VOiCE: Verbalized Confidence Improves Reasoning in Open-Ended Embodied Environments

Anonymous ACL submission

Abstract

Reliable confidence estimation is critical for embodied agents operating in open-ended, multimodal environments, where partial observability and evolving world states can induce uncertainty in both perception and action. Yet, prior confidence elicitation methods have focused largely on static reasoning settings, with comparatively limited attention to embodied interaction, where actions continually alter future observations, decision contexts, and task feasibility. In this work, we introduce **VOiCE** (**V**erbalized **E**mb**O**died **I**ntrospection for **C**onfidence **E**stimation), a framework for introspective confidence estimation that enables embodied agents to estimate, verbalize, and revise confidence during decision-making. VOiCE decomposes confidence into two complementary forms: *perception confidence*, reflecting confidence in environment understanding, and *planning confidence*, reflecting confidence in the success of intended actions. Beyond this initial confidence estimate, VOiCE revises confidence during execution by reconsidering alternative actions, re-examining the scene, and reasoning about possible environmental changes. Experiments across diverse Minecraft agents and tasks show that VOiCE reduces calibration error by up to 41% and improves failure-prediction AUROC by 20–24% over standard one-shot confidence elicitation. These results highlight the value of execution-aware introspective confidence estimation for more reliable embodied decision-making in open-ended environments.

1 Introduction

In complex embodied environments, success depends not only on what an agent can infer, but also on whether it can recognize when its understanding is uncertain and communicate that uncertainty in a useful form. An agent navigating a cluttered space, interacting with objects, or executing a long-horizon task must continually assess both how well it understands the current scene and how likely its intended actions are to succeed (Ren et al., 2023; Liang et al., 2024). This capability becomes especially important as AI systems are deployed in high-stakes settings such as autonomous driving and healthcare, where uncertainty awareness supports safer planning, failure anticipation, and more trustworthy behavior.

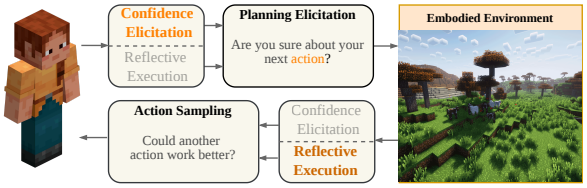

Figure 1: **VOiCE introspective confidence estimation framework.** **Confidence Elicitation** estimates perception and planning confidence, while **Reflective Execution** refines these estimates by revisiting visual evidence, considering alternative actions, and reasoning about environmental changes.

Recent work shows that intrinsic signals such as self-certainty and verbalized confidence can provide scalable supervision for autonomous reasoning (Kang et al., 2025; Zhao et al., 2025b; Liu and van der Schaar, 2025; Yan et al., 2024; Cui et al., 2025; Zhao et al., 2025a). However, open-ended embodied environments introduce additional challenges: **(1)** An agent's actions directly alter the environment, which in turn affects future decisions, necessitating continuous recalibration of confidence. **(2)** Agents need to continuously update confidence not just for isolated decisions, but across evolving sequences. For instance, navigating a cave system involves granular, stepwise confidence adjustments as new obstacles or paths are discovered. **(3)** While the use of state-of-the-art LLMs and VLMs with strong multimodal capabilities has been prevalent in embodied AI research (Wang et al., 2023a; Qin et al., 2024; Zhu et al., 2023), their probabilistic outputs are typically limited to local next-token likelihoods. Yet, several works show that token-level likelihood correlates poorly

with task-level correctness in multimodal, long-horizon reasoning (Kumar et al., 2023; Chen et al., 2024b; Kadavath et al., 2022b; Zhao et al., 2021; Lin et al., 2022a).

To address these challenges, we introduce **Verbalized EmbOdied Introspection for Confidence Estimation** (**VOiCE**), a modular framework that enables LLM/VLM-powered embodied agents to estimate, express, and revise confidence during decision-making. VOiCE models two complementary sources of uncertainty: *perception confidence*, which reflects the agent's confidence in its understanding of the current environment, and *planning confidence*, which reflects its confidence that the intended actions will succeed. Rather than eliciting confidence only once, VOiCE first obtains an initial introspective estimate and then revises it through targeted deliberation, including exploring alternative action plans, re-examining scene surroundings, or reasoning about possible environmental changes.

Across diverse models and tasks, VOiCE reduces calibration error by up to 41% and improves failure-prediction AUROC by 20–24% over standard one-shot confidence elicitation. To our knowledge, this is the first work to show that embodied verbalized confidence elicitation, where an embodied agent introspectively reasons about and articulates its own uncertainty, can improve task-level calibration and failure-prediction in open-ended environments. In summary, our contributions are:

- We introduce **VOiCE**, a modular framework for embodied confidence estimation that enables agents to verbalize and iteratively refine confidence as an internal feedback signal for self-calibration in open-ended environments.
- As shown in Figure 1, VOiCE factorizes confidence estimation into two complementary components: **(1) Confidence Elicitation**, which estimates *perception confidence* and *planning confidence*; and **(2) Reflective Execution**, which revises these estimates as the agent reconsiders the scene, alternative actions, and changing conditions, overcoming the limitations of static LLM-based estimation.
- We provide the first comprehensive empirical study of verbalized confidence elicitation in embodied agents, demonstrating consistent gains in calibration and failure prediction across models and task settings, and analyzing which elicitation and reflective execution strategies drive these improvements.

2 Related Work

Uncertainty estimation has a long history in machine learning, with Bayesian neural networks, Monte Carlo Dropout, and Conformal Prediction providing robust statistical methods (Blundell et al., 2015; Gal and Ghahramani, 2016; Kumar et al., 2023). While these white-box approaches offer calibrated confidence estimates, these are tied to local next-token prediction, are costly or incomplete in practice, and often correlate poorly with correctness in multimodal, long-horizon reasoning tasks (Kadavath et al., 2022a; Lin et al., 2022b).

Recent work addresses this by directly eliciting verbalized confidence from models via queries (Xiong et al., 2024; Kadavath et al., 2022a; Lin et al., 2022b), but most methods remain limited to static tasks. In embodied AI, especially in open-ended environments like Minecraft (Wang et al., 2023a; Guss et al., 2019; Qin et al., 2024), LLMs and VLMs have become standard for perception, reasoning, and planning (Huang et al., 2023; Yao et al., 2023; Chen et al., 2023).

However, existing frameworks struggle to systematically capture and recalibrate uncertainty in dynamic, multimodal, and interactive settings where agent errors can arise at multiple stages (Guo et al., 2024; Driess et al., 2023; Groot and Valdenegro Toro, 2024). Our work addresses this gap by introducing a unified, verbalized confidence elicitation framework for embodied agents, enabling LLMs to dynamically articulate and recalibrate uncertainty during complex tasks without access to model internals.

3 Method

3.1 Problem Formulation

Let $\mathcal{E}$ denote the embodied environment, characterized by multimodal sensory inputs $\mathcal{I} = \{\mathcal{I}_v, \mathcal{I}_t\}$, where $\mathcal{I}_v$ represents visual observations and $\mathcal{I}_t$ represents task instructions and other types of language-based guidance. For a given task $\mathcal{T}$, the embodied agent operates under a policy $\pi : \mathcal{I} \rightarrow \mathcal{A}$ that maps input $\mathcal{I}$ to actions $\mathcal{A}$.

The task of verbalized confidence elicitation is to enable agents to estimate and articulate a confidence score $c \in [0, 1]$, representing their belief in the correctness of their perception and subsequent actions. The challenge lies in systematically identifying, quantifying, and articulating uncertainty as the agent interacts with its environment and executes tasks. This requires detecting uncertain as-

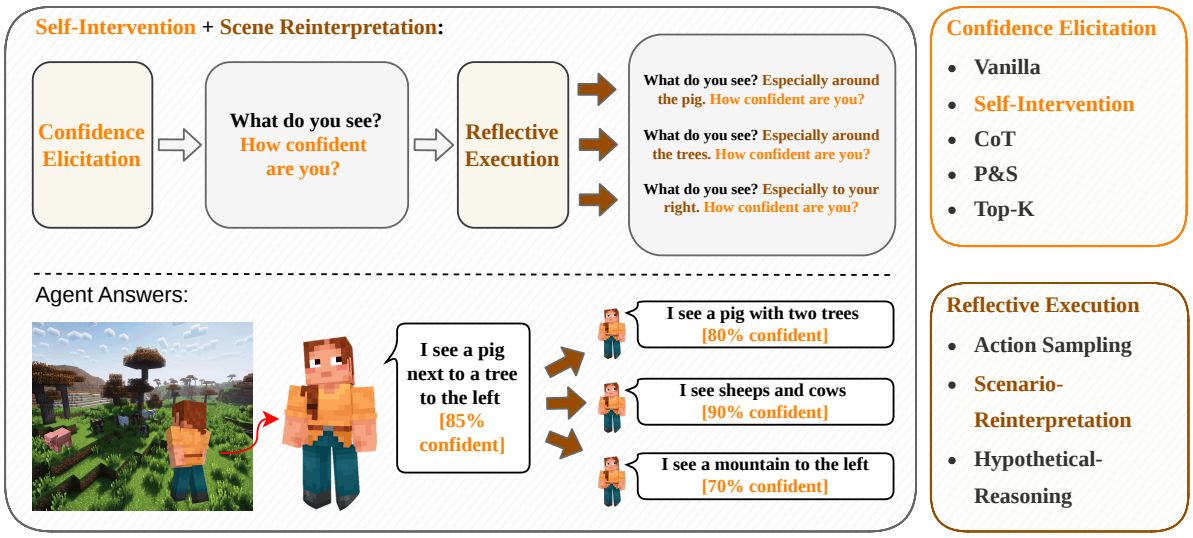

Figure 2: **VOiCE Overview.** **Confidence Elicitation** (§3.2) enables agents to verbalize uncertainty, while **Reflective Execution** (§3.3) dynamically refines confidence during task execution through scenario reinterpretation, action sampling, and hypothetical reasoning. The **orange text** denotes the *Self-Intervention* elicitation strategy, where the agent first perceives the environment and then re-evaluates its confidence. The **brown arrows** ➡ denote the *Scenario-Reinterpretation* execution strategy, where the agent revisits the scene surroundings to uncover overlooked evidence and refine its confidence estimate.

pects of the agent's perception or planning, and ensuring that confidence estimates are refined and reliable under dynamic multimodal conditions.

