.59	27.14	25.00	24.50	19.09	12.90	24.14	25.00	46.2	17.4
GPT-4o+NAP	86.98	43.57	46.42	53.50	40.45	41.94	57.76	64.00	79.0	42.0
GPT-4o+NACoT	87.60	45.00	51.02	39.96	45.91	49.19	62.07	63.00	77.6	35.0
Humans	91.67	85.71	93.87	98.50	97.27	95.16	97.41	99.00	97.6	93.4

Across both closed-source (GPT-4o) and open-source (LLaMA2, Qwen2.5-VL-7B) models, the baseline agents exhibit substantial performance degradation when noise is present in the questions. This suggests that presumptions and implicit biases in human queries pose systematic challenges even for agents with strong question-answering capabilities. These findings highlight that noisy user inputs significantly hinder the deployment of embodied agents in realistic EQA scenarios.

By incorporating NAP or NACoT for self-correction, all evaluated models demonstrate consistent performance improvements across memory, perceptual, semantic, and attribute noise types. For example, on Qwen2.5-VL-7B, the baseline DR and CR are 28.6% and 9.4%, respectively. Applying NAP substantially

improves robustness, while NACoT further increases DR to 69.5% and CR to 31.4%, indicating that structured verification effectively mitigates noise-induced degradation.

Although high-performance closed-source models such as GPT-4o achieve stronger absolute performance, the overall trend remains consistent across heterogeneous model families: noisy questions significantly impair baseline agents, and structured Self-Correction mechanisms systematically improve detection and correction performance.

Nevertheless, even the best-performing configuration (GPT-4+NACoT) still lags behind human performance, particularly under fine-grained attribute noise. Compared to object-level inconsistencies, detecting and correcting attribute-level errors requires more precise visual grounding and fine-grained relational reasoning. This observation suggests that attribute noise exposes a fundamental limitation in current vision-language reasoning capabilities, rather than a shortcoming of the self-correction framework itself.

Overall, these results demonstrate that the proposed mechanism generalizes across diverse model architectures and is not restricted to proprietary systems.

Furthermore, our NAP and NACoT improve agents' robustness to noises by significantly boosting the detection and correction rates of noisy questions. Additionally, although both agents achieve a relatively high detection rate, their correction rate remains much lower than that of humans. Correcting noisy questions requires triggering agents' reasoning capability at a more fine-grained level, we leave this as an open problem for future works. Last, we provide qualitative examples to show the effectiveness of our NAP and NACoT. As shown in Figure 4, the original LLama2 and GPT-4o fail to detect noise in the questions and provide hallucination responses, whereas our NAP and NACoT can effectively mitigate this issue.



Figure 4. Qualitative comparison of visual question answering results under noisy questions. Results show that NAP and NACoT more reliably identifies correct objects and attributes, and mitigates errors caused by misleading or noisy questions.

5.5. Generality across different EQA architectures

To further validate the plug-and-play applicability of the proposed Self-Correction mechanism, we evaluate NAP and NACoT on an additional embodied question answering agent (3D-Mem [24]) beyond Explore-EQA [10]. These agents adopt substantially different reasoning pipelines, including differences in perception backbones, memory integration strategies, and language–vision interaction mechanisms.

Importantly, the Self-Correction mechanism is integrated strictly as an inference-time wrapper without modifying perception, navigation, or memory modules. This ensures that the original architectural designs remain unchanged and allows us to isolate the effect of structured noise verification.

As reported in Table 4, both NAP and NACoT consistently improve Accuracy, DR, and CR across heterogeneous agents. While the magnitude of improvement depends on the base model’s reasoning strength, gains remain consistently positive under noisy conditions.

These results indicate that the effectiveness of the proposed self-correction strategy stems from structured consistency verification at the language–vision alignment level, rather than from architecture-specific design choices, demonstrating its architecture-agnostic and plug-and-play nature.

Table 4. Cross-architecture evaluation of the proposed Self-Correction mechanism on NoisyEQA. NAP and NACoT are integrated as inference-time wrappers without modifying the original pipelines. Consistent improvements in Accuracy, DR, and CR are observed across heterogeneous EQA agents.

Method	Accuracy	DR	CR
Explore-EQA	23.2	28.6	9.4
Explore-EQA + NAP	46.4	65.3	28.7
Explore-EQA + NACoT	50.5	69.5	31.4
3D-Mem	51.1	44.3	32.8
3D-Mem + NAP	57.7	66.8	44.1
3D-Mem + NACoT	62.4	72.6	51.9

5.6. Ablation on verification order

To evaluate the rationality of the fixed verification sequence in NACoT, we conduct an ablation study by perturbing the order of verification stages. Specifically, we compare the original order (object $\rightarrow$ attribute $\rightarrow$ semantic) with two alternative variants: attribute $\rightarrow$ object $\rightarrow$ semantic and semantic $\rightarrow$ object $\rightarrow$ attribute.

As shown in Table 5, both alternative orders lead to consistent degradation in Accuracy, DR, and CR. This confirms that object-level grounding functions as a structural prerequisite for subsequent attribute and semantic verification. Performing fine-grained checks before establishing correct object presence increases reasoning instability and amplifies downstream inconsistencies.

