PROBABILISTIC-SYMBOLIC INTENT DISAMBIGUATION FOR ROBUST EMBODIED CONTROL UNDER AMBIGUOUS SPEECH

Anonymous authors
Paper under double-blind review

ABSTRACT

Speech provides a natural and flexible interface for embodied control, yet its inherent ambiguity challenges accurate task understanding and reliable execution. Meanwhile, embodied interaction continuously provides new environmental information that can further constrain plausible interpretations of the original spoken instruction. Robust speech-conditioned embodied control therefore requires an agent to continuously refine its intent understanding during execution and actively explore discriminative environmental information when competing intents lead to action disagreement. To this end, we propose a probabilistic-symbolic intent disambiguation framework, where a probabilistic measure quantifies how speech ambiguity affects action decisions, while symbolic reasoning localizes and resolves the critical intent differences underlying action disagreement through embodied interaction. Specifically, we introduce speech-induced action uncertainty to quantify how uncertainty in speech interpretation propagates to the current action decision, providing a criterion for when disambiguation is needed. We then develop a unified symbolic intent–world modeling method that maps candidate speech intents and observations into a shared symbolic space for action reasoning and intent disambiguation. Building on this representation, we introduce world-constraint alignment for speech reinterpretation, which localizes intent differences to guide targeted exploration and induces world constraints from interaction feedback to refine speech interpretations. We further introduce SPEAR-Bench, a benchmark that systematically varies perceptual, paralinguistic, and content ambiguity to evaluate task understanding and embodied control under ambiguous speech. Experiments on SPEAR-Bench and REI-Bench show that our method consistently improves task success while reducing interaction cost compared with strong baselines for embodied control under ambiguous speech.

1 INTRODUCTION

Embodied speech control allows users to interact with agents through natural and flexible spoken instructions, yet this flexibility also introduces inherent ambiguity in task interpretation. Noise and pronunciation variations can obscure task-relevant linguistic content (Ying et al., 2026); paralinguistic cues such as tone and pauses can alter how an utterance is interpreted (Sasu et al., 2025); and references, omissions, or self-corrections can leave the intended task ambiguous (Jiang et al., 2026). Unlike static language understanding, embodied control unfolds through interaction, where observations acquired during execution can provide additional information for constraining plausible interpretations of the original instruction (Choi et al., 2025). Consequently, speech ambiguity need not be fully resolved before acting: an agent can refine its intent understanding as execution proceeds and seek discriminative environmental information when unresolved interpretations begin to affect action decisions. Robust embodied speech control therefore requires task execution and intent disambiguation to be coordinated throughout interaction.

Existing work characterizes uncertainty over different components of agent decision making (Oh et al., 2026), including entropy over action distributions (Ding et al., 2026), probabilistic modeling of environmental assumptions under partial observability (Seo et al., 2026), and uncertainty over task parameters combined with information value to determine whether clarification is needed (Suri et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2026). However, these measures quantify uncertainty in actions, environments, or task parameters themselves, rather than answering a question that is particularly important for embodied control under ambiguous speech: to what extent does the current action decision depend on unresolved interpretations of the spoken instruction? Similar levels of intent uncertainty can induce nearly identical actions in one interaction state but substantial action disagreement in another; conversely, high action uncertainty may persist even under a fixed intent when multiple actions are plausible. We therefore need a decision-oriented measure of how unresolved speech interpretation affects the current action decision, providing a criterion for when speech ambiguity requires resolution.

To resolve intent uncertainty induced by ambiguous speech, existing approaches primarily exploit paralinguistic cues in the original speech to improve intent inference (Sasu et al., 2025), or use observed environmental states and collaborative behavior to reinterpret noisy spoken instructions (Ying et al., 2026). More broadly, task understanding and interaction are coupled in embodied control: environmental observations can refine task interpretation, which in turn guides subsequent interaction (Liu et al., 2025; Xiang et al., 2026b). When the original speech and available context remain insufficient to resolve the ambiguity, the key challenge is to identify which exploratory actions can best distinguish competing task intents and what constraints on speech reinterpretation can be derived from the resulting interaction feedback.