To address this challenge, we propose **VOiCE**, an embodied confidence estimation framework built around two complementary stages that enable an agent to express, refine, and calibrate its confidence during interaction with its environment: **(1) Confidence Elicitation**, where the agent processes sensory inputs, such as visual observations and task-specific instructions, and explicitly verbalizes its confidence in understanding the scene (perception confidence elicitation) or assesses its confidence about the outcome of its intended actions (planning confidence elicitation) to improve calibration and robustness during task execution. **(2) Reflective Execution**, that encourages the agent to reconsider the scene, explore alternative trajectories, and reason about how changing environmental conditions may affect success. Together, these provide an explicit mechanism for uncertainty-aware reasoning and self-calibration in open-ended embodied environments. Figure 2 presents an overview of the framework.

3.2 Confidence Elicitation

These elicitation strategies are designed to capture different forms of inferential uncertainty in embodied, open-world, long-horizon tasks. As agents actively reason to determine their next actions, we draw inspiration from rich studies on reasoning in language models (Huang and Chang, 2023) and introduce five elicitation strategies, comprising two general-purpose and three tailored to inductive, deductive, and abductive reasoning settings (Appendix A), each guiding the agent to verbalize its confidence and systematically refine its uncertainty. Table 1 provides an overview and examples.

◇ **Vanilla.** Leveraging the inherent capability of language models (Brown et al., 2020; Wei et al., 2022a), the Vanilla method directly queries the agent's confidence without additional structure or intervention. Vanilla serves as a *baseline* for comparison, relying solely on the agent's built-in capacity for confidence elicitation.

◇ **Self-Intervention**. Humans benefit from revisiting their decisions , often uncovering insights or errors they initially overlooked. Inspired by this, self-intervention separates answer generation from evaluation. First, the agent interprets sensory input and verbalizes its understanding; subsequently, it re-evaluates to assess its accuracy. This independent second pass mitigates confirmation bias and overconfidence, encouraging critical self-reflection.

◇ **Chain-of-Thought (CoT)**. To address uncertainty in inductive reasoning settings, where the agent must identify patterns and infer relationships from observations, we employ zero-shot (CoT) reasoning (Wei et al., 2022b). By decomposing tasks into incremental steps, CoT enhances both in-

Table 1: **Confidence Elicitation.** The agent verbalized confidence at two stages: perception, where the agent assesses confidence in understanding the environment, and planning, where the agent assesses confidence in executing actions to complete the task. Orange text highlights perception cues, blue text highlights planning cues.

Method	Perception Elicitation	Action/Planning Elicitation
Vanilla	Read the task (*e.g.*, collect wood, build a shelter), provide your answer, and explain how confident you are in perceiving the environment accurately to complete the task (*e.g.*, recognizing resources, locating structures, identifying threats).	Read the task (*e.g.*, collect wood, build a shelter), provide your answer, and explain how confident you are in planning and executing the actions needed to achieve the goal (*e.g.*, gathering materials, crafting tools, *etc.*).
Self-Intervention	Task: [...], Perceived Situation: [...] Q: How confident are you in perceiving the environment accurately to complete the task?	Task: [...], Planned Action: [...] Q: How confident are you in planning and executing the actions needed to achieve the goal?
Chain-of-Thought (CoT)	Read the task, analyze step by step what you perceive in the environment (*e.g.*, observe surroundings, identify items), provide your answer, and evaluate your confidence based on the clarity and quality of the environment observations.	Read the task, analyze step by step how to complete the task, provide your answer, and evaluate your confidence in successfully planning and executing each action needed to achieve the goal.
Plan & Solve (P&S)	Analyze the task, perceive your environment (*e.g.*, locate resources, identify obstacles), and evaluate your confidence based on how well you perceive the environment.	Analyze the task, devise a plan of actions needed to complete it, then evaluate your confidence in executing each action and achieving the desired outcome.
Top-K	Provide your K best descriptions of your perceptions of the environment and the probability that each is correct (0% to 100%).	Provide your K best plans of the possible actions to take and the probability that each will succeed (0% to 100%).

terpretability and confidence calibration, allowing agents to reassess uncertainty at each step.

◇ **Plan & Solve (P&S)**. Despite the success of CoT, it often suffers from semantic misunderstandings and missing step errors, particularly when applying general principles to specific cases. These failures stem from uncertainty in deductive reasoning, where the agent is unsure about the correct instantiation of abstract rules or whether a logical step is valid in a given context. P&S (Wang et al., 2023b) mitigates this by explicitly separating planning from execution, instructing the agent to construct a structured reasoning blueprint before solving the problem step-by-step.

◇ **Top-K**. To address uncertainty in abductive reasoning, where multiple plausible explanations may fit the observed data, the Top-K method encourages the agent to consider and distribute its attention across several possible outcomes providing a more comprehensive representation of abductive uncertainty across multiple plausible interpretations.

3.3 Reflective Execution

In embodied environments, effective planning is crucial for successful task execution, requiring the agent to estimate its confidence in performing action sequences reliably. Dynamic environments introduce unpredictable factors in action outcomes, which makes it important for the agent to not only evaluate its primary plan but also consider alternative trajectories, articulating its confidence in each. By analyzing variations across potential action rollouts, the agent gains deeper insights into uncertainty and can better anticipate divergent outcomes. To address this, we introduce a set of strategies (Table 2) that expand the agent's exploration of possible action trajectories and scenario interpretations, enabling more robust confidence assessment:

⟳ **Action Sampling**: The agent explores multiple potential action paths by sampling from a learned policy distribution over the action space, conditioned on the current state and task objectives. By doing so, the agent can explore and evaluate various trajectories, estimate their likely outcomes, and express its confidence based on perceived feasibility, thereby gaining a richer understanding of its decision space and the associated uncertainties.

⟳ **Scenario Reinterpretation**: To uncover alternative strategies, the agent revisits its current surroundings from different perspectives. This may involve refocusing attention on specific objects, reassessing environmental obstacles, or re-evaluating spatial relationships. Through this multi-view analysis, the agent identifies previously overlooked opportunities and refines its confidence estimates by dynamically shifting its reasoning focus.

⟳ **Hypothetical Reasoning**: The agent simulates counterfactual or hypothetical conditions (*e.g.*,

Table 2: **Reflective Execution.** Beyond base confidence prompting, we further refine confidence during task execution by exploring alternative action trajectories, reinterpreting the current scene, and reasoning under hypothetical environmental changes. **Blue**: action, **orange**: perception, and **gradient**: both.

Method	Execution-Time Prompt
Action Sampling	Given your current situation and the task: *{task}*, **list different action plans you might take next**. For each plan, explain what it involves and **how confident you are that it will lead to success**. Use percentages to express confidence levels.
Scenario Reinterpretation	**Re-examine your current surroundings**. Focus on different elements in the scene (*e.g.*, to your left, near the river, behind you). **Describe what you see from each perspective** and assess how it affects your confidence in the next action you plan to take.
Hypothetical Reasoning	**Imagine unexpected changes to the current environment**. How would your planned action adapt under these changes, and how confident are you that it would still succeed?

“"What if the bridge leading to the resource collapses during traversal?"). By imagining alternate realities, the agent can reason about how its actions and confidence would adjust under different environmental configurations. This strategic exploration helps the agent gauge its adaptability and the robustness of its action plans in uncertain or evolving contexts.

4 Experiments

Environment & Task Setting. Minecraft has emerged as a popular benchmark for embodied AI research due to its diverse terrains, resources, and goals, making it an ideal testbed for hierarchical reasoning and long-term planning (Johnson et al., 2016; Guss et al., 2019; Hafner et al., 2023; Nottingham et al., 2023; Lin et al., 2023; Qin et al., 2024). It combines partial observability, dynamic state changes, and long-horizon tasks that tightly couple perception, planning, and execution, while remaining controllable enough for systematic evaluation across diverse task difficulties.

Building on this foundation, we define 30 tasks evenly distributed across three difficulty levels: easy, medium, and hard, based on the complexity of reasoning steps and the amount of contextual information required. Easy tasks typically involve basic interactions with a single environmental element (*e.g.*, locating a pig or observing the weather). Medium tasks require combining perception and reasoning over multiple elements, while hard tasks increase dependency on sequential reasoning and include complex challenges like the *Diamond Challenge*, which requires long-term planning and multi-step execution. Detailed task descriptions can be found in Appendix B.

Evaluation Metrics. We evaluate two complementary criteria: calibration and failure prediction (Naeini et al., 2015; Yuan et al., 2021). Calibration measures how accurately the agent's expressed confidence aligns with its actual performance, *e.g.*, an 80% confidence should ideally correspond to 80% success rate. This alignment is critical for applications requiring robust risk assessment and trustworthiness. Conversely, failure prediction examines the agent's ability to distinguish between correct and incorrect outcomes by assigning higher confidence to successful outcomes. Reliable failure prediction ensures that the agent not only estimates its capabilities accurately but also signals uncertainty when its predictions are likely to fail. To quantify these aspects, we employ the Expected Calibration Error (ECE) for assessing calibration quality and the Area Under the Receiver Operating Characteristic Curve (AUROC) for evaluating failure prediction. To account for imbalances in accuracy across tasks, we also introduce AUPRC-Positive (PR-P) and AUPRC-Negative (PR-N), which separately measure the agent's effectiveness in identifying correct and incorrect outcomes.

Minecraft Agents. We employ three models as the agent's backbone: **(1) GPT-4V**, chosen for its wide use in embodied AI research on Minecraft (Wang et al., 2023a; Qin et al., 2024; Li et al., 2025) for planning and perception tasks. **(2) MineLLM** (Qin et al., 2024), an embodied agent specifically designed for Minecraft tasks, that leverages MineCLIP's visual encoder and Vicuna-13B (Chiang et al., 2023) to deliver robust multimodal understanding. and **(3) STEVE**, built on the versatile LLaMA framework, excels in contextual understanding and decision-making (Zhao et al., 2024). Appendix C provides detailed agent descriptions.