These results indicate that the proposed verification sequence reflects an underlying dependency hierarchy in noise handling rather than an arbitrary prompt formatting choice.

Table 5. Ablation study on verification order using Qwen2.5-VL-7B. Deviating from the original order leads to consistent performance degradation.

Verification Order	Accuracy	DR	CR
Object → Attribute → Semantic (Original)	50.9	69.5	31.4
Attribute → Object → Semantic	46.2	61.8	25.9
Semantic → Object → Attribute	43.9	58.6	23.4

5.7. Statistical robustness and bootstrap analysis

To evaluate whether the reported improvements are statistically reliable given the moderate dataset size (500 samples), we conduct a bootstrap-based robustness analysis. Specifically, we perform 10 resampling iterations with replacement over the full NoisyEQA benchmark. For each iteration, we compute Accuracy, DR, and CR, and report 95% percentile confidence intervals.

We adopt paired bootstrap resampling across methods to preserve sample-level correlations, ensuring that performance comparisons are made on identical resampled subsets. As shown in Table 6, the confidence intervals for NAP and NACoT consistently lie above those of the baseline. Importantly, the confidence intervals of performance differences do not overlap with zero, indicating statistically reliable improvements.

Furthermore, the magnitude of improvement (e.g., +27.0% in Accuracy and +40.9% in DR for NACoT over baseline) substantially exceeds the width of the corresponding confidence intervals, suggesting that the observed gains are not attributable to sampling variance.

We also observe consistent trends when conducting bootstrap analysis within individual noise categories, despite smaller subgroup sizes. These results collectively demonstrate that the performance improvements of the Self-Correction mechanism are statistically robust rather than due to random fluctuations.

Table 6. Bootstrap-based 95% confidence intervals for Qwen2.5-VL-7B variants computed over 10 resamples.

Method	Accuracy (95% CI)	DR (95% CI)	CR (95% CI)
Qwen2.5-VL-7B	23.9 [21.6, 26.2]	28.6 [25.9, 31.4]	9.4 [7.8, 11.1]
Qwen2.5-VL-7B + NAP	46.7 [44.1, 49.3]	65.3 [62.5, 68.0]	28.7 [25.9, 31.5]
Qwen2.5-VL-7B + NACoT	50.9 [48.2, 53.6]	69.5 [66.8, 72.1]	31.4 [28.5, 34.2]

5.8. Failure case analysis of correction

To better understand the gap between DR and CR, we perform a quantitative analysis of NACoT correction failures. We manually categorize failed correction cases into three primary types:

- Fine-grained attribute ambiguity (38%): Subtle distinctions in color, material, counting, or spatial relations that make corrected interpretation reconstruction uncertain. For example, differentiating between a light-gray table and a white table under varying lighting conditions may lead the model to conservatively avoid correction.

- Semantic proximity errors (33%): Confusion between semantically related objects despite successful inconsistency detection. For instance, after identifying that the queried object is incorrect, the model may hesitate between “sofa” and “armchair,” which share similar visual and categorical properties.
- Visual ambiguity or incomplete evidence (29%): Occlusion, limited viewpoints, or strong visual resemblance preventing reliable correction. A representative case involves confusion between a window and a framed painting on the wall, where reflective surfaces and rectangular framing produce visual similarity.

The results indicate that most correction failures arise from fine-grained discrimination challenges rather than failure to detect inconsistency. Notably, in a subset of cases the model correctly detects noise but conservatively abstains from proposing a corrected alternative due to insufficient visual certainty. While such cases are counted as non-correction under CR, they reflect safety-aware reasoning behavior.

This analysis explains the systematic gap between DR and CR: detection requires identifying inconsistency, whereas correction demands reconstructing a precise and confidently grounded alternative interpretation.

Although NACoT explicitly verifies spatial attributes, Position Noise often involves premise-level relational inconsistencies (e.g., incorrect room grounding). The structured decomposition in NACoT (object $\rightarrow$ attribute $\rightarrow$ semantic) first confirms object existence before verifying attributes.

When the object is correctly detected, the reasoning process may implicitly continue under the original spatial premise, especially if spatial cues are subtle. As a result, the model may proceed with answering rather than fully challenging the spatial assumption embedded in the question.

In contrast, NAP encourages holistic consistency checking without staged decomposition, which may occasionally make the model more sensitive to questioning incorrect spatial premises.

This behavior reflects the interaction between structured reasoning and premise-level spatial errors, rather than a limitation of the NACoT framework itself. It also suggests that future extensions could incorporate explicit premise validation steps to better handle relational spatial noise.

5.9. Limitations and future extensions

The current version of NoisyEQA assumes static environments, consistent with standard embodied question answering benchmarks. In real-world deployments, however, dynamic environmental changes—such as object relocation after a user forms a memory—may introduce additional memory–environment conflicts. Modeling such temporal inconsistencies would require explicit temporal reasoning, persistent state tracking, and environment updates across time steps. These challenges are orthogonal to the language-level noise studied in this work. Extending NoisyEQA to incorporate dynamic environment changes represents an important future direction toward more realistic embodied evaluation settings.

5.10. Computational cost analysis

We further analyze the computational overhead introduced by the proposed Self-Correction mechanism. All experiments are conducted using LLaMA-2 as the language backbone under identical hardware settings.

