VAR: Benchmarking Active Reasoning with Noisy Visual Feedback

Anonymous ACL submission

Abstract

Real-world reasoning rarely reduces to static question answering: agents must actively gather information from tools and sensors that are often noisy and incorrect. However, most existing active reasoning benchmarks either focus on environments where feedback is largely reliable or inject noise without providing an explicit, calibrated uncertainty signal about tool outputs, making it difficult to analyze how LLMs should reason with uncertain evidence. We introduce VAR, a novel benchmark for active reasoning under noisy visual feedback that is explicitly designed to evaluate text-only LLM reasoners: a fixed, off-the-shelf VLM is treated as a stochastic visual sensor, and the LLM must solve VQA problems solely by querying this sensor. For each sensor query, we draw multiple samples and expose a coarse uncertainty signal via self-consistency, enabling the reasoner to probe from different angles and decide what to ask next and when to stop. Our construction is automatic and scalable: starting from diverse VQA sources and two modern VLMs, we select instances where the sensor is inconsistent yet human-solvable. VAR thus provides a controlled playground to study how different LLMs exploit uncertainty signals for robust reasoning.

1 Introduction

Humans typically solve problems by *actively* gathering information from tools and sensors that are noisy and sometimes wrong: we form expectations about the world, compare new observations against these beliefs, and cross-check conflicting evidence before committing to a decision. Recent work on LLM agents and active reasoning begins to move in this direction, letting models plan tool calls, ask follow-up questions, or navigate simulators instead of answering in a single shot (Shridhar et al., 2020; Nakano et al., 2021; Yao et al., 2022a; Abdulhai et al., 2023; Li et al., 2024; Hu et al., 2024; Zhou et al., 2025). However, these settings usually assume that the environment behaves as an oracle, or they provide no explicit and reliable signal about the uncertainty of tool outputs, leaving open how LLMs should act when feedback itself is unreliable. People, in contrast, rarely trust a single observation at face value, especially when they are explicitly asked to double-check or when the stakes are high. For instance, when a navigation app suggests an exit that conflicts with road signs, we slow down and cross-check; when a medical monitor shows an abnormal reading that contradicts clinical signs, doctors re-measure or order a different test. A similar pattern arises in vision: consider the image in Figure 1 and the question *"Where is the woman standing with respect to the man?"* A visual-language model repeatedly answers that the woman is *behind* the man, yet also describes them as walking *arm-in-arm*. Even without seeing the image, a human reasoner can detect this inconsistency, realize that walking arm-in-arm while one person is behind the other is unlikely, and infer that they are most plausibly standing next to each

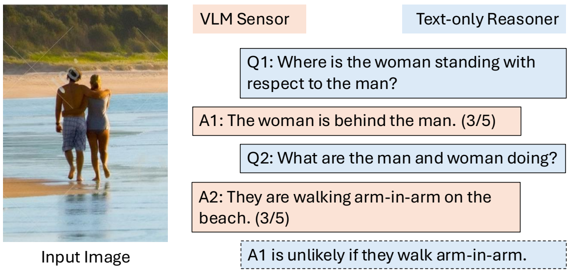

Figure 1: **Active reasoning under noisy visual feedback**. A text-only reasoner (right) receives VLM outputs (left) with consistency scores. The reasoner identifies inconsistencies (conflict between "standing behind" and "walking arm-in-arm") and resolves them by asking auxiliary questions. The VAR benchmark uses this setup to evaluate how well LLMs reason and query to handle uncertain or inconsistent evidence.

other. This kind of belief-driven cross-validation is largely missing from current LLM benchmarks.

In this paper, we study how *text-only* LLMs perform active reasoning when their only access to an image is uncertain feedback from a fixed VLM sensor. Rather than pushing the frontier of visual recognition, we treat an off-the-shelf VLM as a black-box noisy sensor and use its self-consistency across samples as a cheap, scalable, and externally verifiable uncertainty signal. Building on prior "blind" VQA setups (Li et al., 2025), the LLM never sees the image: it must solve each instance by deciding what perception questions to ask, how to aggregate the answers, and when to trust or override them. For each sensor query we draw multiple VLM samples and expose a coarse consistency score, enabling the reasoner to probe from different angles, maintain an internal belief state, and recover from misleading feedback. To study this systematically, we introduce VAR, a benchmark and analysis suite for active reasoning under noisy visual feedback. VAR is constructed automatically from diverse VQA sources and two modern VLMs with distinct failure modes, selecting instances where the sensors are self-inconsistent yet "blind" human annotators can still solve the task. We use VAR to evaluate a range of LLM reasoners and interaction strategies, finding that all models still lag far behind humans, but strong reasoners reliably improve over direct VLM predictions and leverage consistency signals to reduce miscalibration, especially on perception and chart-understanding tasks. These results highlight both the promise and current limitations of uncertainty-aware reasoning with noisy sensors.

2 Related Work

Our work connects to modular visual reasoning frameworks that separate perception from reasoning, to LLM-based agents for active information seeking and tool use, and to RAG systems that use external knowledge sources for question answering. Due to space limits, we provide an extended discussion of these connections in the App. A.

3 VAR Construction

3.1 Problem Setup: Active Reasoning with a Noisy Visual Sensor

We formulate VQA as an active reasoning problem for a text-only LLM reasoner R interacting with a noisy visual sensor S. Each instance consists of an image x, an original VQA question $q^{(0)}$, and a ground-truth answer y. The sensor S is *stateless*: every call $S(x, q)$ depends only on (x, q). At interaction step t, the reasoner observes the interaction history

$$h^{(t)} = \big(q^{(0)},\, (q^{(1)}, a^{(1)}, u^{(1)}),\, \ldots,\\ (q^{(t-1)}, a^{(t-1)}, u^{(t-1)})\big), \qquad (1)$$

and chooses an action of concluding with an answer or issuing a self-contained query $q^{(t)}$ to the sensor. To model noisy perception and expose uncertainty, we sample sensor K times with the same prompt:

$$\{a_1^{(t)}, \ldots, a_K^{(t)}\} \sim S(x, q^{(t)}). \qquad (2)$$

We then (i) sample one reply $a^{(t)}$ uniformly from $\{a_i^{(t)}\}_{i=1}^K$ as the realized sensory feedback, and (ii) estimate an uncertainty score by computing the empirical consistency of this answer:

$$u^{(t)} = \sum_{i=1}^{K} \mathbf{1}\big[\text{sem\_eq}(a_i^{(t)}, a^{(t)})\big], \qquad (3)$$

where $\text{sem\_eq}(\cdot, \cdot)$ is implemented by an LLM-based aggregator that judges whether two responses are semantically equivalent. The pair $(a^{(t)}, u^{(t)})$ is appended to the history to form $h^{(t+1)}$. The interaction loop continues until R outputs a parsable answer or the step budget T is reached.