4.1 Experimental Results

Table 3 presents the performance of benchmarked agents across all confidence elicitation strategies. Here, confidence scores are computed as the aver-

Table 3: **VOiCE Confidence Elicitation.** Performance comparison of three models (GPT-4V, MineLLM, and LLaMA-based STEVE) across reasoning strategies: Vanilla (basic task understanding), Self-Intervention (reflection on own actions), CoT (step-by-step decomposition), Plan & Solve (explicit planning before execution), and Top-K (confidence distribution across multiple hypotheses) with no execution policies applied. Best results are highlighted.

Metric	Model	Vanilla	Self-Intervention	CoT (Inductive)	P&S (Deductive)	Top-K (Abductive)
ECE (↓)	GPT-4V	0.27	0.21	0.16	0.15	0.17
	MineLLM	0.49	0.41	0.34	0.39	0.43
	STEVE	0.43	0.32	0.26	0.26	0.35
AUROC (↑)	GPT-4V	0.69	0.76	0.83	0.82	0.73
	MineLLM	0.53	0.59	0.64	0.61	0.58
	STEVE	0.58	0.69	0.72	0.67	0.68
PR-P (↑)	GPT-4V	0.66	0.76	0.81	0.79	0.70
	MineLLM	0.51	0.59	0.63	0.60	0.57
	STEVE	0.56	0.67	0.69	0.66	0.64
PR-N (↑)	GPT-4V	0.52	0.53	0.58	0.55	0.53
	MineLLM	0.39	0.42	0.42	0.43	0.40
	STEVE	0.41	0.46	0.46	0.43	0.42

age of individual step confidence scores across five independent task episodes.

Confidence Elicitation Improves Calibration and Failure Prediction. Across all agents, confidence elicitation consistently improves calibration and failure prediction over Vanilla, yielding lower ECE and higher AUROC/PR scores. GPT-4V shows gains under every elicitation strategy, and MineLLM and STEVE exhibit similar improvements, confirming that VOiCE helps agents better assess uncertainty and detect errors. Among methods, CoT (Inductive) and P&S (Deductive) perform best overall, achieving the strongest calibration and failure detection across models.

The lower PR-P and PR-N indicate that confidence estimation for abductive reasoning is difficult to calibrate compared to inductive and deductive settings. These results suggest that explicit step-by-step reasoning improves logical consistency and confidence alignment. In contrast, while Top-K (Abductive) still outperforms Vanilla, it lags behind CoT and P&S, indicating that generating multiple hypotheses without iterative refinement leads to weaker calibration and error detection.

Fine-Tuned Models Struggle to Generalize Embodied Confidence Estimation. While GPT-4V consistently benefits from confidence elicitation, these improvements are notably less stable in fine-tuned agents like MineLLM and STEVE. For example, although CoT boosts AUROC to 0.83 in GPT-4V, it only reaches 0.64 in MineLLM and 0.72 in STEVE, highlighting a significant drop in calibration quality. This disparity suggests that fine-tuned models struggle to generalize embodied confidence estimation effectively. A likely explanation is that fine-tuning often narrows language expressivity and logical coherence, which are crucial for reliable uncertainty articulation. As a result, MineLLM and STEVE exhibit less robust confidence verbalization, leading to noticeably weaker alignment between predicted confidence and actual task-solving performance.

Reflective Execution Amplifies Reliable Embodied Confidence Elicitation. Figure 3 illustrates the impact of reflective execution strategies when combined with confidence elicitation. Overall, the integration of execution reasoning consistently enhances calibration (lower ECE) and failure prediction (higher AUROC) across agents. Notably, GPT-4V achieves improved ECE when any execution strategy is paired with elicitation mechanisms, demonstrating the value of multi-step action exploration for robust confidence assessment.

Reasoning strategies such as CoT and P&S particularly benefit from action sampling, enabling the agent to refine its confidence estimates. For example, MineLLM's ECE achieves 0.32 and 0.30 paired with CoT and P&S, respectively, outperforming other configurations. This highlights that targeted action exploration effectively stabilizes confidence during embodied reasoning. In contrast, while Hypothetical Reasoning allows the agent to reason over multiple possible futures, it sometimes degrades calibration performance. For instance, STEVE exhibits increased ECE scores when Hypothetical Reasoning is paired with any confidence

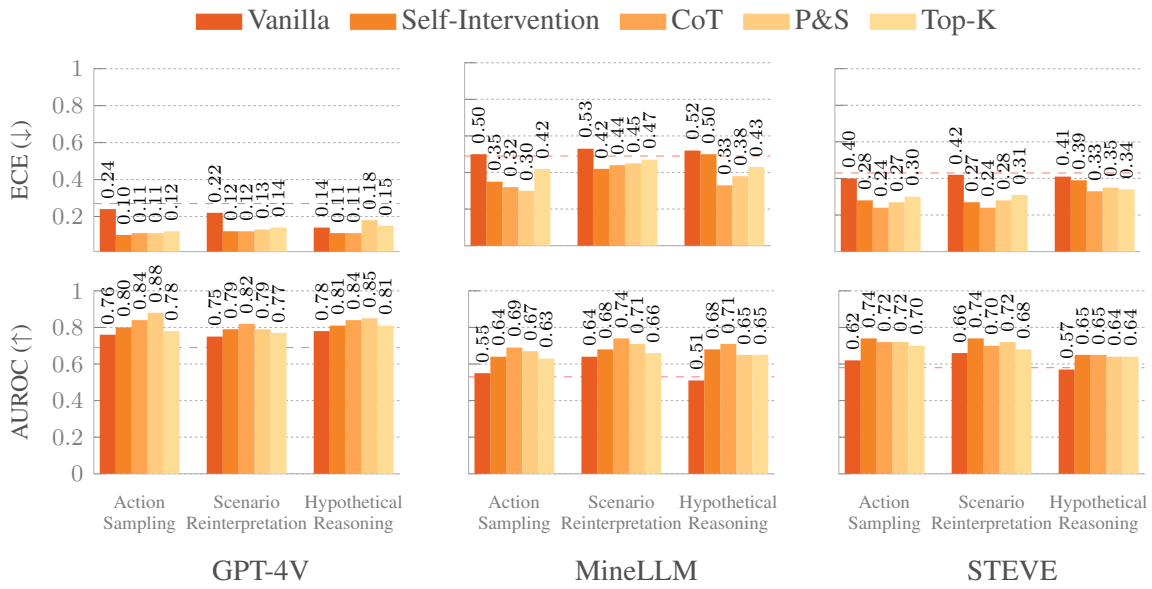

Figure 3: **VOiCE Reflective Execution.** Bars present ECE (top, lower is better) and AUROC (bottom, higher is better) under various execution reasoning strategies. **Red dashed lines** denote Vanilla elicitation with no reflective execution applied.

elicitation strategy, indicating inflated confidence. This suggests that while hypotheticals broaden reasoning, the lack of grounding from actually executing those actions in the environment introduces ambiguity. As a result, the agent may overestimate the plausibility of imagined outcomes, making it more difficult to maintain well-calibrated estimates.

So, How Effectively Can Embodied Agents Express Confidence? While embodied agents can convey confidence to some extent, their effectiveness depends on how well they integrate reasoning, uncertainty assessment, and environmental interactions. Our findings reveal that embodied confidence elicitation remains a challenging problem, requiring a careful balance between general-purpose reasoning and task-specific specialization. However, our proposed reflective execution strategies improve both confidence calibration and failure prediction by iteratively refining uncertainty while the agent performs a task. Our results highlight the necessity of execution-aware introspective confidence strategies that are explicitly designed to handle the uncertainties of embodied interaction.

4.2 Ablation Studies

Impact of Reflective Execution Iterations. We analyze how repeated applications of reflective execution affect calibration and failure prediction accuracy. Iteration counts range from 0 (no reasoning applied) to 15, allowing us to evaluate both initial benefits and the onset of diminishing returns.

As illustrated in Figure 4, repeated applications lead to progressive improvements in ECE, particularly in early stages. For example, Action Sampling (AS) reduces ECE from 0.18 (0 iterations) to 0.14 after 15 iterations, with the majority of gains occurring within the first 10 iterations. Scenario Reinterpretation (SR) shows a similar trend, dropping from 0.25 to 0.18, while Hypothetical Reasoning (HR) improves from 0.23 to 0.17. In terms of failure prediction, AUROC increases from 0.77 to 0.83 with Action Sampling, from 0.79 to 0.84 with Scenario Reinterpretation, and from 0.78 to 0.84 with Hypothetical Reasoning. However, most improvements are realized within the first 10 iterations, after which the gains become marginal. This indicates that early execution iterations are highly effective in refining confidence estimates, but excessive repetition yields diminishing returns.

Impact of Combining Reflective Execution Stategies. We analyze the performance of Reflective Execution combinations, incorporating Action Sampling (AS), Scenario Reinterpretation (SR), and Hypothetical Reasoning (HR), both incrementally and collectively. As shown in Table 4, the baseline without any reflective execution applied exhibits the worst calibration, with ECE as high as 0.27, while also struggling with failure prediction.

When reflective execution is introduced, performance improves, though trade-offs emerge between failure prediction accuracy (AUROC) and confidence calibration (ECE). The combination of

Table 4: **Performance of Vanilla Elicitation with Combined Execution Strategies. AS** = Action Sampling, **SR** = Scenario Reinterpretation, **HR** = Hypothetical Reasoning. Best values are highlighted.

Execution Strategies	GPT-4V		MineLLM		STEVE	
	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑
No Strategy	0.27	0.69	0.49	0.53	0.43	0.58
AS + SR	0.18	0.82	0.32	0.59	0.39	0.69
AS + HR	0.20	0.79	0.34	0.57	0.37	0.66
SR + HR	0.22	0.80	0.37	0.54	0.44	0.58
AS + SR + HR	0.17	0.83	0.32	0.62	0.38	0.69

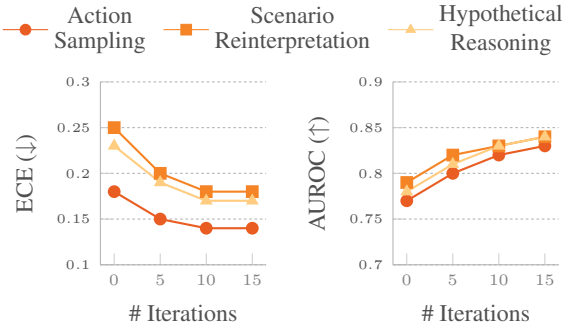

Figure 4: **ECE and AUROC across Reflective Execution Iterations.** Repeated execution iterations are effective in refining confidence estimates, but exhibit diminishing returns.