Running the full NoisyEQA benchmark (500 questions) requires approximately 25 hours and 51 minutes for the baseline agent, corresponding to an average of 186 seconds per question. Under the same setting, NAP requires approximately 191 seconds per question, while NACoT requires approximately 198 seconds per

question. This corresponds to a latency increase of approximately 2%–3% for NAP and 6%–7% for NACoT.

Table 7 summarizes the computational cost on 500 questions using a single NVIDIA RTX 3090 Graphics Processing Unit (GPU). NAP and NACoT slightly increase the processing time compared with the baseline, while the peak GPU memory remains stable at 16.3 GB. NACoT does not require additional model calls, indicating limited computational overhead. These results suggest that the proposed Self-Correction mechanism incurs a modest and bounded computational cost while providing substantial improvements in robustness under noisy queries.

Table 7. Detailed computational cost analysis of the proposed Self-Correction mechanism on a single NVIDIA RTX 3090 GPU.

Method	Total Time (500 Qs)	Average Time (s)	Peak GPU Memory (GB)	Additional Calls
Baseline	25h 51m	186	16.3	No
NAP	26h 31m	191	16.3	No
NACoT	27h 30m	198	16.3	No

6. Conclusion

This paper introduces the NoisyEQA benchmark, a dataset specially designed to evaluate robots’ ability to handle noisy human questions, particularly those from non-expert users such as the elderly. To mimic real-world noisy questions, we define Memory, Perception, and Semantic noise, and generate 500 noisy questions via a novel LLM-based framework. Evaluating baseline agents on NoisyEQA reveals their limitations when addressing noisy questions from these diverse user groups. A Self-Correction mechanism, presented in two formats, is proposed to assist agents in detecting and correcting noise before providing answers. To provide a more comprehensive evaluation of real-world agents, we propose two new metrics for noise detection and correction. This work paves the way for building reliable robots for real-world applications, and a more detailed, fine-grained noise benchmark is underway.

Data availability statement

The data or datasets generated or analyzed in this study are available at <https://aihabitat.org/datasets/hm3d/> and <https://github.com/gibrantaowu/NoisyEQA>.

Declaration of generative AI and AI-assisted technologies

The authors did not use generative AI or AI-assisted technologies in the writing of this manuscript.

Acknowledgments

This work is jointly supported by MOE Singapore Tier 1 Grant RG83/25, RS36/24 and a Start-up Grant from Nanyang Technological University.

Author’s contribution

Tao Wu: conceptualization, methodology, visualization, writing—original draft; Chuhao Zhou: formal analysis, visualization, writing—review and editing; Haozhi Cao: formal analysis, visualization, writing—review and editing; Yen Heng Wong: data curation; Lin Gu: investigation, formal analysis, visualization, writing—original draft and supervision; Jianfei Yang: investigation, resources, visualization, writing—review and editing and funding. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

Appendix

In this appendix, we first illustrate that NoisyEQA achieves AI transparency and accountability in Section A1. Then, more details of existing datasets and evaluation metrics are elaborated in Section A2 and Section A3, respectively. Eventually, we provide more experiment results and comprehensive analysis regarding to noise impacts and the effectiveness of our methods in Section A4.

A1. Transparency and accountability

A1.1. AI transparency in NoisyEQA

Transparency in AI systems is essential for reliable [11] and interpretable decision-making, especially in noisy scenarios addressed in NoisyEQA. In this work, we propose the Self-Correction mechanism to tackle the transparency gap in EQA agents. This mechanism enables agents to identify noise in user queries before generating corrected responses. By separating the detection and correction phases, the reasoning process of the system becomes explicit, allowing users to understand both the identification of noises and the rationale behind corrections. This enhances the transparency in agents and makes the responses more trustworthy and interpretable. In contrast, correcting without an explicit detection of noise makes the decision-making process obscure and harder for users to trust the corrected responses. Furthermore, our detailed evaluation framework, incorporating DR and CR metrics, offers a more transparent and fine-grained understanding of the agents’ performance. By separating noise detection and correction, our approach aligns with transparency principles. It enables users and developers to trace the reasoning process, identify potential vulnerabilities, and correct agents’ decisions in time.

A1.2. AI accountability in NoisyEQA

Accountability in AI systems is crucial to ensure that every phase of development, deployment, and operation are traceable to responsible actions [25]. In the context of NoisyEQA, the Self-Correction mechanism assures accountability by compelling agents to both detect noise and suggest actionable corrections. It guarantees that agents’ decisions are validated and can be scrutinized at every step, minimizing the risk of error propagation. Additionally, the Self-Correction mechanism can incorporate a human-in-the-loop processes, which allows humans to review and confirm critical decisions. In this

way, the agents completely meet regulatory standards and ethical requirements, ensuring the robust and trustworthy responses. The stepwise accountability, achieved through the combination of detection, explanation, and correction, promotes the development of responsible agents capable of effectively managing real-world noise.