3.2 Dataset Creation

Dataset Sources. We construct our benchmark from multiple-choice VQA datasets that collectively cover diverse visual reasoning abilities. Concretely, we include three broad categories: perception on natural images, using SeedBench (Li et al., 2023) for everyday scene understanding; chart understanding, using FigureQA (Kahou et al., 2017); and knowledge-based VQA, using MMStar (Chen et al., 2024), CVQA (Romero et al., 2024), and ReasonVQA (Tran et al., 2025) to capture reasoning settings that require discipline-specific, cultural, and encyclopedic knowledge.

VLM Sensors. We consider two modern VLM sensors, Qwen2.5-VL-7B and InternVL3.5-8B, chosen because preliminary experiments indicate that they exhibit complementary behaviors and failure modes. We report the VLM behavior results in App. H and App. I.

Two-stage Filtering. To ensure that the reasoning tasks are both challenging and solvable, we

		InternVL3.5-8B						Qwen2.5-VL-7B					
		Perception		Charts		Knowledge		Perception		Charts		Knowledge	
Reasoner	Strategy	Acc ↑	ECE ↓	Acc ↑	ECE ↓	Acc ↑	ECE ↓	Acc ↑	ECE ↓	Acc ↑	ECE ↓	Acc ↑	ECE ↓
Human	-	70.00	-	95.00	-	75.00	-	80.00	-	100.00	-	81.67	-
GPT-4o	BP	**46.00**	-	56.00	-	**57.33**	-	46.00	-	**58.00**	-	**60.67**	-
	UoT	22.00	-	**58.00**	-	49.33	-	42.00	-	52.00	-	**60.67**	-
	ReAct	34.00	-	48.00	-	44.00	-	46.00	-	56.00	-	60.00	-
GPT-oss-20B	BP	38.00 ±9.59	0.32	43.60 ±3.29	0.41	48.27 ±2.24	0.24	**51.60** ±7.80	0.13	32.00 ±5.66	0.56	57.33 ±4.32	0.14
	UoT	35.20 ±5.93	0.21	33.60 ±7.92	0.49	45.60 ±3.50	0.13	44.80 ±6.72	0.20	26.00 ±1.41	0.64	49.07 ±3.15	0.13
	ReAct	32.40 ±6.39	0.39	38.40 ±4.10	0.48	46.80 ±2.59	0.24	44.00 ±3.16	0.34	43.60 ±5.90	0.38	54.13 ±3.70	0.18
Qwen2.5-72B	BP	42.00 ±3.74	0.35	50.00 ±8.94	0.36	44.80 ±2.86	0.21	50.00 ±8.25	0.15	45.20 ±4.82	0.40	51.87 ±3.31	0.19
	UoT	37.60 ±9.10	0.40	38.80 ±5.76	0.41	40.00 ±2.63	0.26	44.40 ±5.73	0.30	39.60 ±5.18	0.40	50.80 ±3.47	0.16
	ReAct	38.00 ±3.74	0.42	42.40 ±7.13	0.38	44.27 ±1.86	0.27	43.60 ±2.61	0.40	48.40 ±6.54	0.33	50.13 ±3.27	0.26
Qwen3-32B	BP	32.80 ±9.65	0.24	43.60 ±6.07	0.16	34.93 ±3.88	0.18	36.40 ±6.23	0.14	39.20 ±5.93	0.19	39.33 ±3.72	0.17
	UoT	31.20 ±3.35	0.27	38.40 ±4.34	0.32	32.00 ±2.23	0.13	28.40 ±7.27	0.18	41.20 ±3.63	0.17	36.80 ±3.34	0.21
	ReAct	32.00 ±7.75	0.32	42.00 ±6.32	0.37	38.27 ±3.97	0.13	42.80 ±6.57	0.23	45.20 ±6.42	0.29	46.80 ±3.42	0.15
Qwen3-8B	BP	27.60 ±3.58	0.25	31.20 ±4.82	0.31	24.53 ±2.88	0.16	28.80 ±1.79	0.22	28.00 ±4.69	0.37	32.00 ±3.33	0.13
	UoT	30.80 ±4.60	0.23	24.40 ±6.23	0.31	24.93 ±4.08	0.16	30.40 ±5.55	0.24	24.80 ±2.68	0.40	28.40 ±2.84	0.25
	ReAct	28.00 ±3.74	0.38	32.40 ±1.67	0.45	28.27 ±3.49	0.21	40.00 ±4.69	0.23	28.40 ±4.98	0.55	36.00 ±3.43	0.23
Qwen2.5-7B	BP	30.80 ±5.02	0.43	36.00 ±10.30	0.20	31.33 ±3.36	0.31	38.80 ±6.26	0.23	28.00 ±4.00	0.32	36.53 ±3.23	0.18
	UoT	31.20 ±8.44	0.31	25.20 ±5.40	0.34	31.33 ±2.97	0.26	38.40 ±5.90	0.30	26.80 ±5.76	0.42	35.47 ±2.69	0.17
	ReAct	29.20 ±5.93	0.42	36.80 ±3.90	0.48	33.07 ±2.85	0.29	44.80 ±8.32	0.29	25.60 ±5.55	0.64	40.53 ±2.87	0.26
LLama3.1-8B	BP	26.40 ±9.21	0.34	41.20 ±3.90	0.29	32.13 ±4.60	0.18	39.60 ±5.55	0.27	38.80 ±3.63	0.32	37.47 ±3.05	0.17
	UoT	30.80 ±5.40	0.23	35.60 ±6.99	0.39	34.53 ±2.85	0.13	37.60 ±2.61	0.20	36.00 ±6.93	0.35	28.40 ±4.48	0.26
	ReAct	30.40 ±2.61	0.42	51.60 ±6.39	0.28	34.93 ±4.46	0.16	42.80 ±2.28	0.21	**58.00** ±3.74	0.11	34.53 ±2.69	0.23
VLM	-	27.09 ±6.02	0.48	32.18 ±4.77	0.54	27.03 ±2.95	0.47	29.09 ±5.17	0.40	38.00 ±6.13	0.43	31.94 ±3.41	0.27

Table 1: Accuracy and calibration on three task categories: **Perception** (SeedBench), **Charts** (FigureQA), and **Knowledge** (macro-average over MMStar, CVQA, and ReasonVQA).

adopt a two-stage filtering strategy. First, we apply a consistency-based filter to isolate questions on which the sensors themselves struggle: for each question and sensor, we sample 11 independent responses, compute a consistency score K as the number of times the ground-truth option is predicted, and retain only questions with $1 \leq K \leq 5$. This guarantees that the VLM is occasionally correct but overall unreliable, so the reasoner cannot succeed by simply forwarding the original question and must engage in additional information seeking. Second, human annotators assume the role of a blind LLM reasoner: they interact with the VLM sensor via text-only queries (without seeing the image), and we keep only questions that are correctly solved by at least one annotator. Finally, we obtain 50 questions per (sensor, dataset) pair, yielding 500 challenging reasoning instances that span a diverse range of VQA skills and VLM failure modes.