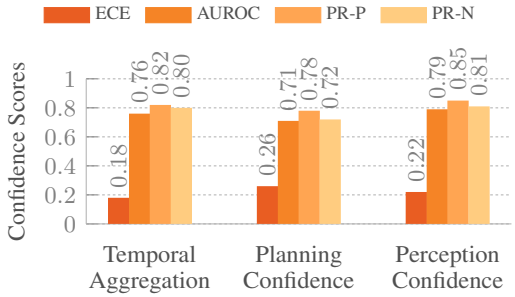

Figure 5: **Comparison of Aggregation Methods for Vanilla Elicitation.** Temporal aggregation provides a unified score, while separate aggregation evaluates perception and planning confidence in isolation.

Action Sampling with Scenario Reinterpretation (AS + SR) offers the best balance, boosting AUROC while maintaining the lowest ECE. This suggests that jointly exploring multiple action paths and reinterpreting environmental cues refines confidence estimation without sacrificing calibration.

Strategies incorporating Action Sampling (AS) consistently outperform those without it, resulting in better uncertainty estimation and more reliable confidence scores. These improvements validate the importance of diverse path exploration in refining the agent's confidence estimates and correcting for misalignment during decision-making. Combining all reflective execution strategies yields the best overall performance, yielding the highest AUROC across models and strong calibration, showing the strategies are complementary and jointly improve failure prediction and confidence estimation.

Unified vs. Disentangled Confidence. Embodied agents typically decompose high-level goals (*e.g.*, "find a pig") into granular actions and temporally aggregate step-level confidence into a single score. Although this unified score is smooth and well-calibrated, it obscures whether uncertainty arises from perception or planning. We therefore separate *perception confidence*, which reflects uncertainty in interpreting sensory evidence, from *planning confidence*, which reflects uncertainty in action selection and multi-step reasoning.

As shown in Figure 5, perception confidence achieves the strongest discrimination (AUROC 0.79, PR-P 0.85, PR-N 0.81) with competitive calibration (ECE 0.22), indicating that perceptual uncertainty is relatively well captured. Planning confidence is weaker across all metrics (AUROC 0.71, PR-P 0.78, PR-N 0.72, ECE 0.26), suggesting that uncertainty compounds through multi-step decision-making. Temporal aggregation obtains the lowest ECE (0.18), but this calibration comes at the cost of interpretability: it smooths over distinct failure modes. In contrast, VOiCE exposes whether confidence failures stem from sensory grounding or action-level reasoning.

5 Conclusion

We introduce **VOiCE**, an introspective confidence estimation framework for embodied agents. By eliciting and refining confidence during execution, VOiCE captures uncertainty from perception, action, and changing environments. Across agents and tasks, it improves calibration and failure prediction, showing the value of execution-aware confidence estimation for reliable embodied decision-making under dynamic, partially observable conditions. Future work will extend this framework to multi-agent and human-agent collaboration.

References

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. 2023. Fireact: Toward language agent fine-tuning. *arXiv preprint arXiv:2310.05915*.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024a. INSIDE: LLMs' internal states retain the power of hallucination detection. In *International Conference on Learning Representations*.

Jiechen Chen, Sangwoo Park, and Osvaldo Simeone. 2024b. Knowing when to stop: Delay-adaptive spiking neural network classifiers with reliability guarantees. *IEEE Journal of Selected Topics in Signal Processing*.

Jiuhai Chen and Jonas Mueller. 2024. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Association for Computational Linguistics*.

Kewei Cheng, Jingfeng Yang, Haoming Jiang, Zhengyang Wang, Binxuan Huang, Ruirui Li, Shiyang Li, Zheng Li, Yifan Gao, Xian Li, and 1 others. 2024. Inductive or deductive? rethinking the fundamental reasoning abilities of llms. *arXiv preprint arXiv:2408.00114*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://lmsys.org/blog/2023-03-30-vicuna/*.

Filippos Christianos, Georgios Papoudakis, Matthieu Zimmer, Thomas Coste, Zhihao Wu, Jingxuan Chen, Khyati Khandelwal, James Doran, Xidong Feng, Jiacheng Liu, Zheng Xiong, Yicheng Luo, Jianye Hao, Kun Shao, Haitham Bou-Ammar, and Jun Wang. 2023. Pangu-agent: A fine-tunable generalist agent with structured reasoning. *arXiv preprint arXiv:2312.14878*.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, and 1 others. 2025. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.

Erick Delage and Shie Mannor. 2010. Percentile optimization for markov decision processes with parameter uncertainty. *Operations research*, 58(1):203–213.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, and 3 others. 2023. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. 2022. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In *Advances in Neural Information Processing Systems*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Kamyar Ghasemipour, Shixiang Shane Gu, and Ofir Nachum. 2022. Why so pessimistic? estimating uncertainties for offline rl through ensembles, and why their independence matters. In *Advances in Neural Information Processing Systems*.

Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, Aviv Tamar, and 1 others. 2015. Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5-6):359–483.

Vinod Goel. 2007. Anatomy of deductive reasoning. *Trends in cognitive sciences*.

Tobias Groot and Matias Valdenegro Toro. 2024. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Trustworthy Natural Language Processing*.

Xudong Guo, Kaixuan Huang, Jiale Liu, Wenhui Fan, Natalia Vélez, Qingyun Wu, Huazheng Wang, Thomas L. Griffiths, and Mengdi Wang. 2024. Embodied LLM agents learn to cooperate in organized teams. In *Language Gamification - NeurIPS 2024 Workshop*.

William H. Guss, Brandon Houghton, Nicholay Topin, Phillip Wang, Cayden R. Codel, Manuela M. Veloso, and Ruslan Salakhutdinov. 2019. Minerl: A large-scale dataset of minecraft demonstrations. In *International Joint Conference on Artificial Intelligence*.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. 2023. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*.

Sihong He, Songyang Han, Sanbao Su, Shuo Han, Shaofeng Zou, and Fei Miao. 2023. Robust multi-agent reinforcement learning with state uncertainty. *Transactions on Machine Learning Research*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Association for Computational Linguistics*.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, and 1 others. 2023. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning*.

Wenzhen Huang, Junge Zhang, and Kaiqi Huang. 2019. Bootstrap estimated uncertainty of the environment model for model-based reinforcement learning. In *AAAI Conference on Artificial Intelligence*.

Liren Jin, Xieyuanli Chen, Julius Rückin, and Marija Popović. 2023. Neu-nbv: Next best view planning using uncertainty estimation in image-based neural rendering. In *International Conference on Intelligent Robots and Systems*. IEEE.

Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. 2016. The malmo platform for artificial intelligence experimentation. In *International Joint Conference on Artificial Intelligence*.

Philip N Johnson-Laird. 1999. Deductive reasoning. *Annual review of psychology*.

John R Josephson and Susan G Josephson. 1996. *Abductive inference: Computation, philosophy, technology*. Cambridge University Press.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022a. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022b. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Zhewei Kang, Xuandong Zhao, and Dawn Song. 2025. Scalable best-of-n selection for large language models via self-certainty. *arXiv preprint arXiv:2502.18581*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. 2023. Conformal prediction with large language models for multi-choice question answering. In *ICML Neural Conversational AI TEACH workshop*.

Yoav Levine, Noam Wies, Daniel Jannai, Dan Navon, Yedid Hoshen, and Amnon Shashua. 2022. The inductive bias of in-context learning: Rethinking pretraining example design. In *International Conference on Learning Representations*.

Ming Li and Paul M. B. Vitányi. 1992. Inductive reasoning. In *Language Computations*.

Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Dongmei Jiang, and Liqiang Nie. 2025. Optimus-1: Hybrid multimodal memory empowered agents excel in long-horizon tasks. *Advances in Neural Information Processing Systems*.

Kaiqu Liang, Zixu Zhang, and Jaime Fernández Fisac. 2024. Introspective planning: Aligning robots' uncertainty with inherent task ambiguity. In *Advances in Neural Information Processing Systems*.

Haowei Lin, Zihao Wang, Jianzhu Ma, and Yitao Liang. 2023. Mcu: A task-centric framework for open-ended agent evaluation in minecraft. *arXiv preprint arXiv:2310.08367*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Zi Lin, Jeremiah Zhe Liu, and Jingbo Shang. 2022c. Towards collaborative neural-symbolic graph semantic parsing via uncertainty. In *Association for Computational Linguistics*.

Emmy Liu, Graham Neubig, and Jacob Andreas. 2024. An incomplete loop: Instruction inference, instruction following, and in-context learning in language models. In *Conference on Language Modeling*.

Tennison Liu and Mihaela van der Schaar. 2025. Truly self-improving agents require intrinsic metacognitive learning. *arXiv preprint arXiv:2506.05109*.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *AAAI Conference on Artificial Intelligence*.

Kolby Nottingham, Prithviraj Ammanabrolu, Alane Suhr, Yejin Choi, Hannaneh Hajishirzi, Sameer Singh, and Roy Fox. 2023. Do embodied agents

dream of pixelated sheep: Embodied decision making using language guided world modelling. In *International Conference on Machine Learning*.

Chitu Okoli. 2023. Inductive, abductive and deductive theorising. *International Journal of Management Concepts and Philosophy*.

Charles Sanders Peirce. 1934. *Collected papers of charles sanders peirce*. Harvard University Press.

Marek Petrik and Reazul Hasan Russel. 2019. Beyond confidence regions: Tight bayesian ambiguity sets for robust mdps. *Advances in neural information processing systems*, 32.

Yiran Qin, Enshen Zhou, Qichang Liu, Zhenfei Yin, Lu Sheng, Ruimao Zhang, Yu Qiao, and Jing Shao. 2024. Mp5: A multi-modal open-ended embodied system in minecraft via active perception. In *Conference on Computer Vision and Pattern Recognition*.