A2. More details of existing datasets

As shown in Figure A1, we observed that in existing OpenEQA [3] and ExploreEQA [10] datasets, “Position” constitutes the largest proportion of all attributes in the questions. This aligns with the natural tendency of humans that we usually rely on positional information to identify different objects in a scene, especially when multiple similar objects are presented. Followed the observation, our dataset also treats “Position” as a primary attribute. By this way, our dataset can better simulate realistic interactions, where users frequently depend on location for precise objects identification, enhancing its applicability and usability in real-world scenario.

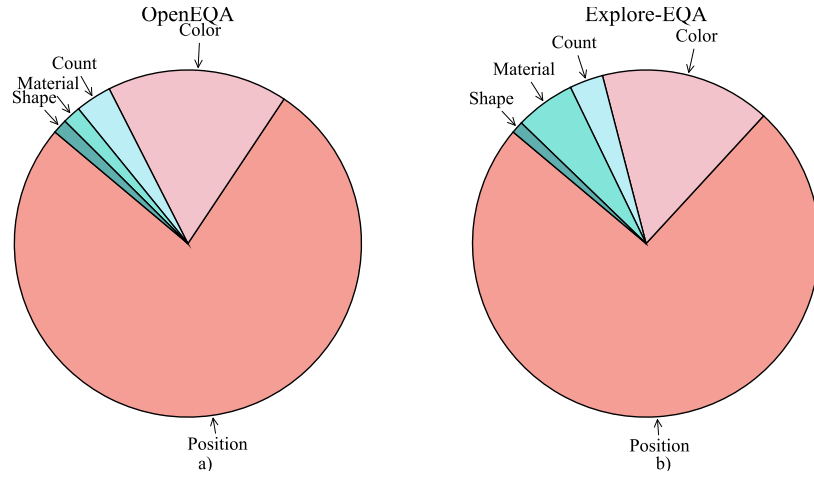


Figure A1. Distribution of different EQA noise types in the OpenEQA and Explore-EQA datasets. The pie charts illustrate the relative proportions of noise categories, including positional, perceptual, semantic, memory-related, and latent hallucination noise.

A3. Evaluation details

A3.1. Evaluation scale

To comprehensively evaluate the quality of responses from question-answering agents in the presence of noisy questions, we elaborately design a scoring scale ranging from 1 to 5. The evaluation scale assesses three key aspects: noise detection, noise correction, and accuracy of the generated response. The detailed criteria for each score are outlined as follows:

- (1) Score of 1: The question’s noise remains uncorrected, resulting in a response that is completely irrelevant or different from the correct answer. This indicates a fundamental misunderstanding for both the noise and the question’s core intent.
- (2) Score of 2: Although the noise in the question is not addressed or corrected, the response aligns with the correct answer in meaning. This suggests that the agent achieves the correct answer by coincidence, but fails to handle the noise in the question.

- (3) Score of 3: The agent recognizes the question is unanswerable due to excessive noise and responds with a statement such as “No, it’s not mentioned.” This demonstrates the agent is able to detect unanswerable questions, rather than forging an inaccurate or irrelevant answer.
- (4) Score of 4: The agent effectively corrects the noise in the question, but the response is only partially aligned with the correct answer. It shows that the agent fully understands the question but not achieves a completely correct answer.
- (5) Score of 5: The agent both corrects the noise in the question accurately and generates a response that perfectly matches the correct answer. It demonstrates that the agent has a full understanding of the question and can provide a high-quality answer.

A3.2. Effectiveness of LLM-Match

Inspired by OpenEQA [3], we adopt LLM-Match as an evaluation metric in our work. Specifically, the LLM-Match achieves an impressive Spearman correlation coefficient of 0.909 to human evaluations, highlighting its strong alignment with human judgment. Furthermore, OpenEQA shows that GPT-4 achieves superior performance in scoring tasks, with a correlation of 0.88 to human evaluations, significantly outperforming earlier models such as GPT-3.5 and LLaMA-2. These findings motivates us to leverage LLM-Match with GPT-4 in our evaluation framework.

A4. More experiment results

A4.1. Impact of noise on response confidence

Figure 1d in the manuscript shows that the noise in the question will significantly decrease the generation accuracy. Except for it, we further show that the impact of noise can deteriorate the response confidence as well. Recalling Figure 3a, the response confidence helps the agents to select the most reliable view to answer human questions. Therefore, the deterioration of response confidence will hinder the agents to choose the optimal view from the scene and indirectly make the responses worse. As presented in Figure A2, the agent consistently shows lower response confidence across all types of noisy question compared with the clean ones. Notably, more remarkable confidence drop is observed in categories of Latent Hallucination Noise and Memory Material Noise. We conjecture that it can be attributed to the essential of these tow noise types: they either introduce non-existent object or objects with totally incorrect attribute, which incorporates a greater impact on view selection. Conversely, Perception Noise or Semantic Noise separately involves visually similar or semantically related objects, which introduces milder impacts on view selection but needs more complex reasoning ability from agents to detect and correct. The facts revealed by Figure 1d and Figure A2 underline the dual impact of noise on both accuracy and response confidence, underscoring the necessity of our work to systematically investigate these noises and to develop noise-robust agents in real-world.