4 Experiments and Analysis

4.1 Baselines and Metrics

We evaluate a range of LLM reasoners spanning proprietary and open-source families: GPT-4o, GPT-oss-20B, Qwen2.5 (7B, 72B), Qwen3 (8B, 32B), and LLaMA-3.1-8B. For each model, we consider three reasoning strategies. (i) *ReAct* (Yao et al., 2022b), where the model alternates between textual thoughts and actions, and actions correspond to querying the VLM sensor. (ii) *Uncertainty of Thoughts (UoT)* (Hu et al., 2024), adapted from the original information-seeking setting to VQA by modifying the prompt to describe our setup instead of medical diagnosis. (iii) *Belief-state prompting (BP)*, our tuned prompting strategy that instructs the model to implicitly track its internal belief over answer options during the interaction and to commit to a final choice only when sufficiently certain. We report *accuracy* and *Expected Calibration Error* (ECE) as our primary metrics. For non-proprietary models, we sample 5 trajectories per instance at temperature 1.0, and compute ECE by treating the empirical consistency of the predicted option across the 5 runs. For GPT-4o, due to cost constraints, we obtain a single deterministic response per instance with temperature 0.0. In addition, we collect a human baseline by randomly sampling 20 questions for each setting and asking a different human annotator to solve them, using the same interaction protocol.

4.2 Observations and Analysis

Tasks are genuinely hard even for strong LLMs. Across both sensors and all reasoning strategies,

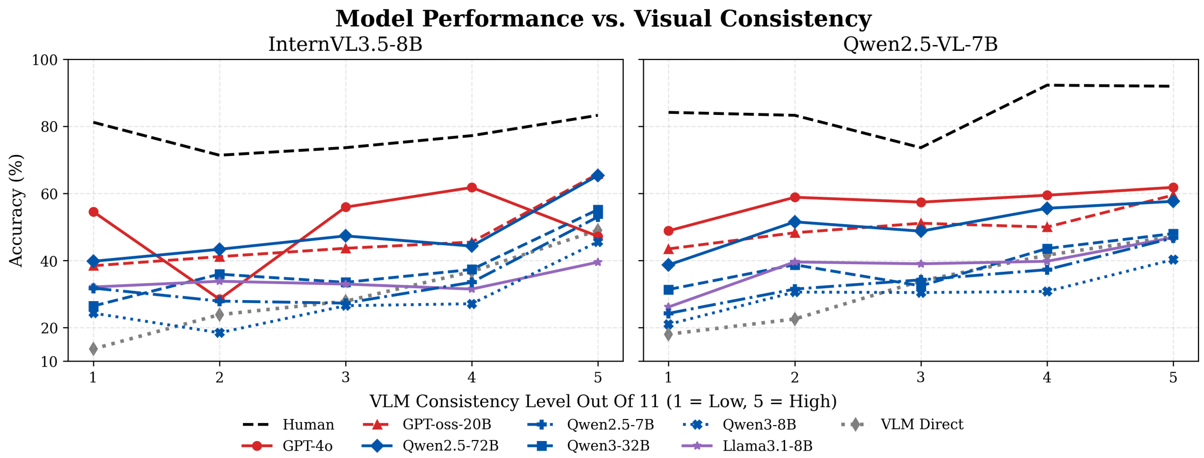

Figure 2: Model performance using BP vs. VLM consistency level. The x-axis shows binned levels of how often the VLM gives the correct answer (out of 11 samples), with 1 being the lowest consistency and 5 being the highest. All LLM reasoners improve over the raw VLM baseline (gray), with humans (dashed black) maintaining the highest performance across all consistency levels.

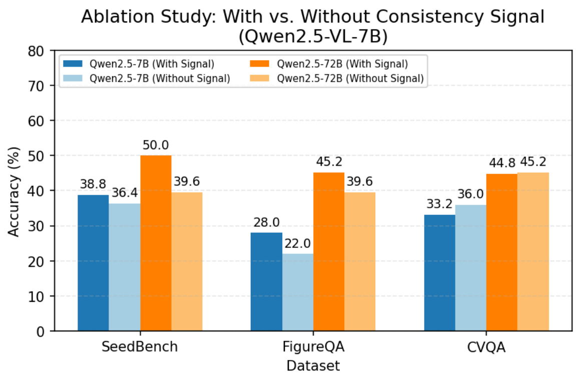

Figure 3: Ablation study showing performances with and without the consistency signal under BP.

accuracies for GPT-4o and the strongest open models stay in the 40-60% range, while human performances are at 70-100%, indicating our task creates unique challenges that even frontier models struggle to robustly reason over noisy visual evidence.

LLM reasoners improve raw VLM baselines. The inference-only LLM reasoners are able to outperform direct VLM prediction baselines with improved accuracy and reduced calibration error, indicating the LLMs can integrate and partially correct the inaccurate sensory inputs.

Effect of VLM consistency. Figure 2 shows the accuracy of humans, VLM-direct, and LLM-based reasoners as a function of the VLM self-consistency level. Across both sensors, LLM performance increases with consistency; however, the gap between VLM-direct and LLM-based systems narrows in the easiest bins, and only the largest reasoners achieve substantial gains over the VLM baseline there. In contrast, human blind performance degrades only mildly in the lowest-consistency bins, indicating a large performance gap in the most difficult settings. In addition, reasoners paired with the better-calibrated Qwen2.5-VL benefit more from increasing consistency than those using InternVL3.5, while human accuracy remains relatively stable with even miscalibrated signals.

Ablation study of the consistency signal. Figure 3 compares BP performance of Qwen2.5-7B/72B with and without the VLM's answer-consistency metadata. We observe sizable gains on perception and diagram datasets, especially for Qwen2.5-72B, indicating that strong reasoners can exploit this uncertainty cue. In contrast, the effect is negligible on the knowledge-based benchmark: the VLM can be confidently wrong, making consistency a noisy or even adversarial proxy for reliability. This presents a challenge as a reasoning task and suggests an interesting direction for future work where reasoners learn domain-specific strategies.