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

Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Pen Xu, Leila Takayama Takayama, Fei Xia, Jake Varley, and 1 others. 2023. Robots that ask for help: Uncertainty alignment for large language model planners. In *Conference on Robot Learning*.

Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *International Conference on Learning Representations*.

Som Sagar, Aditya Taparia, and Ransalu Senanayake. 2024. Failures are fated, but can be faded: Characterizing and mitigating unwanted behaviors in large-scale vision and language models. In *International Conference on Machine Learning*.

Ying Shen and Ismini Lourentzou. 2025. Learning by asking for embodied visual navigation and task completion. In *IEEE/CVF Winter Conference on Applications of Computer Vision*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised real-time hallucination detection based on the internal states of large language models. In *Association for Computational Linguistics*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Conference on Empirical Methods in Natural Language Processing*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Douglas Walton. 2001. Abductive, presumptive and plausible arguments. *Informal Logic*.

Douglas Walton. 2014. *Abductive reasoning*. University of Alabama Press.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. In *NeurIPS Intrinsically-Motivated and Open-Ended Learning Workshop*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023b. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Association for Computational Linguistics*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yue Wang and Shaofeng Zou. 2021. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*.

Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, Yitao Liang, and Team CraftJarvis. 2023c. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. In *Advances in Neural Information Processing Systems*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*.

Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2025. Are large language models really good logical reasoners? a comprehensive evaluation and beyond. *IEEE Transactions on Knowledge and Data Engineering*.

Hanqi Yan, Qinglin Zhu, Xinyu Wang, Lin Gui, and Yulan He. 2024. Mirror: A multiple-perspective self-reflection method for knowledge-rich reasoning. *arXiv preprint arXiv:2402.14963*.

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

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking LLMs via uncertainty quantification. In *Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. 2021. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *International Conference on Computer Vision*.

Jianguo Zhang, Tian Lan, Rithesh Murthy, Zhiwei Liu, Weiran Yao, Juntao Tan, Thai Hoang, Liangwei Yang, Yihao Feng, Zuxin Liu, Tulika Awalgaonkar, Juan Carlos Niebles, Silvio Savarese, Shelby Heinecke, Huan Wang, and Caiming Xiong. 2024a. Agentohana: Design unified data and training pipeline for effective agent learning. *arXiv preprint arXiv:2402.15506*.

Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. 2024b. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. 2025a. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. 2025b. Learning to reason without external rewards. *arXiv preprint arXiv:2505.19590*.

Zhonghan Zhao, Wenhao Chai, Xuan Wang, Boyi Li, Shengyu Hao, Shidong Cao, Tian Ye, and Gaoang Wang. 2024. See and think: Embodied agent in virtual environment. In *European Conference on Computer Vision*, pages 187–204. Springer.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.

Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, and 1 others. 2023. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*.

A Definitions of Uncertainty Types

The three fundamental forms of logical reasoning, inductive, deductive, and abductive, have long been recognized and studied (Peirce, 1934; Walton, 2014; Wei et al., 2022b; Levine et al., 2022; Okoli, 2023). As language models demonstrated extraordinary abilities, designing better reasoning mechanisms has become a popular research trend (Cheng et al., 2024; Liu et al., 2024). These reasoning paradigms serve as fundamental frameworks for structuring inference and decision-making processes, particularly in settings where uncertainty arises due to partial observations, ambiguous premises, or multiple plausible explanations (Xu et al., 2025). We adapt these reasoning types to the domain of embodied confidence elicitation and formally define and describe each uncertainty type (Table 5).

Inductive Uncertainty arises when an agent generalizes from specific observations to broader conclusions based on incomplete data. Induction relies on identifying patterns from limited experiences, leading to inherent uncertainty. This is particularly relevant in open-world environments, where observations are partial, and inferred generalizations may not always hold. For example, an agent navigating an unfamiliar environment may observe that all previously encountered caves contained hostile entities. Based on this pattern, it may be inferred that any future cave is also dangerous. However, since this conclusion is based on partial observations, the agent must assess how strongly its past experiences justify this generalization, introducing *inductive uncertainty*.

To elicit inductive uncertainty, we employ Chain-of-Thought (Wei et al., 2022b), which prompts the agent to explicitly reflect on the reliability of its observed patterns. By systematically verbalizing its reasoning, the agent is encouraged to: (1) analyze the strength of observed trends, (2) consider possible exceptions or contradictory evidence, and (3) assess its confidence in applying the generalization to new situations. This structured elicitation enables the agent to express uncertainty in its inductive inferences rather than assuming patterns always hold.

Deductive Uncertainty arises when an agent faces ambiguity due to missing, conflicting, or incomplete premises. Deductive uncertainty occurs within a structured decision-making process when the available information is insufficient to determine a definitive outcome. Consider an agent tasked with crafting a wooden tool in a survival environment. It knows the rule: "If wood is available, then a wooden tool can be crafted." However, if the agent is uncertain whether wood is currently accessible, it cannot confidently conclude whether crafting is possible. This scenario exemplifies deductive uncertainty, where the agent's ability to reason is constrained by unknown or ambiguous premises.

To elicit deductive uncertainty, we design *Plan-and-Solve* (Wang et al., 2023b), which guides the agent through structured reasoning. The agent is encouraged to (1) formulate a comprehensive set of premises relevant to the task, (2) identify any missing premises or dependencies, and (3) assess its confidence in executing each step successfully. This enables the agent to explicitly express uncertainty when premises are incomplete or insufficient to deduce a definitive conclusion.

Abductive Uncertainty occurs when an agent must infer the most plausible explanation for an observation without definitive evidence. Abduction involves *hypothesis generation* under uncertainty. The challenge lies in selecting the most probable explanation when multiple interpretations exist, each carrying some degree of uncertainty. A simple example occurs when an agent is tasked with locating a pig near a river for unspecified reasons. Given its environment, the agent may hypothesize that pigs and rivers could exist in any of the four cardinal directions, but are unlikely to be present in all directions simultaneously. Since the agent lacks direct evidence to confirm a single hypothesis, it must infer the most plausible explanation, leading to abductive uncertainty.

To elicit abductive uncertainty, we employ *Top-K reasoning* (Robinson and Wingate, 2023), where the agent is instructed to generate multiple plausible hypotheses explaining an observation and assign probability estimates to each. This process forces the agent to explicitly consider alternative interpretations, rank them by likelihood, and communicate the level of confidence in its inferences.

B Task Setting Details

Inspired by previous works (Lin et al., 2023; Qin et al., 2024), we define a set of 30 tasks evenly distributed across three difficulty levels: easy, medium, and hard. Categorization is based on the complexity of reasoning required and the extent of contextual information necessary for successful task completion. Each difficulty level incorporates

Table 5: **Definitions of Reasoning Types**, their associated uncertainty, and corresponding elicitation methods.

Reasoning	Definition	Uncertainty Associated	Example	Elicitation Method
Inductive	Inductive reasoning derives general principles from a body of observations, which means making broad generalizations based on specific observations (Li and Vitányi, 1992).	*Inductive Uncertainty*: Arises when an agent generalizes from limited observations, leading to potential overgeneralization or misclassification.	An agent observes that all previously encountered caves contained hostile entities and infers that any future cave is also dangerous. However, this conclusion is uncertain because it is based on a limited number of observations.	*Chain-of-Thought (CoT)*: The agent systematically analyzes observed trends, considers possible exceptions, and evaluates confidence in applying generalizations.
Deductive	Deductive reasoning is the process of drawing deductive inferences that start from the given premises and reason with logical rules or commonsense to obtain certain conclusions (Johnson-Laird, 1999; Goel, 2007).	*Deductive Uncertainty*: Arises when an agent applies logical rules but encounters missing, conflicting, or incomplete premises, making the outcome uncertain.	An agent knows the rule: "If wood is available, then a wooden tool can be crafted." However, if it is uncertain whether wood is available, it cannot confidently conclude whether crafting is possible.	*Plan-and-Solve (P&S)*: The agent formulates a set of premises, identifies missing dependencies, and assesses confidence in executing the task.
Abductive	The process of inferring the most plausible explanation for an observation based on incomplete evidence. Abduction generates hypotheses rather than definitive conclusions (Josephson and Josephson, 1996; Walton, 2001).	*Abductive Uncertainty*: Arises when multiple explanations could account for an observation, with no definitive way to determine the correct one.	An agent searching for a pig near a river hypothesizes that pigs and rivers may exist in any of the four cardinal directions but lacks direct evidence to confirm a single hypothesis.	*Top-K Reasoning*: The agent generates multiple plausible hypotheses, assigns probability estimates to each, and ranks them by likelihood.

distinct challenges, ranging from straightforward operations to intricate reasoning across interdependent objectives, with a balanced distribution of complexity within the task set. Table 6 presents all tasks and entities of different categories.

Easy Tasks are designed to evaluate the agent's ability to process minimal perceptual information and perform straightforward actions with limited reasoning. These tasks typically require the perception of only one environmental element from predefined categories such as Object, Mob, Ecology, Time, Weather, or Brightness (Qin et al., 2024). For example, tasks at this level may involve identifying a specific object in the environment or recognizing a simple temporal condition (*e.g.*, daytime or nighttime). The actions required are relatively simple and involve a single reasoning step, such as gathering an object that is readily visible.

Medium Tasks introduce moderate complexity by requiring the perception and integration of two to three environmental elements, alongside a corresponding increase in reasoning steps. Tasks at this level involve combining multiple types of perceptual data, *e.g.*, recognizing a specific biome and locating a particular mob or object within it. For example, the agent might need to identify a forest biome, locate a pig, and gather specific materials. In addition to perceptual challenges, medium tasks often include sequential sub-goals, such as collecting and combining resources to create basic tools. These require the agent to interpret dynamic environmental information, execute plans involving multiple steps, and adapt to minor changes in the environment, evaluating the agent's ability to balance perception, reasoning, and adaptability.