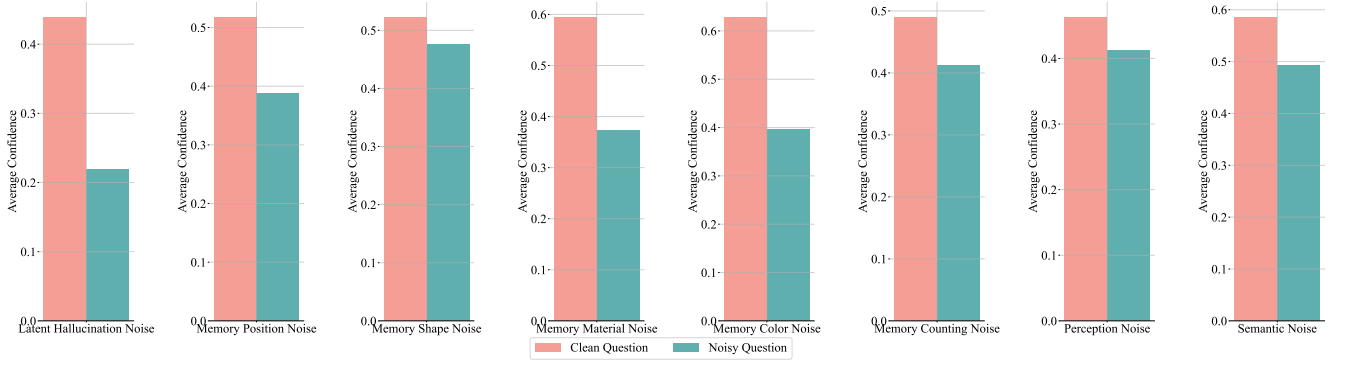


Figure A2. Impact of different EQA noise types on the agent’s answer confidence. The bar charts compare average confidence scores between clean and noisy questions across noise categories, showing a consistent confidence decline under noise.

A4.2. More analysis on noise detection and correction

We visualize DR and CR regarding to various types of noise in Figure A3 to intuitively show the effectiveness of our method. Besides, we conduct a more comprehensive analysis on the difference in effectiveness of our method when dealing with different noises.

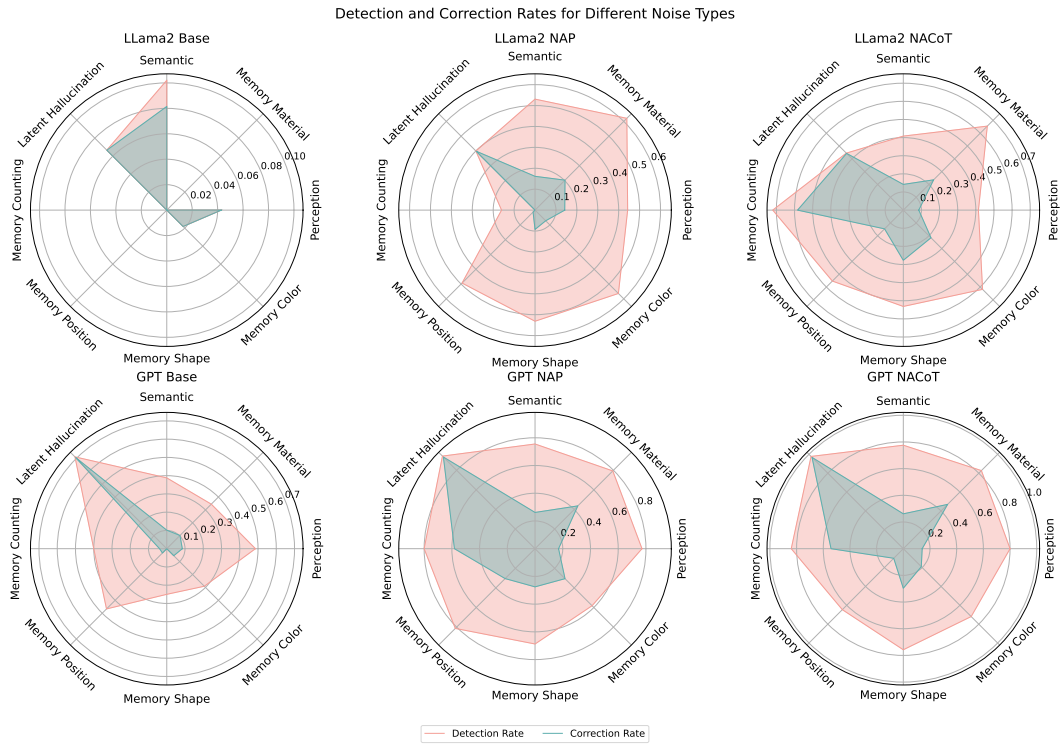


Figure A3. Detection and correction rates under different EQA noise types. NACoT improves performance for latent hallucination and memory-related noise, whereas perception and semantic noise remain difficult to correct despite successful detection.

Specifically, our method demonstrates significant effectiveness in handling Latent Hallucination Noise. As shown in Figure A3, the introduce of NAP and NACoT significantly enhances detection capabilities of both agents. Besides, as long as the noise can be detected, the model can successfully correct it. This indicates that the challenge in handling latent hallucination noise primarily lined in detection and it can be easily corrected once identified.

In addition, for memory noise regarding to Material and Shape, the outstanding improvements are achieved in both DR and CR. The Self-Correction mechanism allows the agent to identify any misalignment between visual contents and noisy questions by re-checking. It shows that this mechanism is effective for material-related and shape-related memory noise, since they always introduce relatively straightforward difference between the scene and human questions, which can be handled by agents without complex reasoning.