Additional results and analysis. Please see App. B and App. G for additional analysis. See App. C for detailed per-dataset results for the knowledge-based reasoning category.

5 Conclusion

We introduced VAR, a scalable benchmark for evaluating and studying text-only LLM reasoners that must solve VQA problems by querying a noisy VLM sensor equipped with uncertainty feedback.

6 Limitations

Our study has several limitations. First, VAR currently contains around 500 questions, although we extensively cover a diverse set of datasets and reasoning scenarios; expanding to larger and more open-ended settings is an important next step. Second, for computational reasons, we restrict attention to moderately sized VLM sensors, leaving it to future work to examine stronger and more diverse sensors and how their error modes and calibration affect downstream reasoning. Third, we only evaluate inference-time prompting baselines and do not train the LLM reasoners themselves; analyzing how supervised or reinforcement learning could shape querying policies and dataset-specific strategies for exploiting sensor uncertainty is an exciting direction for future work.

References

Marwa Abdulhai, Isadora White, Charlie Snell, Charles Sun, Joey Hong, Yuexiang Zhai, Kelvin Xu, and Sergey Levine. 2023. Lmrl gym: Benchmarks for multi-turn reinforcement learning with language models. *arXiv preprint arXiv:2311.18232*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14953–14962.

Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. 2024. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in large language models. *arXiv preprint arXiv:2402.03271*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pages 874–880.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888.

Zhaonan Li, Shijie Lu, Fei Wang, Jacob Dineen, Xiao Ye, Zhikun Xu, Siyi Liu, Young Min Cho, Bangzheng Li, Daniel Chang, and 1 others. 2025. Unbiased visual reasoning with controlled visual inputs. *arXiv preprint arXiv:2512.22183*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, and 1 others. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, and 1 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2020. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11888–11898.

Duong T Tran, Trung-Kien Tran, Manfred Hauswirth, and Danh Le Phuoc. 2025. Reasonvqa: A multi-hop reasoning benchmark with structural knowledge for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18793–18803.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022a. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022b. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31.

Zhanke Zhou, Xiao Feng, Zhaocheng Zhu, Jiangchao Yao, Sanmi Koyejo, and Bo Han. 2025. From passive to active reasoning: Can large language models ask the right questions under incomplete information? *arXiv preprint arXiv:2506.08295*.

A Related Work

Modular visual reasoning. Classical modular VQA approaches separate perception from reasoning by composing neural modules or executing programs over scene graphs and other structured visual abstractions, as in neural-symbolic VQA systems (Yi et al., 2018). More recent work such as ViperGPT and VisProg uses LLMs to synthesize code or programs that call visual tools while keeping a clean interface between text-based reasoning and perception (Surís et al., 2023; Gupta and Kembhavi, 2023). Closest to our setup, recent "blind" VQA frameworks (e.g., Li et al., 2025) let a text-only reasoner query a VLM sensor through restricted perception primitives; our benchmark follows this separation but focuses on explicitly *noisy* VLM feedback and empirical self-consistency signals.

LLM agents and tool use. A growing line of work studies LLM agents that interact with external tools or environments, such as web browsers, simulators, or APIs. Early systems like WebGPT and related web agents train models to browse, search, and cite online resources for question answering (Nakano et al., 2021). ReAct interleaves chain-of-thought with actions, enabling LLMs to plan and execute multi-step tool calls in textual environments (Yao et al., 2022b), while Toolformer explicitly learns to call APIs or modular experts when needed (Schick et al., 2023). More recent agent frameworks (e.g., Voyager and other long-horizon agents) extend this paradigm to embodied and open-ended environments with persistent state and skills (Wang et al., 2023). In most of these settings, tools are treated as approximately reliable oracles (e.g., a browser that returns correct pages or an API that returns accurate results), whereas in our benchmark the only tool for accessing the image is an explicitly noisy, stateless VLM sensor, with explicit consistency information associated with each query.

Retrieval-augmented generation (RAG). RAG systems augment LLMs with non-parametric memory by retrieving textual evidence and conditioning generation on the retrieved passages (Lewis et al., 2020). Dense retrievers such as DPR learn neural encoders that retrieve relevant documents, while Fusion-in-Decoder aggregates many retrieved passages directly in the decoder (Karpukhin et al., 2020; Izacard and Grave, 2021). VAR can be viewed as a perceptual analogue of RAG: instead of retrieving text, the reasoner queries a VLM to obtain multiple, possibly inconsistent visual descriptions and must aggregate them using an explicit uncertainty signal.

Active reasoning and information seeking under uncertainty. Active reasoning works study models that gather additional information before committing to an answer, often by interacting with tools or environments over multiple turns. LMRL Gym introduces benchmarks for multi-turn reinforcement learning with LLM-based agents, highlighting challenges of long-horizon control and partial observability (Abdulhai et al., 2023). In high-stakes domains, MedIQ evaluates question-asking LLMs that propose targeted follow-up queries to reduce diagnostic uncertainty in clinical reasoning (Li et al., 2024). Zhou et al. move from passive to active reasoning and ask whether LLMs can pose the right questions under incomplete information, proposing benchmarks and methods for query selection when key evidence is missing (Zhou et al., 2025). VAR instantiates a similar active reasoning loop in the visual domain: a text-only reasoner must decide which perception queries to issue to a noisy VLM sensor, interpret self-consistency as a coarse uncertainty signal, and determine when the accumulated evidence is sufficient to commit to an answer. To the best of our knowledge, we are the first to explicitly model uncertainty in the *environmental feedback* itself within an active reasoning benchmark; moreover, our construction is fully automatic and inexpensive, and does not require hand-crafted text simulators or environment-specific infrastructure.

B Additional analysis: general trend in model scales, and prompting strategies

Across all benchmarks, human annotators are strongest, GPT-4o is the best LLM reasoner, and all open-source models lag behind both. Within each family, larger models consistently outperform their smaller counterparts, with Qwen2.5-72B generally the strongest open-source reasoner across sensors and tasks. Trends are similar for both InternVL3.5-8B and Qwen2.5-VL-7B sensors, though absolute performance is higher with the latter. Prompting strategy also matters: belief-state prompting (BP) usually yields the best accuracy and ECE, ReAct is typically a close second and occasionally wins on chart QA, while UoT rarely dominates and can even hurt calibration.