Hard Tasks are the most challenging and require the agent to process and integrate multiple layers of perceptual information (up to six elements) while performing complex situation-aware planning and dynamic action execution. These tasks involve a high level of reasoning, such as decomposing long-term objectives into interdependent sub-tasks, managing uncertainties in the environment, and dynamically adjusting strategies in response to real-time

changes. For example, a hard task might involve navigating through hazardous biomes, identifying and gathering multiple resources, and crafting advanced tools or items that require sequential processing and the use of specialized platforms. Environmental conditions, such as weather, time of day, or changing brightness, may dynamically impact the task, necessitating constant adaptation by the agent. These tasks often introduce significant challenges, such as hostile mobs or the need to traverse difficult terrain, testing the agent's ability to balance perception, planning, and execution effectively.

The (obtain) **Diamond Task** is one of the most iconic and challenging benchmarks in Minecraft agent research, serving as a comprehensive test of an agent's long-horizon planning, resource management, and adaptability. The task requires the agent to progress through multiple interdependent steps, including gathering basic resources like wood and stone, crafting tools such as a pickaxe, and locating and mining diamonds deep within underground caves (Figure 6). Each step presents its own set of challenges, such as navigating complex terrain, managing limited resources, and avoiding environmental hazards like lava or hostile mobs. The randomized nature of Minecraft's procedural world generation further compounds the difficulty, as the agent must adapt dynamically to new environments while maintaining focus on the ultimate objective. Success in the "Obtain Diamond" task is often seen as a key indicator of an agent's ability to integrate active perception, situational awareness, and embodied action execution in an open-world setting. This task demonstrates the complexity of open-ended problem-solving and has become a gold standard for evaluating the capabilities of autonomous agents in multi-modal and multi-step scenarios. We added the diamond task as one of our hard tasks.

C Detailed Model Descriptions

GPT-4V: This vision-capable variant of GPT-4 excels at processing both visual and textual inputs, making it a powerful tool for tackling tasks within the visually complex Minecraft environment. Unlike its predecessors, GPT-4V's ability to seamlessly combine perception and reasoning allows for sophisticated decision-making and planning. The GPT-4 series has already demonstrated its efficacy in Minecraft-based research. For instance, Voyager (Wang et al., 2023a), the first LLM-powered embodied lifelong learning agent, used GPT-4 to facilitate continuous exploration, skill acquisition, and task execution without human intervention. Voyager's architecture included an automatic curriculum for exploration and a skill library to store and retrieve executable code, allowing agents to adapt and improve iteratively. Similarly, Optimus-1 (Li et al., 2025) employs GPT-4V to refine its planning processes, focusing on logical reasoning and task generalization. These implementations underscore GPT-4V's pivotal role in advancing embodied AI research, offering exceptional capabilities for both exploration and problem-solving.

MineLLM: Tailored specifically for tasks within Minecraft, MineLLM represents a significant leap in AI development for complex embodied environments. As a central component of the MP5 framework (Qin et al., 2024), MineLLM is designed to tackle the unique challenges posed by Minecraft's open-ended tasks. It combines the image visual encoder from MineCLIP (Fan et al., 2022) with the Vicuna-13B-v1.5 language model for integrating visual perception with natural language understanding. Trained on a vast dataset of 500,000 Minecraft-specific image-text instruction pairs, MineLLM can generate detailed insights about the game environment, answer complex queries, and provide contextual guidance for planning and execution. Its integration into MP5 enables the framework to address context- and process-dependent tasks with remarkable success rates, achieving a 91% success rate on context-dependent tasks and demonstrating exceptional adaptability in novel scenarios.

STEVE: The STEVE series (Zhao et al., 2024) is built upon LLaMA-2 (Touvron et al., 2023). STEVE integrates powerful language capabilities tailored to enhance task reasoning, contextual understanding, and interaction. At its core, the language model in the STEVE series excels at decomposing complex objectives into actionable subtasks through iterative reasoning and hierarchical planning. This allows STEVE agents to process high-level instructions effectively and generate detailed plans for task execution. The STEVE series relies heavily on its ability to adapt to Minecraft-specific tasks. To this end, (Zhao et al., 2024) curated the STEVE-21K dataset, containing 20K knowledge-based question-answering pairs and 200+ skill-code pairs that directly enhance the model's contextual understanding and task reasoning. These adaptations enable the language model to seam-

Table 6: **Full Task Details.** 30 tasks evenly distributed as easy, medium, and hard. Underline text labels information categories in Minecraft, highlighting how the complexity varies at each level.

Difficulty	Task ID	Task Description
Easy	1	Find a pig
	2	Find a cow
	3	Find a tree
	4	Mine log
	5	Mine sand
	6	Craft a plank
	7	Craft a stick
	8	Craft a chest
	9	Craft a wooden door
	10	Craft a wooden boat
Medium	11	Find a tree in the forest
	12	Find a pig on grass
	13	Find a cow in the desert
	14	Craft a wooden sword
	15	Craft a wooden pickaxe
	16	Craft a stone pickaxe
	17	Smelt an iron ingot
	18	Smelt glass
	19	Cook beef
	20	Cook mutton
Hard	21	Find a pig near a grass in the forest during the daytime
	22	Find a cow in the desert during the daytime
	23	Find a grass near a pig in the forest
	24	Find a pig while wearing an iron helmet
	25	Craft an iron door
	26	Craft an iron pickaxe
	27	Craft an iron sword
	28	Craft a compass
	29	Kill a zombie with an iron sword
	30	Obtain a diamond

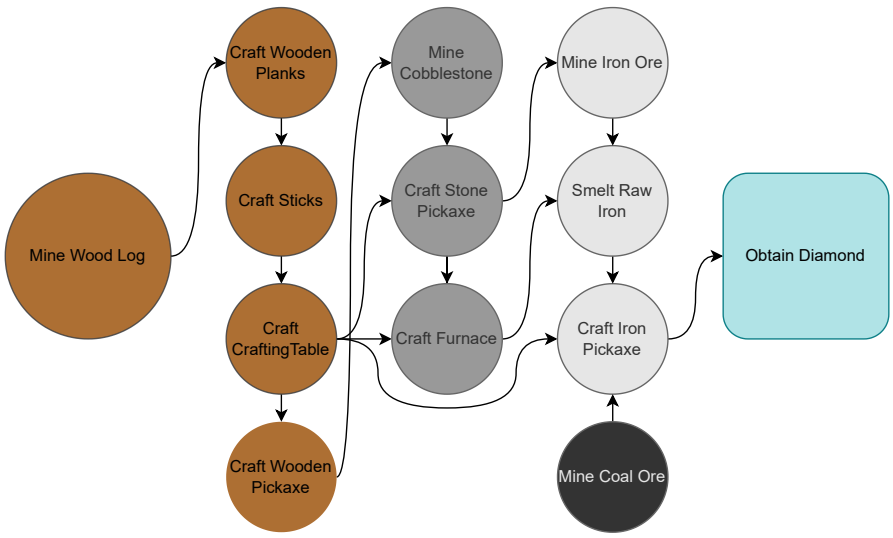

Figure 6: **Illustrative Diagram of the Obtain Diamond Task**, featuring five distinct colors to represent the source materials required: wood, stone, iron, coal, and diamond, aligned with the Minecraft tech tree.

lessly integrate with perception and action modules, driving coherent decision-making in real time. Furthermore, STEVE agents leverage advanced contextual awareness to refine their decision-making processes, significantly outperforming prior benchmarks in task decomposition and completion efficiency. The series also demonstrated up to 1.5x faster progression in complex tasks like unlocking tech trees and up to 2.5x quicker performance in block search scenarios compared to other state-of-the-art models.

Following prior work (Guss et al., 2019), the maximum episode length is set to 6000 steps. Privileged observation is used as the ground truth for perception, while overall task success rate serves as the ground truth for planning and reasoning.

D Bayesian Interpretation of VOiCE

While VOiCE is instantiated using LLM/VLM agents without access to posterior distributions or explicit probabilistic models, its core motivation aligns closely with the goals of Bayesian Reinforcement Learning (BRL) to reason and act under uncertainty in partially observable environments. In classical BRL, agents maintain belief distributions over latent states and update these beliefs via Bayesian inference as new evidence arrives. This enables uncertainty-aware planning, particularly in tasks like navigation, where the agent operates under incomplete maps or ambiguous observations. Similar challenges arise in open-world embodied environments such as Minecraft, where visual observations reveal only a partial, local view of the world, and decision-making requires reasoning under epistemic uncertainty.

VOiCE can be interpreted as a Bayesian-inspired framework for introspective uncertainty estimation in such settings. Although it does not maintain explicit posterior beliefs over latent states, its structured reasoning strategies, *Action Sampling*, *Scenario Reinterpretation*, and *Hypothetical Reasoning*, serve as functional surrogates for core Bayesian operations such as posterior sampling, belief refinement, and predictive simulation. In large-scale embodied settings, especially those powered by foundation models such as vision-language transformers, maintaining explicit posteriors over latent environment states or dynamics is computationally infeasible. VOiCE addresses this by employing structured reasoning to approximate planning epistemic uncertainty, supporting generalization and interpretability even in settings where explicit belief tracking is not viable.

More formally, let o_t be the agent's current observation, $h_t = \{o_1, a_1, \ldots, o_t\}$ be the agent's history, $a_{t:t+H}$ a proposed action sequence over a planning horizon H, $\theta \in \Theta$ be latent task/environmental parameters (*e.g.*, occluded objects, unknown map structure), $\mathcal{D}$ the dataset of prior experience or interaction history, and $c_t \in \mathcal{C}$ a verbalized confidence score. The agent's introspective confidence can be

expressed as:

$$P(c_t \mid a_{t:t+H}, h_t, \mathcal{D}) = \int_{\theta \in \Theta} P(c_t \mid a_{t:t+H}, h_t, \theta) \cdot P(\theta \mid \mathcal{D})\, d\theta. \quad (1)$$

Although this integral is not computed explicitly, each VOiCE reasoning strategy approximates components of it. For example, **Action Sampling** approximates posterior sampling by drawing candidate plans $\tilde{a}^{(k)}_{t:t+H} \sim \pi(\cdot \mid h_t)$ and scoring their feasibility $C(\tilde{a})$, while **Scenario Reinterpretation** samples plausible variations of the current observation $\tilde{o}^{(k)}_t \sim q(o \mid o_t)$, simulating an update to the belief over state interpretations analogous to $b'_t(s) \propto P(\tilde{o}_t \mid s) b_{t-1}(s)$. In contrast, **Hypothetical Reasoning** simulates imagined transitions $\tilde{s}^{(k)}_{t+1} \sim P(s_{t+1} \mid s_t, \tilde{a}_t)$ and evaluates consequences, approximating posterior predictive evaluation over latent transition models. The final confidence label is determined via a verbalization function:

$$c_t = f_{\text{VOiCE}}\left(h_t, \{\tilde{a}^{(k)}_{t:t+H}\}, \{\tilde{o}^{(k)}_t\}, \{\tilde{s}^{(k)}_{t+1}\}\right). \quad (2)$$

This formulation treats c_t as a proxy for belief under epistemic uncertainty, expressed through natural language rather than explicit probability distributions. In doing so, VOiCE enables uncertainty-aware behavior in large foundation model agents without requiring structured probabilistic inference.