However, for Perception Noise and Semantic noise, the correction rate remains low even when the noise is detected. Particularly, these two noises incorporate either visually or semantically similar objects to replace the original ones. In contrast to non-existent objects or objects with significant differences, we conjecture that the agent requires more comprehensive and complex reasoning ability to correct such fine-grained factual errors. This underscores that noises related to subtle visual or semantic changes remain challenging, suggesting a promising direction for future research to develop more advanced mechanisms that are capable of addressing these nuanced distinctions.

A4.3. Comparison of NACoT with generic structured prompting paradigms

Although NAP and NACoT are implemented as prompting strategies, their contribution is not merely an instance of generic prompt engineering. Standard reasoning prompts, such as Chain-of-Thought (CoT), generally assume that the input question is valid and focus on improving answer derivation through intermediate reasoning steps. In contrast, our setting explicitly considers the possibility that the question itself may contain inconsistencies with the observed environment.

More importantly, the proposed NACoT is not simply a longer or more structured prompt. Its key design is a taxonomy-aligned verification process that decomposes reasoning into three explicit stages: object presence checking, attribute consistency verification, and semantic consistency verification. These stages are directly grounded in the noise taxonomy defined in NoisyEQA, namely Memory Noise, Perception Noise, and Semantic Noise. As a result, NACoT guides the model to first verify whether the query premise is valid before committing to answer generation, rather than reasoning directly from the original question.

To further distinguish this design from generic structured prompting, we additionally compare NACoT with standard CoT and a generic verification prompt under the same model and evaluation setting. As shown in Table A1, while both generic structured reasoning and generic verification provide modest improvements over the baseline, NACoT consistently achieves substantially higher detection and correction performance. This indicates that the gain does not come from structured prompting alone, but from the proposed noise-aware and taxonomy-aligned verification logic.

Therefore, the contribution of NACoT lies not in prompting format itself, but in formalizing noisy-question handling as a structured inference procedure tailored to embodied question answering under imperfect user queries.

Table A1. Comparison of NACoT with generic structured prompting paradigms on Qwen2.5-VL-7B under identical settings. NACoT achieves substantially higher DR and CR than standard CoT and generic verification prompting.

Prompting Paradigm	Detection Rate (DR) ↑	Correction Rate (CR) ↑
Qwen2.5-VL-7B	28.6	9.4
+ Standard CoT [26]	34.2	12.8
+ Generic Verification Prompt [27]	41.5	18.2
+ NACoT (Ours)	69.5	31.4

A5. Prompting template

Table A2. Prompting templates for NAP and NACoT.

NAP

The question may contain inaccurate or noisy information.

Before answering, verify that all mentioned objects and attributes are consistent with the visible scene.

If inconsistencies are found, explain what is incorrect and why, then provide a corrected response grounded strictly in visual evidence.

If no noise is detected, answer based only on visible information.

NACoT

Question: {question}

Please analyze the image(s) step by step and determine whether the question contains noise.

Step 1: Object Identification and Presence Check

- List all key objects explicitly mentioned in the question.
- Verify whether each object is clearly visible in the scene.
- If an object does not exist in the scene, mark it as **Object-Level Noise**.

Step 2: Attribute Verification

- For each mentioned object, check whether its attributes (position, color, material, shape, counting) match the visual evidence.
- If any attribute conflicts with what is visible, mark it as **Attribute-Level Noise**.

Step 3: Semantic and Functional Consistency Check

- Determine whether the object name could be a semantically related but incorrect substitution (e.g., “stove” vs. “fireplace”).
- If the mentioned object is visually similar but semantically inconsistent with the scene context, mark it as **Semantic Noise**.

Decision and Response Rules

- If no noise is detected, provide a concise answer grounded in visible evidence.
- If noise is detected, clearly explain:
 1. What part of the question is incorrect.
 2. Why it is inconsistent with the scene.
 3. Provide the corrected, noise-free answer.
- Do not hallucinate missing objects or attributes.
- Use only visible evidence from the provided image(s).

A6. Discussion on reliability and variability

Confidence-based view selection does not guarantee that the chosen frame always contains all critical visual cues required to answer a query. When relevant evidence is absent from the selected view, NACoT may conservatively detect object absence or attribute inconsistency during structured verification. Although this may occasionally stem from incomplete observation rather than genuine inconsistency, such conservative behavior mitigates hallucinated responses and supports the safety-oriented objective of NoisyEQA.

Furthermore, different models may produce different confidence distributions and therefore select different frames. Rather than enforcing deterministic frame selection, we preserve this variability to reflect realistic embodied deployment settings. Since all evaluation metrics are computed in an end-to-end manner, the reported accuracy, detection rate, and correction rate already incorporate the downstream impact of view-selection differences on reasoning and prediction.