C Knowledge Category Breakdown

		InternVL3.5-8B						Qwen2.5-VL-7B					
		MMStar		CVQA		ReasonVQA		MMStar		CVQA		ReasonVQA	
Reasoner	Strategy	Acc $\uparrow$	ECE $\downarrow$	Acc $\uparrow$	ECE $\downarrow$	Acc $\uparrow$	ECE $\downarrow$	Acc $\uparrow$	ECE $\downarrow$	Acc $\uparrow$	ECE $\downarrow$	Acc $\uparrow$	ECE $\downarrow$
Human	-	70.00	-	85.00	-	70.00	-	70.00	-	90.00	-	85.00	-
GPT-4o	BP	**64.00**	-	**56.00**	-	52.00	-	**56.00**	-	**66.00**	-	60.00	-
	UoT	44.00	-	48.00	-	**56.00**	-	54.00	-	52.00	-	74.00	-
	ReAct	52.00	-	48.00	-	32.00	-	46.00	-	64.00	-	70.00	-
GPT-oss-20B	BP	50.00 ±6.00	0.25	42.00 ±2.00	0.30	52.80 ±2.28	0.18	52.00 ±5.10	0.13	45.60 ±6.54	0.17	**74.40** ±9.94	0.11
	UoT	39.60 ±6.84	0.23	49.20 ±4.15	0.06	48.00 ±6.78	0.09	39.60 ±3.85	0.10	38.80 ±6.87	0.22	68.80 ±5.22	0.08
	ReAct	49.20 ±4.60	0.24	43.20 ±5.40	0.30	48.00 ±3.16	0.18	49.20 ±7.69	0.19	42.40 ±7.40	0.29	70.80 ±3.03	0.06
Qwen2.5-72B	BP	48.80 ±5.02	0.20	44.80 ±6.10	0.16	40.80 ±3.35	0.28	46.40 ±6.23	0.20	44.80 ±4.60	0.27	64.40 ±6.23	0.08
	UoT	42.00 ±2.83	0.22	41.20 ±3.35	0.30	36.80 ±6.57	0.25	44.40 ±7.13	0.20	46.40 ±4.34	0.19	61.60 ±6.23	0.08
	ReAct	46.80 ±2.28	0.26	44.00 ±3.16	0.30	42.00 ±4.00	0.25	42.00 ±5.83	0.30	39.20 ±3.03	0.30	69.20 ±7.29	0.17
Qwen3-32B	BP	38.80 ±7.56	0.19	34.80 ±6.72	0.13	31.20 ±5.76	0.21	35.60 ±5.18	0.10	35.60 ±7.27	0.20	46.80 ±6.72	0.22
	UoT	31.60 ±2.61	0.10	38.80 ±3.63	0.14	25.60 ±4.98	0.15	32.00 ±4.90	0.16	33.20 ±7.95	0.25	45.20 ±3.63	0.23
	ReAct	42.40 ±5.18	0.18	38.00 ±9.17	0.12	34.40 ±5.55	0.10	47.20 ±5.76	0.13	34.80 ±5.40	0.19	58.40 ±6.54	0.14
Qwen3-8B	BP	32.80 ±6.42	0.12	20.00 ±4.90	0.21	20.80 ±3.03	0.14	32.00 ±7.35	0.11	26.40 ±4.34	0.18	37.60 ±5.18	0.10
	UoT	27.60 ±8.29	0.18	25.20 ±7.56	0.19	22.00 ±4.90	0.10	30.40 ±3.29	0.17	23.20 ±6.87	0.25	31.60 ±3.85	0.34
	ReAct	36.00 ±7.62	0.16	24.40 ±6.54	0.30	24.40 ±2.97	0.16	32.00 ±2.00	0.30	34.00 ±9.17	0.22	42.00 ±4.24	0.17
Qwen2.5-7B	BP	33.60 ±7.54	0.30	26.00 ±3.74	0.39	34.40 ±5.55	0.23	38.00 ±7.75	0.12	33.20 ±5.02	0.30	38.40 ±2.97	0.12
	UoT	37.20 ±5.02	0.14	29.20 ±5.76	0.30	27.60 ±4.56	0.33	32.80 ±6.10	0.25	32.40 ±2.61	0.21	41.20 ±4.60	0.06
	ReAct	42.40 ±1.67	0.20	31.20 ±4.15	0.34	25.60 ±7.27	0.35	34.00 ±7.07	0.36	36.40 ±3.58	0.34	51.20 ±3.35	0.08
LLama3.1-8B	BP	32.40 ±6.69	0.18	34.00 ±6.16	0.19	30.00 ±10.39	0.18	31.60 ±6.39	0.21	36.80 ±2.28	0.20	44.00 ±6.16	0.12
	UoT	35.20 ±5.93	0.07	36.80 ±4.15	0.13	31.60 ±4.56	0.19	23.60 ±6.54	0.29	27.20 ±7.29	0.29	34.40 ±9.21	0.20
	ReAct	35.60 ±2.19	0.15	38.80 ±9.12	0.14	30.40 ±9.53	0.19	32.80 ±5.40	0.27	29.60 ±5.90	0.30	41.20 ±1.10	0.12
VLM	-	25.4 ±4.82	0.51	26.91 ±5.68	0.42	28.73 ±4.76	0.47	33.45 ±5.59	0.28	28.73 ±6.53	0.35	33.64 ±5.57	0.18

Table 2: Knowledge-based category breakdown showing accuracy and calibration on: **MMStar**, **CVQA**, and **ReasonVQA**.

D Belief-State Prompting Strategy (BP)

Belief-State Prompt Used to Evaluate BP method

Belief Tracking

- Initialize: If 2 options $\rightarrow$ 50% each; If 4 options $\rightarrow$ 25% each
- Update beliefs after each question based on response patterns and the VLM's confidence
- Aim for $\geq$90% confidence before finalizing your answer
- Only provide final answer when truly confident you've eliminated alternatives
- Responses with low confidence score will need to be further cross-validated because the VLM is providing conflicting information.

OUTPUT FORMAT

At every step, you must include the following and with the correct format:

- **Thought:** Before every question or final answer, explicitly state your thought process by outputting 'Thought: <complete description of your rationales>'.
- **Action:** Then output exactly one of:
  - 'My question is: <fully self-contained question>'
  - 'The answer is: (A)' or '(B)', '(C)', '(D)'

CRITICAL FORMAT RULES

- The phrase "My question is: " must appear exactly when asking a question
- The phrase "The answer is: " must appear exactly when providing final answer
- Output only ONE thought & ONE action per turn
- Each "My question" must include all necessary context (e.g., "about the largest red shape", "regarding the texture of the object on the right").
- You're allowed to ask unlimited questions, so you should not rely too heavily on early questions.