We address these challenges with a probabilistic-symbolic intent disambiguation framework for accurate task understanding and reliable embodied execution under ambiguous speech. The probabilistic component quantifies decision-relevant uncertainty, while the symbolic component supports intent disambiguation through embodied interaction. To measure when speech ambiguity begins to affect action decisions, we introduce speech-induced action uncertainty, which uses mutual information to quantify the extent to which uncertainty in speech intent influences uncertainty in the current action choice. When task intents induce different actions, effective disambiguation requires identifying the intent differences that matter for the current decision. We use a shared symbolic representation of intents and world states to localize these differences and guide targeted exploration. Building on this representation, we introduce world-constraint alignment for speech reinterpretation, which derives environmental constraints from exploration and execution feedback to refine subsequent speech interpretation and action planning. To systematically evaluate these capabilities, we introduce SPEAR-Bench, which covers diverse forms of speech ambiguity across perceptual, paralinguistic, and content dimensions, providing a benchmark for studying embodied control under ambiguous speech. Extensive comparisons and analyses on SPEAR-Bench and REI-Bench demonstrate that our method consistently improves task success while requiring fewer interaction steps.

Our contributions are threefold:

- We present speech-induced action uncertainty, a decision-oriented measure based on mutual information that quantifies how uncertainty over speech intent affects the current action decision, providing a principled criterion for when disambiguation is needed.
- We develop a world-constraint alignment framework for speech reinterpretation, using decision-relevant intent differences to guide exploration and inducing environmental constraints from interaction feedback to refine speech interpretation and action planning.
- We construct SPEAR-Bench, a new benchmark that studies how speech ambiguity arising from three different sources affects embodied control. Extensive comparisons and analyses on SPEAR-Bench and REI-Bench show that our method consistently improves task success while requiring fewer interaction steps.

2 Related Work

Uncertainty Quantification for Embodied Planning. Prior work has quantified uncertainty in agent decision-making from several perspectives. KnowNo uses conformal prediction to construct action prediction sets and determine when human assistance is needed (Ren et al., 2023), while AdaNav uses action entropy to guide adaptive reasoning (Ding et al., 2026). PCE organizes latent environmental assumptions into a decision tree and evaluates candidate actions using scenario likelihood, goal-directed gain, and execution cost (Seo et al., 2026). For tool-using agents, SAGE represents uncertainty over task parameters and uses the expected value of perfect information to assess clarification questions (Suri et al., 2026). ILS uses calibration error to select informative

clarification queries and improve reliability in interactive LLMs (Li et al., 2026). These approaches use uncertainty estimates to guide assistance, reasoning, planning, or information acquisition. Our work specifically quantifies the dependence between candidate speech-induced task intents and their induced action distributions under the current interaction context, providing a criterion for when further intent disambiguation is needed.

Interactive Disambiguation for Embodied Instructions. Recent work addresses ambiguous embodied instructions by leveraging environmental context, speech cues, and interaction feedback. GraphCoT-VLA combines structured environment representations with reasoning to interpret ambiguous manipulation instructions and generate actions (Huang et al., 2026). SIFToM uses environmental states and collaborative behavior to infer intended tasks from noisy spoken instructions (Ying et al., 2026), while prosody-based methods exploit paralinguistic cues for referent and intent disambiguation (Sasu et al., 2025). In contrast, our framework localizes the intent differences responsible for current action disagreement, uses them to guide targeted exploration, and converts interaction feedback into world constraints for subsequent speech reinterpretation and planning.

Benchmarks for Ambiguous Embodied Instructions. Existing benchmarks study ambiguity in embodied instructions from different perspectives. ULN examines language underspecification in navigation (Feng et al., 2022), while AmbiK focuses on ambiguous task descriptions (Ivanova et al., 2025), and REI-Bench studies vague referring expressions in embodied planning (Jiang et al., 2026). As these benchmarks operate on textual instructions, they overlook the effects of speech-specific factors, such as acoustic perception and prosody, on task understanding and embodied control. Recent speech-oriented studies investigate how noise and pronunciation variation affect robotic task understanding (Ying et al., 2026), or how prosodic cues support intent disambiguation (Sasu et al., 2025), but primarily evaluate speech understanding or intent inference rather than their downstream effects during interactive embodied control. SPEAR-Bench instead controls perceptual, paralinguistic, and content ambiguity at multiple levels over matched tasks and environments, enabling controlled evaluation of how different sources of speech ambiguity and their combinations affect intent understanding, action decisions, and task execution.