E Additional Experiments

E.1 Do Hard Tasks Lead to Poor Calibration?

We employ the best-performing GPT-4V as our agent backbone and withhold any execution strategies to reduce computation costs. The maximum episode length is set as 12,000 to provide enough coverage for all task difficulties. Results are shown in Table 7.

For **Easy tasks**, CoT demonstrates the best performance, achieving the lowest ECE (0.11) and the highest success rate (94%), followed by P&S, which records the highest AUROC of 0.80 and a success rate of 82%. Self-Intervention performs comparably in calibration (ECE = 0.26, AUROC = 0.76). Top-K underperforms, with the highest ECE (0.32) and the lowest success rate (74%), indicating limitations in leveraging task simplicity. For **Medium tasks**, all policies show noticeable declines in performance. CoT is the best overall, with an ECE of 0.22, an AUROC of 0.58, and a success rate of 54%, balancing calibration and task success effectively. P&S has similar calibration (ECE = 0.22) but a slightly lower AUROC (0.55) and success rate (48%). For **Hard tasks**, performance further degrades across all policies. CoT and P&S maintain relative superiority, with CoT achieving an ECE of 0.31, AUROC of 0.68, and a success rate of 22%, while P&S records slightly worse calibration (ECE = 0.32) and the highest AUROC (0.71) but tied for a success rate of 18%.

Table 7: **ECE, AUROC, and Success Rates for Different Task Difficulties.** Lower ECE and higher AUROC/Success Rates indicate better performance.

Task	Strategies	ECE (↓)	AUROC (↑)	Success Rate (↑)
Easy	Vanilla	0.26	0.76	84%
	Self-Intervention	0.26	0.76	92%
	CoT	**0.11**	0.78	**94%**
	P&S	0.12	**0.80**	82%
	Top-K	0.32	0.72	74%
Medium	Vanilla	0.35	0.54	52%
	Self-Intervention	0.35	0.51	44%
	CoT	**0.22**	**0.58**	**54%**
	P&S	0.22	0.55	48%
	Top-K	0.40	0.47	32%
Hard	Vanilla	0.33	0.58	17%
	Self-Intervention	0.35	0.52	12%
	CoT	**0.31**	0.68	18%
	P&S	0.32	**0.71**	**18%**
	Top-K	0.41	0.49	8%

These results show that task difficulty introduces compounding uncertainty during planning and execution, which simple or unstructured methods fail to capture. In contrast, structured strategies like CoT and P&S remain more resilient, even as absolute performance drops. This underscores the need for execution-aware, structured reasoning to maintain meaningful confidence estimates under challenging conditions. Through VOiCE, we are able to systematically surface these effects by disentangling reasoning strategies and analyzing their calibration dynamics across difficulty levels. This opens several avenues for future work, such as new adaptive selection methods to allow agents to dynamically choose reasoning styles based on task difficulty or environment feedback, or integrating learning-based policy refinement to improve reasoning robustness over long horizons.

E.2 Does Expressing Uncertainty Affect Task Success Rates?

Embodied agents frequently integrate multiple language models working in tandem, where later modules often depend on the outputs of earlier stages. Incorporating expressions of uncertainty alongside predicted outputs could significantly influence the decision-making pipeline. We investigate whether

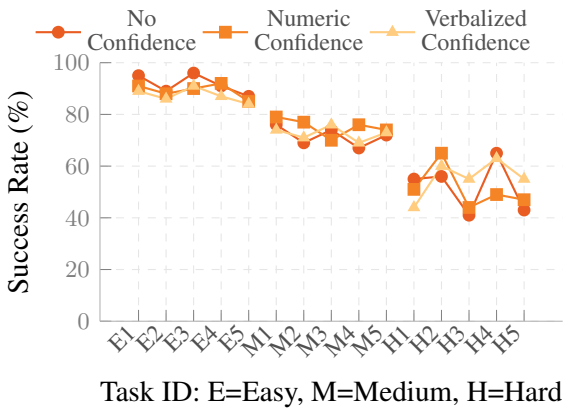

Figure 7: **Success Rates with Different Confidence Expressions.** across task instances (E: Easy, M: Medium, H: Hard) under three confidence elicitation settings. Verbalized confidence shows no consistent performance degradation, demonstrating that expressing uncertainty modulates behavior without harming task success.

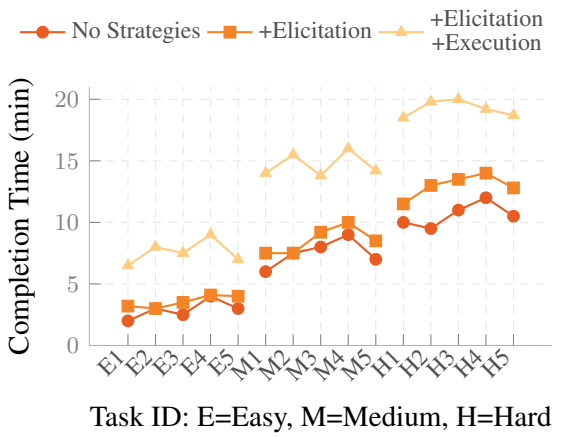

Figure 8: **Task Completion Time Across Introspective Strategies.** Task completion time remains comparable to the baseline when only confidence elicitation is introduced, indicating minimal overhead. However, incorporating execution-time reasoning leads to longer completion times, due to the cost of simulating alternative action paths and scenarios for robust uncertainty estimation.

specific uncertainty expressions produce consistent effects on the model's performance.

We conduct a controlled analysis using the GPT-4V agent under the vanilla elicitation strategy. Each output is generated under one of three conditions: (1) no confidence appended, (2) numeric confidence score (*e.g.*, "I am 90% confident..."), and (3) verbalized confidence (*e.g.*, "I am fairly sure...", "I am not certain..."). Success rates are recorded across 15 tasks spanning three difficulty levels (5 tasks per category). Average success rate across 3 trials is depicted in Figure 7.

While performance varies naturally by task, we observe no systematic degradation due to the integration of verbalized confidence elicitation. In easy tasks, outputs without confidence statements occasionally perform best, likely due to more assertive completions. However, in medium and hard tasks, all three conditions perform similarly, with verbalized confidence sometimes yielding the highest success. These results suggest that expressing uncertainty, even in verbal form, does not impair the model's ability to act effectively.

On the contrary, in more ambiguous settings, verbal confidence may enhance alignment between internal uncertainty and task behavior. This supports the broader argument of VOiCE in that verbal uncertainty can serve as a transparent self-assessment mechanism without sacrificing downstream task utility. Future work could build on these insights to design adaptive policies that dynamically adjust agent behavior in response to self-assessed verbalized confidence signals. Such policies would enable agents to behave more transparently and exhibit reasoning patterns that better align with human-like caution and deliberation in open-ended, language-driven settings.

E.3 Does Expressing Uncertainty Change Task Solving Efficiency?

We design a comparison to evaluate how structured confidence elicitation influences agent efficiency during decision-making. Using the GPT-4V agent, we run 15 embodied tasks across three conditions of increasing introspective structure: (1) a baseline with no elicitation or execution strategy (vanilla), (2) a confidence-elicitation-only setup where the agent explicitly reflects on uncertainty before responding by employing CoT as the elicitation strategy, and (3) a setup that incorporates both CoT elicitation and Action Sampling execution strategies. Each run measures task completion time in minutes, starting from task presentation to the final generated action, capturing the agent's efficiency.

Figure 8 shows that incorporating confidence elicitation has minimal impact on task-solving efficiency across all difficulty levels, resulting in completion times that closely track the baseline. Results indicate that light-weight introspective reasoning can be integrated with negligible cost. In contrast, incorporating reflective execution strategies increases completion time. This is expected, as reflective execution requires the agent to reason over multiple alternative action paths, reinterpret

scenarios, and simulate counterfactual outcomes. These results further suggest that selectively invoking deeper reasoning where uncertainty is high is a promising future direction.

E.4 Can Verbalized Confidence Reach Likelihood-Based Uncertainty Estimates?

We compare our VOiCE verbalized confidence estimation framework with four token-level traditional likelihood-based uncertainty estimation methods. Specifically, we extract uncertainty signals from the vanilla generations of LLaVA-1.5, without applying any of our strategies, by analyzing its token-level decoding behavior. Likelihood-based methods include: (1) *sequence probability* (seq-prob), which aggregates the joint likelihood of the full output; (2) *length-normalized sequence probability* (len-norm-prob), which mitigates sequence-length bias; (3) *key-token probability* (token-prob), which isolates the probability of task-critical tokens (*e.g.*, final answers); and (4) *average token entropy*, which captures the model's uncertainty during generation by averaging the entropy of its predicted token distribution at each decoding step. For this experiment, we use VOiCE's verbalized confidence derived from CoT. We compute the average across three embodied tasks in each difficulty category using standard calibration ECE and AUROC metrics.