DECISION POINTS

Ask another question when:

- You're uncertain which option is correct
- You got conflicting answers that need clarification (The responses with low confidence score, such as 1/5 or 2/5)
- You want to verify your leading hypothesis
- You haven't yet tested all the key differences between options

Provide your final answer when:

- One option is clearly supported by multiple reliable observations
- You've tested the main alternatives and ruled them out
- You are $\geq$90% confident about the final option
- You've asked enough questions to feel justified in your conclusion

KEY REMINDERS

- Think about what you actually learned and what you're still uncertain about
- Ask follow-ups when you notice contradictions
- Make decisions when you feel reasonably certain
- Each question should be fully self-contained with all necessary context

---

Begin now with your first question. Think about what distinguishing features matter for this problem, and ask an exploratory question that will help you understand the scene.

E Uncertainty of Thoughts (UoT) Setup

UoT: Prompt for Generating Candidate Questions In Visual Setting

You are helping a blind reasoner answer an MCQ about an image. You will generate {n_candidates} diverse visual questions that could help distinguish between the MCQ options.
MCQ Question: {mcq_question}
Options: {options}
Round: {round_num}

Previous questions and answers:
{history}

Generate exactly {n_candidates} diverse visual questions that could help distinguish between the MCQ options. Each question should:

1. Be self-contained (include all necessary context)
2. Target a specific visual feature that differs between options
3. Help eliminate at least one MCQ option

Output a JSON array of exactly {n_candidates} question strings:

```
["Question 1 about specific visual feature",
 "Question 2 about another feature", ...]
```

Output ONLY the JSON array, no other text.

UoT: Prompt for Information Gain Estimation In Visual Setting

You are estimating how informative a visual question would be for solving an MCQ.
MCQ Question: {mcq_question}
Options: {options}
Current belief: {belief_str}

Candidate visual question: "{candidate_question}"

Imagine a vision model sees the image and answers this question. Predict 2-3 most likely responses and how each would change your beliefs over the MCQ options. Output ONLY the JSON object in the exact format below (no other text):

```
{
  "outcomes": [
    {
      "response": "likely answer 1",
      "probability": 0.5,
      "new_belief": {example_belief}
    },
    {
      "response": "likely answer 2",
      "probability": 0.3,
      "new_belief": {example_belief2}
    },
    {
      "response": "uncertain/unclear",
      "probability": 0.2,
      "new_belief": {current_belief}
    }
  ]
}
```

IMPORTANT:

- Probabilities must sum to 1.0
- Each new_belief must have exactly {n_options} values summing to 1.0
- Output only valid JSON

F ReAct Setup

ReAct: Vanilla Setup in Visual Setting

You are a visually-impaired reasoner using the ReAct (Reasoning + Acting) framework to answer a multiple-choice question about an image. You interact with a Vision Language Model (VLM) that can see the image.

ReAct Framework

ReAct combines explicit **reasoning traces** with **actions**. For each turn, you must:

1. **Thought**: Explicitly reason about what you currently know, what information is missing, and what to do next
2. **Action**: Either ask a visual question OR provide your final answer

OUTPUT FORMAT

Every response MUST follow this exact structure:
When you need visual information:

```
Thought: <your reasoning about the current
situation and what visual info would help>
My question is: <your visual question - must
be self-contained with all necessary context>
```

When you are ready to answer:

```
Thought: <your reasoning summarizing what
you learned and why you chose this answer>
The answer is: (A/B/C/D/E)
```

CRITICAL FORMAT RULES

- The phrase "My question is: " must appear exactly when asking a question
- The phrase "The answer is: " must appear exactly when providing final answer
- These phrases are how the pipeline recognizes your output—if they're missing or slightly different, the system will fail
- Output only ONE thought + ONE action per turn
- Each "My question" must include all necessary context (e.g., "about the largest red shape," "regarding the texture of the object on the right").

VLM RESPONSES

The VLM responses will include a confidence score (e.g., "Confidence: 3/5"). Use your Thought to evaluate responses critically and decide your next action.

EXAMPLE TRACE

Question: What color is the car in the image?
(A) Red (B) Blue (C) Green (D) Yellow

```
Thought: I need to determine the color of
the car to answer this question. I should
directly ask about the car's color.
My question is: What is the color of the
car visible in the image?
```

Response: The car appears to be blue.
Confidence: 4/5

```
Thought: The VLM reports the car is blue
with high confidence (4/5). This matches
option (B). I have sufficient evidence.
The answer is: (B)
```

Begin now. Think about what visual information would help distinguish between the answer options, then ask your first question.

G Number of Queries per Setting

Number of Queries and Reasoning Strategy. Figure 4 reveals distinct trade-offs between reasoning strategies across datasets. Belief-state prompting (BP) and Uncertainty of Thoughts (UoT) both exhibit higher query counts (4.1–5.4 rounds) relative to ReAct (2.1–3.4 rounds), indicating that uncertainty-aware prompting encourages more active information seeking. Across all three datasets, BP consistently queries the VLM more aggressively than UoT, particularly on SeedBench (4.4 vs. 4.3) and CVQA (5.4 vs. 4.7), suggesting BP's belief-tracking mechanism promotes deeper exploration. Notably, this increased querying correlates with the improved accuracy gains observed in the main paper: both BP and UoT substantially outperform ReAct, demonstrating that deliberate uncertainty quantification and internal state tracking drive effective interaction with noisy sensors.