A7. Noise distribution design

To examine whether the adopted noise proportions are broadly consistent with realistic user behavior, we conducted a small-scale user study with non-expert participants in indoor object-finding and navigation scenarios. Specifically, 50 participants were recruited, and a total of 500 natural-language queries were collected. Participants were shown short video sequences sampled from indoor environments in Habitat and Matterport3D, covering common household spaces such as living rooms, bedrooms, kitchens, and dining areas. They were asked to freely describe or query objects in the scene, including object identity, location, and attributes, without being informed of our predefined noise taxonomy. This setting was designed to elicit natural imperfect queries rather than taxonomy-aware examples.

After collection, all queries were manually categorized according to the three noise types defined in NoisyEQA, namely Memory Noise, Perception Noise, and Semantic Noise. To improve labeling reliability, two annotators independently assigned noise labels to each query. The inter-annotator agreement was 87% in percentage agreement, and all disagreements were resolved through discussion to reach a final consensus label.

Based on these observations, we use the study as a small-scale empirical reference for benchmark design rather than as a definitive estimate of real-world frequency statistics. The final NoisyEQA distribution is therefore constructed as a controlled and diagnostically balanced benchmark, informed by the observed tendencies in user queries while maintaining sufficient coverage across noise categories.

Among the collected noisy queries, memory-related noise accounted for approximately 40%, perceptual confusions 30%, and semantic substitutions 30%. These results suggest that memory inconsistencies constitute a substantial portion of real user noise, while perceptual and semantic confusions also occur with comparable frequency.

Based on these observations, the final NoisyEQA distribution is designed as a controlled benchmark rather than a strict replication of real-world frequency statistics. Specifically, within object-level perturbations, Memory, Perception, and Semantic noise account for 42.6%, 34.0%, and 23.4%, respectively.

At the attribute level, perturbations are distributed across position (51.5%), color (18.9%), shape (11.0%), material (10.0%), and counting (8.6%), reflecting the predominance of positional references in human descriptions.

This hierarchical normalization ensures diagnostic diversity and evaluation stability, while avoiding strong imbalance across noise categories. We believe that large-scale corpus studies could further refine real-world distribution estimates and consider this an important direction for future work.

A8. Threshold definition rationale

Detection and correction are defined using $\sigma \geq 3$ and $\sigma \geq 4$, respectively, based on semantically meaningful score boundaries in the evaluation rubric. The scoring scale is behaviorally structured: a score of 3 corresponds to successful noise detection, where the agent identifies an inconsistency and refrains from directly answering the noisy query, while a score of 4 corresponds to successful noise correction, where the agent not only detects the inconsistency but also provides a corrected interpretation and answer.

These thresholds therefore correspond to qualitatively distinct reasoning stages, rather than arbitrary numerical cutoffs. To further verify that our conclusions are not sensitive to this design choice, we additionally conduct a threshold sensitivity analysis in the revised appendix by varying the score cutoffs for Detection Rate and Correction Rate. As shown in Table A3, although the absolute values of DR and CR vary with stricter or looser thresholds, the relative performance trends remain unchanged across settings, with both NAP and NACoT consistently outperforming the baseline and NACoT remaining the strongest variant. This confirms that the adopted thresholds are both semantically interpretable and empirically robust.

Table A3. Sensitivity analysis of DR and CR under different score thresholds σ on Qwen2.5-VL-7B. Although absolute values vary with threshold strictness, the relative ordering remains unchanged across settings.

Method	DR $\uparrow$			CR $\uparrow$		
	$\sigma \geq 2$	$\sigma \geq 3$	$\sigma \geq 4$	$\sigma \geq 3$	$\sigma \geq 4$	$\sigma = 5$
Qwen2.5-VL-7B	40.2	28.6	19.4	15.1	9.4	4.2
+ NAP (Ours)	72.5	65.3	54.1	38.2	28.7	19.5
+ NACoT (Ours)	79.2	69.5	60.8	40.4	31.4	22.6

References

- [1] Das A, Datta S, Gkioxari G, Lee S, Parikh D, *et al.* Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, USA, June 18–22, 2018, pp. 1–10.
- [2] Sakamoto K, Azuma D, Miyanishi T, Kurita S, Kawanabe M. Map-based modular approach for zero-shot embodied question answering. In *Proceedings of 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Abu Dhabi, UAE, October 14–18, 2024, pp. 10013–10019.
- [3] Majumdar A, Ajay A, Zhang X, Putta P, Yenamandra S, *et al.* OpenEQA: embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, June 16–22, 2024, pp. 16488–16498.