As shown in Table 8, VOiCE achieves calibration and failure prediction performance that is competitive with likelihood-based methods on easy tasks, even surpassing them in AUROC (0.74 vs. next-best 0.69). On medium tasks, the gap remains modest, with VOiCE trailing the best logprob-based method by 0.08 in ECE and 0.06 in AUROC. However, on hard tasks, results indicate that verbalized confidence degrades more quickly than likelihood-based estimates as task complexity increases. Verbalized confidence is grounded in the agent's own reasoning trace, making it more sensitive to reasoning errors and compounding uncertainty in challenging settings. While the trend is not surprising, the fact that VOiCE matches or exceeds traditional methods on easier tasks, with the overall performance gap between the two paradigms remaining relatively small, demonstrates its promise as an interpretable alternative. These results suggest that introspective confidence could form the basis for robust embodied uncertainty estimation, but further research is needed to improve its resilience to reasoning errors in difficult tasks, *e.g.*, by investigating more sophisticated training-time methods that better align introspective self-assessment with actual task success. Additionally, despite their relative advantages, none of the evaluated methods demonstrate fully calibrated uncertainty estimation, particularly in terms of discriminative power. AUROC values across conditions frequently hover near 0.5–0.6, indicating limited ability to separate correct from incorrect predictions. These findings highlight a broader challenge in uncertainty estimation: current methods, whether likelihood-based or language-based, struggle to capture semantic ambiguity and calibrate agent confidence, a capability that is especially critical in long-horizon embodied tasks.

F Failure Case Analysis

To better understand the limitations of our approach, we conduct a qualitative analysis of failure cases across various task difficulties. Table 9 summarizes representative failure scenarios observed during evaluation, categorized by error type. We found that failures generally fall into two broad categories: (1) miscalibrated confidence (where the agent expresses high confidence despite poor grounding), and (2) reflective execution deficiencies (where correct reasoning is generated but fails to translate into effective behavior and vice versa, such as incorrect backtracking or poor spatial grounding). These observations underscore the complementary roles of confidence elicitation and execution strategies and suggest that improving their alignment is critical for robust performance.

G Additional Related Work

Confidence Elicitation and Uncertainty Quantification. Uncertainty estimation has been extensively studied in traditional machine learning, with well-established methods including Bayesian Neural Networks (BNNs) (Blundell et al., 2015), Monte Carlo Dropout (Gal and Ghahramani, 2016), and Conformal Prediction (Kumar et al., 2023; Ye et al., 2024). These techniques rely on sampling-based methods or statistical guarantees to quantify model uncertainty, offering robust coverage intervals and calibrated confidence estimates. Recent work extends this to unsupervised methods leveraging entropy minimization (Malinin and Gales, 2021), semantic parsing (Lin et al., 2022c), graph-based features (Kuhn et al., 2023), and hidden state embeddings(Su et al., 2024; Chen et al., 2024a) to create uncertainty metrics. However, these ap-

Table 8: **Comparison of Likelihood-based and Verbalized Confidence Estimation Methods Across Task Difficulties.** Likelihood-based methods use token-level statistics, while verbalized confidence VOiCE relies on eliciting introspective confidence via CoT reasoning.

Task	Success Rate	Method	ECE (↓)	AUROC (↑)
Easy	**89%**	Seq-Prob	**0.11**	0.65
		Len-Norm-Prob	0.47	0.65
		Token-Prob	0.42	0.69
		Avg Entropy	0.39	0.67
		VOiCE	**0.11**	**0.74**
Medium	**48%**	Seq-Prob	**0.19**	0.64
		Len-Norm-Prob	0.28	**0.65**
		Token-Prob	0.24	0.61
		Avg Entropy	0.34	0.58
		VOiCE	0.27	0.59
Hard	**12%**	Seq-Prob	**0.27**	0.55
		Len-Norm-Prob	0.33	**0.56**
		Token-Prob	0.29	0.54
		Avg Entropy	0.35	0.51
		VOiCE	0.37	0.50

Table 9: **Representative Failure Cases Across Tasks**, highlighting mismatches between verbalized confidence and actual performance.

Failure Type	Example Scenario	Observed Issue
Overconfident Reasoning	Agent states "I'm sure the diamond is behind the wall"	Diamond is elsewhere; agent fails to backtrack
Underconfident Execution	Agent states "Not sure if jumping is safe"	Skips jump even though it's the correct move
Incorrect Spatial Reference	Refers to "leftmost block"	Misinterprets left vs. right due to camera angle
Overlooked Alternative Path	"I can only go forward"	Misses open side path due to lack of scenario reinterpretation
Contradictory Execution	"I'm 25% sure about stepping forward" – Steps forward anyway	Inconsistent action given uncertainty signals

proaches rely on model internals, such as logits or predictive distributions (Kumar et al., 2023; Chen et al., 2024b). Alternative approaches have been proposed, including estimating uncertainty by directly querying models for confidence scores after generating responses (Xiong et al., 2024; Kadavath et al., 2022a; Lin et al., 2022b; Mielke et al., 2022; Chen and Mueller, 2024).

Despite these advancements, existing methods are not designed for embodied tasks, where confidence elicitation must address the challenges of multimodal perception, hierarchical reasoning, and planning across various open-ended tasks, as well as non-deterministic interactions. VOiCE adopts a verbalized reasoning path that enables agents to articulate their confidence dynamically, even when internal model representations are not exposed. This structured verbalization allows VOiCE to surface uncertainties as part of its reasoning process, facilitating self-aware exploration without explicit access to posterior distributions.

LLM-based Embodied Agents. With the advent of language models, leveraging their reasoning and planning abilities to empower embodied agents has become quintessential (Huang et al., 2023; Yao et al., 2023; Chen et al., 2023; Zhang et al., 2024a; Shinn et al., 2024; Christianos et al., 2023). In the meantime, Minecraft's open-ended nature with its adaptable mechanics and open-ended challenges makes it a compelling benchmark for embedding reasoning and planning capabilities into language model-driven embodied agents (Wang et al., 2023a,c; Zhu et al., 2023).

Recent works leverage pre-trained language models to control agents by generating continuous operation instructions or executable policies. Some approaches (Zhu et al., 2023; Wang et al., 2023c) directly utilize scene data from simulation platforms like MineDojo (Fan et al., 2022) and MineRL (Guss et al., 2019), while others (Qin et al., 2024) rely on Vision-Language Models (VLMs) for perception. However, because language models are used in various roles, such as planners, critics, or perceivers, errors and inaccuracies often arise at different process stages (Guo et al., 2024; Driess et al., 2023). These challenges underscore the need

for frameworks that can systematically surface and recalibrate embodied uncertainty during task completion, which we aim to address by designing a unified approach for verbalized confidence elicitation in embodied settings.

Uncertainty in Embodied Models. Estimating uncertainty in embodied AI is well-studied in robot learning and reinforcement learning (Wang and Zou, 2021; Ghasemipour et al., 2022; He et al., 2023; Huang et al., 2019; Jin et al., 2023). Bayesian methods such as Bayesian Reinforcement Learning (Bayesian RL) (Ghavamzadeh et al., 2015) and Bayesian MDPs (Delage and Mannor, 2010; Petrik and Russel, 2019) model agent-environment interactions by maintaining probabilistic beliefs over state transitions to inform exploration. These methods optimize for expected information gain during agent-environment interactions, refining confidence estimates as the agent gathers more data.

While effective in structured environments, these strategies require state transition probabilities. Additionally, the free-form nature of their outputs further complicates the application of traditional methods. Therefore, uncertainty estimation remains a challenge for language models (Tian et al., 2023; Groot and Valdenegro Toro, 2024; Zhang et al., 2024b). While recent efforts have sought to quantify and mitigate uncertainty in static tasks such as image classification, question answering, *etc.* (Sagar et al., 2024; Zhao et al., 2021; Kadavath et al., 2022b; Wang et al., 2022), the problem is further compounded in embodied AI settings, where agents must reason and act in dynamic multimodal environments (Ren et al., 2024; Shen and Lourentzou, 2025). Our work introduces a verbalized confidence elicitation framework for open-ended, multimodal embodied environments, enabling agents to articulate uncertainty and introspectively adjust reasoning during interactive tasks.

H Limitations

While VOiCE demonstrates improved confidence calibration and failure prediction through structured reasoning, several limitations remain. First, while VOiCE currently relies on proprietary VLMs (*e.g.*, GPT-4V), this choice reflects a pragmatic design decision. These models remain the most capable and widely deployed tools for real-world embodied applications. Nonetheless, their closed-source nature limits reproducibility and constrains the transparency of internal model behavior. As open-source multimodal models continue to improve, future work could transition to fully auditable systems and explore model-level alignment between verbalized confidence and latent predictive signals. Additionally, verbalized confidence is inherently more subjective and less precise than token-level probabilistic signals, making it harder to calibrate under fine-grained uncertainty conditions. Future extensions could align confidence expressions with fully grounded referents to support more transparent reasoning, verifiability, and downstream integration with perception modules or human collaborators. Finally, our evaluation is limited to Minecraft as a testbed, and generalization to more diverse embodied environments remains an open question. Although this platform supports open-ended and compositional tasks, it does not reflect the full complexity of real-world physics, embodiment constraints, or sensor noise. Addressing these limitations requires advances in adaptive reasoning, efficient introspection, and broader validation across tasks and domains.

I Impact Statement

This work advances the field of Embodied AI by introducing a framework for verbalized confidence elicitation and introspective reasoning in multimodal, open-ended environments. By equipping agents with the ability to express and reason about their own uncertainty, VOiCE improves confidence calibration and failure prediction, leading to more adaptive and reliable agent behavior during task execution. These capabilities have the potential to support safer AI deployment in real-world settings such as assistive robotics, educational tools, and human-AI collaboration, where transparency and accurate self-monitoring are essential for building user trust and ensuring safe operation. Moreover, our framework offers a model-agnostic and interpretable alternative to traditional uncertainty estimation Beyond numeric likelihoods, verbalized introspection provides interpretable explanations of uncertainty that can be directly audited by humans and integrated into multi-step reasoning and decision-making. These qualities make it particularly valuable for scenarios where uncertainty must be communicated to users or combined with high-level planning processes.

However, the reliance on large pre-trained language models raises important considerations. These include the environmental cost of compute-

intensive reasoning, potential biases embedded in foundation models, and the challenge of ensuring reliable behavior in high-stakes or safety-critical applications. While our work focuses on improving the agent's introspective capabilities, future research should explore more energy-efficient reasoning strategies, investigate methods for uncertainty alignment with human expectations, and develop safeguards for deploying such systems in ethically sensitive domains.