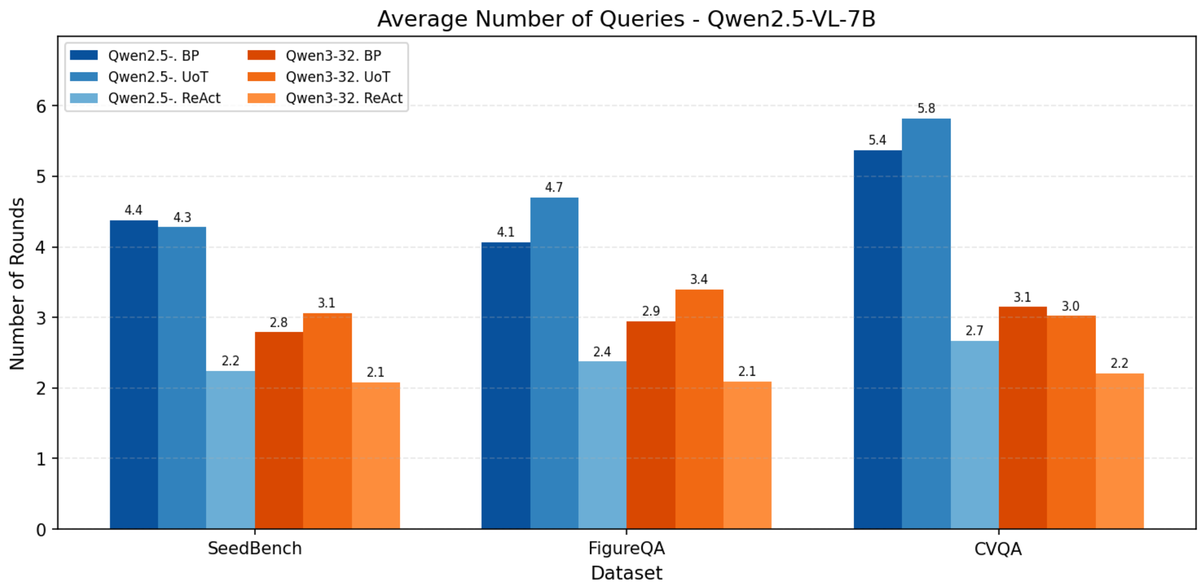

Figure 4: Average Number of Rounds for 3 Datasets Under 3 Different Settings

H Accuracy and Consistency Correlation

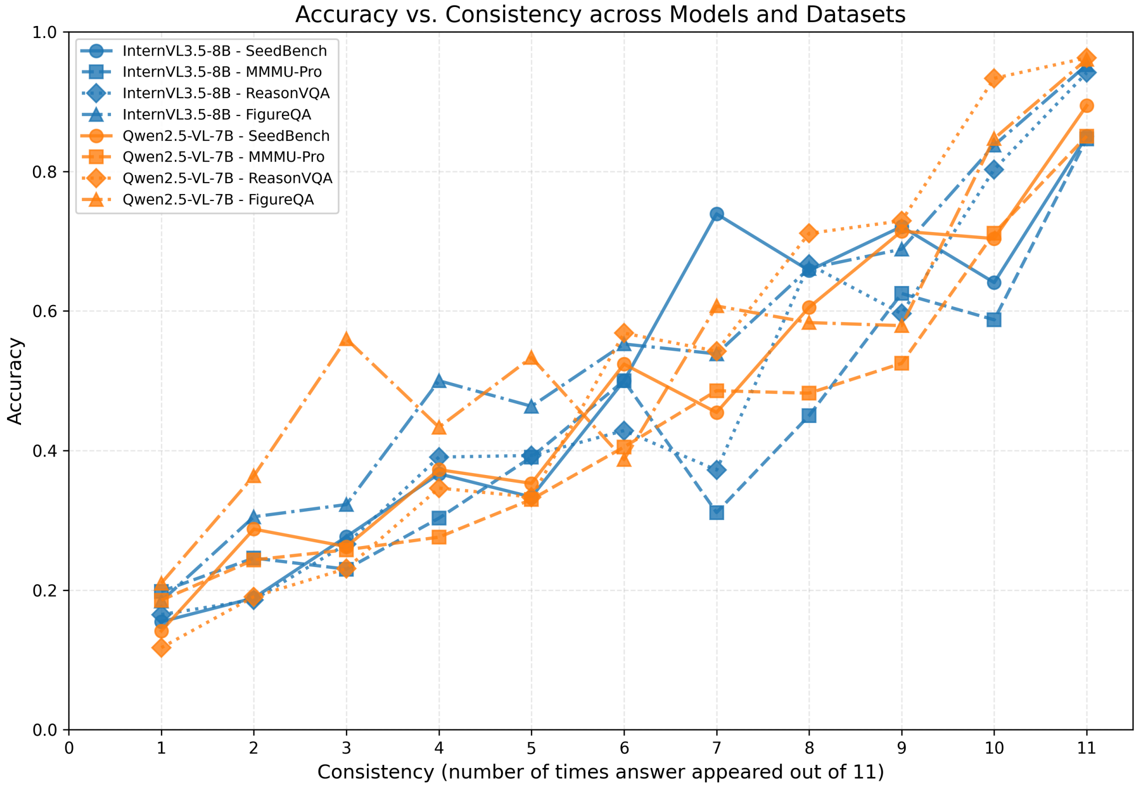

Figure 5: Accuracy vs. Consistency across different datasets for InternVL3.5-8B and Qwen2.5-VL-7B. Solid lines represent the base VLM performance, while dashed lines indicate reasoning performance. The x-axis represents the consistency level (1-11), showing that higher sensor consistency generally correlates with better downstream accuracy.

I Expected Calibration Error (ECE)

Expected Calibration Error (ECE)

	SeedBench	MMMU-Pro	CVQA	ChartQA-Pro	ReasonVQA	FigureQA	Average
InternVL3.5-8B	0.11	0.13	0.16	0.12	0.08	0.07	0.11
Llama3.2-11B-Vision	0.10	0.12	0.10	0.09	0.04	0.14	0.10
Qwen2.5-VL-7B	0.09	0.11	0.08	0.09	0.04	0.11	0.09

Figure 6: Preliminary effort to select the two most distinguished VLMs (InternVL3.5-8B and Qwen2.5-VL-7B) based on the largest gap in Expected Calibration Error (ECE) across diverse datasets.

J Model Traces on Belief-State Prompting

Step 1
To determine the correct answer, I need to gather information about the building's characteristics, such as its name, location, and purpose. I'll start by asking about the building's name to see if the VLM can provide a direct answer.
My question is: "What is the name of the building in the image?"
The building in the image appears to be a theater, characterized by its large entrance and classical architecture. Confidence: 3/5

Step 2
The VLM provided a response that the building appears to be a theater, but this does not match any of the given options. The confidence level is moderate, so I need to ask more specific questions to clarify the building's identity. I will focus on the building's name and purpose to narrow down the options.
My question is: "Can you confirm the name of the building in the image, specifically if it is called Hunnu Mall, Department Store, Ulaanbaatar Department Store, or Sport Center?"
The building in the image appears to be the Ulaanbaatar Department Store. Confidence: 4/5

Step 3
The VLM has provided a more specific and relevant answer, indicating that the building is the Ulaanbaatar Department Store with a confidence level of 4/5. This is a strong indication, but I want to ensure there are no contradictions or ambiguities. I will ask a follow-up question to confirm the building's purpose to further validate this answer.
My question is: "Can you confirm if the building in the image is used as a department store?"
The building appears to have an ornate facade with large arches, which is characteristic of historic architecture from certain European countries. It's less common for modern department stores to adopt such traditional architectural styles, suggesting it might serve another purpose or be a historical building, potentially not currently used as a department store. However, without more specific labels or signs visible on the building, it is difficult to confirm the exact function. Confidence: 3/5