- [4] Antol S, Agrawal A, Lu J, Mitchell M, Batra D, *et al.* VQA: visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, December 7–13, 2015, pp. 2425–2433.
- [5] Wang X, Ma W, Li Z, Kortylewski A, Yuille AL. 3D-aware visual question answering about parts, poses and occlusions. In *Proceedings of the Advances in Neural Information Processing Systems*, New Orleans, USA, December 10–16, 2023, pp. 58717–58735.
- [6] Loftus EF. *Eyewitness testimony*, 2nd ed. Cambridge: Harvard University Press, 1996.
- [7] Loftus EF, Palmer JC. Reconstruction of automobile destruction: an example of the interaction between language and memory. *J. Verbal. Learn. Verbal. Behav.* 1974, 13(5):585–589.
- [8] Lubow R, Rifkin B, Alek M. The context effect: the relationship between stimulus preexposure and environmental preexposure determines subsequent learning. *J. Exp. Psychol. Anim. Behav. Process.* 1976, 2(1):38.
- [9] Salam N, Hamid NA, Rahman S, Nor MFM, Hassim MH. A segment study of the elderly ageing population: Malaysia, Bangladesh, Qatar, Singapore and Japan. In *Proceedings of International Conference On Syariah & Law*, online, April 6, 2021, pp. 89–99.
- [10] Ren AZ, Clark J, Dixit A, Itkina M, Majumdar A, *et al.* Explore until confident: efficient exploration for Embodied Question Answering. *arXiv* 2024, arXiv:2403.15941.
- [11] Edwards L. The EU AI Act: a summary of its significance and scope. 2021. Available: <https://www.adalovelaceinstitute.org/wp-content/uploads/2022/04/Expert-explainer-The-EU-AI-Act-11-April-2022.pdf> (accessed on 16 May 2026).
- [12] Liu Y, Chen W, Bai Y, Luo J, Song X, *et al.* Aligning cyber space with physical world: a comprehensive survey on embodied AI. *arXiv* 2024, arXiv:2407.06886.
- [13] Luo H, Lin G, Shen F, Huang X, Yao Y, *et al.* Robust-EQA: robust learning for embodied question answering with noisy labels. *IEEE Trans. Neural Networks Learn. Syst.* 2023, 35(9):12083–12094.
- [14] Guan T, Liu F, Wu X, Xian R, Li Z, *et al.* HallusionBench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, June 16–22, 2024, pp. 14375–14385.
- [15] Liu H, Xue W, Chen Y, Chen D, Zhao X, *et al.* A survey on hallucination in large vision-language models. *arXiv* 2024, arXiv:2402.00253.
- [16] Manas K, Zwicklbauer S, Paschke A. CoT-TL: low-resource temporal knowledge representation of planning instructions using chain-of-thought reasoning. In *Proceedings of 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Abu Dhabi, UAE, October 14–18, 2024, pp. 9636–9643.
- [17] Hao G, Wu J, Pan Q, Morello R. Quantifying the uncertainty of LLM hallucination spreading in complex adaptive social networks. *Sci. Rep.* 2024, 14(1):16375.
- [18] Liu A, Huang T, Liu X, Xu Y, Ma Y, *et al.* Adversarial attacks for embodied agents. *arXiv* 2020, arXiv:2005.09161.

- [19] Taioli F, Rosa S, Castellini A, Natale L, Del Bue A, *et al.* Mind the Error! Detection and localization of instruction errors in vision-and-language navigation. In *Proceedings of 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Abu Dhabi, UAE, October 14–18, 2024, pp. 12993–13000.
- [20] Gopinathan M, Abu-Khalaf J, Suter D, Masek M. StratXplore: strategic novelty-seeking and instruction-aligned exploration for vision and language navigation. In *Proceedings of 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Abu Dhabi, UAE, October 14–18, 2024, pp. 12093–12100.
- [21] Islam CM, Salman S, Shams M, Liu X, Kumar P. Malicious path manipulations via exploitation of representation vulnerabilities of vision-language navigation systems. In *Proceedings of 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Abu Dhabi, UAE, October 14–18, 2024, pp. 13845–13852.
- [22] Russell S, Moskowitz IS, Raglin A. Human information interaction, artificial intelligence, and errors. In *Autonomy and Artificial Intelligence: A Threat or Savior?*, 1st ed. Cham: Springer, 2017. pp. 71–101.
- [23] Norman DA. Design rules based on analyses of human error. *Commun. ACM* 1983, 26(4):254–258.
- [24] Yang Y, Yang H, Zhou J, Chen P, Zhang H, *et al.* 3D-mem: 3D scene memory for embodied exploration and reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, Nashville, USA, June 11–15, 2025, pp. 17294–17303.
- [25] Verde C, Guinea E, Macedonia N, Leone S, Rica C, *et al.* Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development: draft resolution/Albania, Argentina, Australia, Bahamas, Belgium, Brazil, Bulgaria, Cabo Verde, Canada, Chile, Côte d’Ivoire, Croatia, Czechia, Denmark, Djibouti, Dominican Republic, Equatorial Guinea, Estonia, Fiji, Finland, France, Georgia, Germany, Greece, Ireland, Israel, Italy, Japan, Jordan, Kenya, Liberia, Luxembourg, Maldives, Montenegro, Morocco, Netherlands (Kingdom of the), New Zealand, North Macedonia, Peru, Republic of Korea, Romania, Serbia, Sierra Leone, Singapore, Slovakia, Slovenia, South Africa, Spain, Sweden, Türkiye, United Arab Emirates, United Kingdom of Great Britain and Northern Ireland, United States of America, Uzbekistan and Zambia. 2024. Available: <https://digitallibrary.un.org/record/4040897?v=pdf> (accessed on 16 May 2026).
- [26] Wei J, Wang X, Schuurmans D, Bosma M, Xia F, *et al.* Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of Advances in neural information processing systems*, New Orleans, USA, November 28–December 9, 2022, pp. 24824–24837.
- [27] Weng Y, Zhu M, Xia F, Li B, He S, *et al.* Large language models are better reasoners with self-verification. In *Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, December 6–10, 2023, pp. 2550–2575.