Step 4
The VLM's response introduces some uncertainty about the building's current use, despite the previous indication that it is the Ulaanbaatar Department Store. Given the moderate confidence level, I need to ask a more direct question to clarify the building's current function and name.
My question is: "Are there any signs, labels, or other indicators in the image that specifically identify the building as the Ulaanbaatar Department Store?"
No, the image does not show any visible signage or labels that clearly identify the building as the Ulaanbaatar Department Store. Confidence: 4/5

Step 5
The VLM's response indicates that there are no visible signs or labels identifying the building as the Ulaanbaatar Department Store, which introduces more uncertainty. However, the previous response with a 4/5 confidence level suggested it is the Ulaanbaatar Department Store. I will ask a final clarifying question to see if there are any other distinguishing features or details that can help confirm the building's identity.
My question is: "Can you describe any distinctive features or details of the building that might help identify it as one of the options: Hunnu Mall, Department Store, Ulaanbaatar Department Store, or Sport Center?"
The building has a modern design with large windows, a flat roof, and decorative arches above the entryway. These features align more closely with those typically found in a shopping center like a mall or department store rather than a traditional sport center or single-use building. The presence of what appears to be advertisement screens on its facade further supports this conclusion. However, without more specific signage or interior clues, it's challenging to definitively classify the building as one of the given options. Confidence: 4/5

Step 6
The VLM's response provides more details about the building's modern design and features, which align more closely with a shopping center like a mall or department store. The presence of advertisement screens further supports this. Given the consistent indication that the building is likely a department store and the lack of evidence pointing to the other options, I am now reasonably confident that the building is the Ulaanbaatar Department Store. The correct answer is (C) Ulaanbaatar Department Store
The answer is: (C)

K Use of AI Tools

We used AI-assisted tools (e.g., large language models) for grammatical and stylistic editing, and to help search for and organize relevant background information and literature. All technical content, experimental designs, analyses, and conclusions were created and verified by the authors, who remain responsible for any remaining errors.

L Annotation Instruction

We recruited five expert annotators with prior experience in vision-and-language tasks. Annotators were instructed to carefully read each VAR question and its answer options, use common sense and domain knowledge to select the single best answer. Since our benchmark emphasizes reasoning under uncertainty and resolving inconsistent information, annotators were allowed to consult external tools (e.g., web search) when specialized domain knowledge was required. We first familiarized annotators with the task via a small pilot set of examples, discussed ambiguous cases, and provided feedback and clarifications on the guidelines before they proceeded to the final annotations.

M Annotation Interface

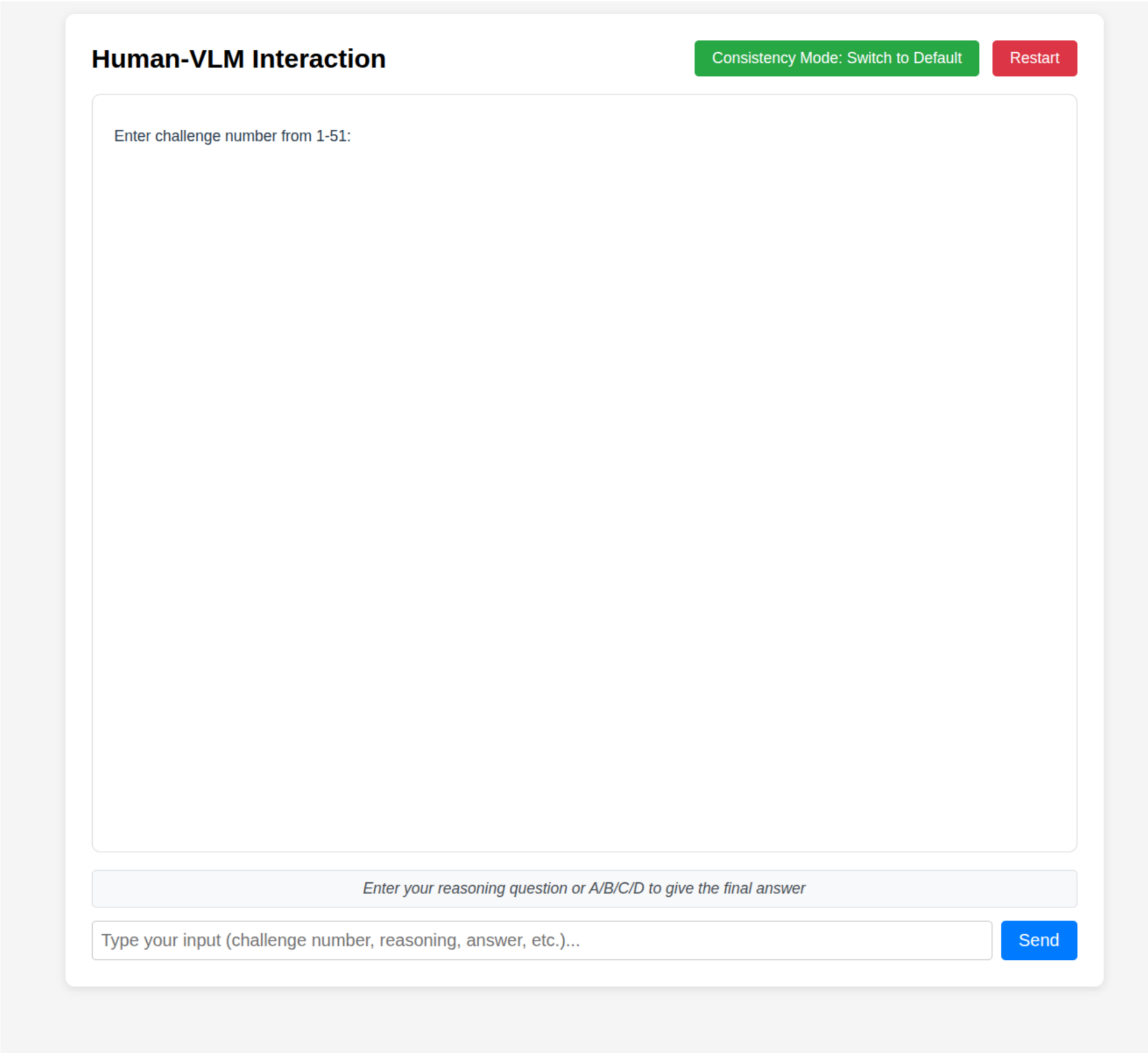

Figure 7: Annotation Interface for Human Annotators