3 Problem Formulation and Uncertainty Quantification

Ambiguous speech can leave the intended task uncertain, and this uncertainty may propagate to action decisions. Embodied agents must therefore infer the user's intended task from an ambiguous spoken instruction and execute the corresponding task in the environment. We first formalize this problem by decomposing speech-conditioned control into intent inference and intent-conditioned planning. We then quantify when unresolved intent uncertainty affects action selection.

3.1 Embodied Speech Control under Intent Uncertainty

We consider a partially observed embodied environment with state space $\mathcal{S}$, action space $\mathcal{A}$, and transition dynamics $\mathcal{T}$. Given a spoken instruction x, the agent interacts with the environment through actions $a_i \in \mathcal{A}$ and observations o_i, which provide partial information about the current world state s_i. The interaction context at step i is $h_i = (o_0, a_0, o_1, \ldots, a_{i-1}, o_i)$. The instruction x expresses a fixed but unobserved task intent $z^\star \in \mathcal{Z}$, where each intent z specifies an embodied goal $g(z)$. Ambiguous speech can support multiple plausible intents, coupling task interpretation with action planning. We introduce a latent intent variable Z, assumed to capture all task-relevant information in x for action selection conditioned on the interaction history, yielding the decomposition

$$\pi(a_i \mid x, h_i) = \sum_{z \in \mathcal{Z}} P(z \mid x, h_i)\pi(a_i \mid z, h_i), \tag{1}$$

where $P(z \mid x, h_i)$ is the agent's belief over possible task intents, and $\pi(a_i \mid z, h_i)$ is the action policy conditioned on intent z and the interaction history. Marginalizing over Z allows action planning without first committing to a single interpretation. As interaction proceeds, new observations update the belief over Z, thereby influencing subsequent action selection.

Embodied speech control requires not only successful task execution but also accurate inference of the task intent. For a trajectory $\tau = (s_0, a_0, s_1, \ldots, a_{T-1}, s_T)$, let $\hat{z}_T = \arg\max_{z \in \mathcal{Z}} P(z \mid x, h_T)$ denote the agent's final intent estimate. We define successful embodied speech control as

$$\mathrm{Succ}(\tau, \hat{z}_T; z^\star) = \mathbf{1}[g(\hat{z}_T) = g(z^\star) \ \wedge \ s_T \models g(z^\star)]\,. \tag{2}$$

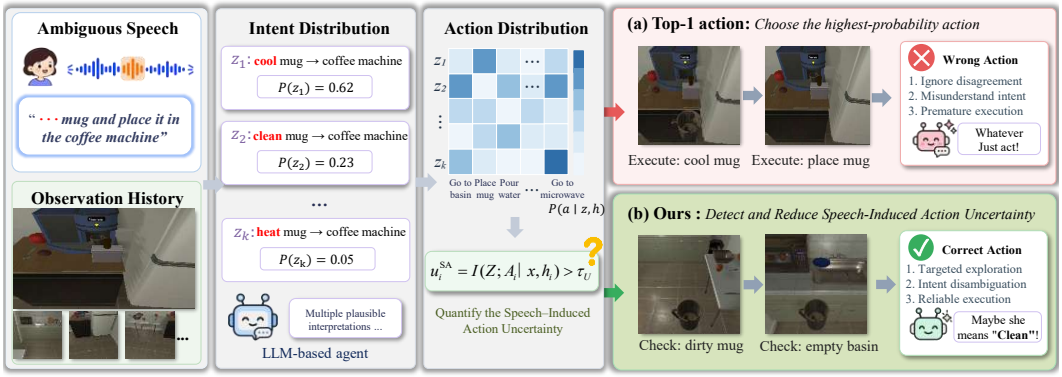


Figure 1: Quantification of speech-induced action uncertainty. Ambiguous speech and observation history induce multiple candidate intents and corresponding action distributions. U_i^{SA} quantifies their action disagreement: (a) selecting the top-1 intent can lead to an incorrect action; (b) $U_i^{\mathrm{SA}} > \tau_U$ detects substantial disagreement and triggers exploration and disambiguation to reduce it.

Because the true intent $z^\star$ and its goal $g(z^\star)$ are not available to the agent, it must infer the intended task from the spoken instruction and interaction history during execution. In practice, both accurate intent recovery and task completion should be achieved with low interaction cost.

3.2 Quantifying Speech-Induced Action Uncertainty

Equation 1 connects uncertainty over task intent to the distribution over next actions. However, intent uncertainty does not indicate whether it affects the current action decision: different plausible intents may induce similar distributions over the next action. We therefore quantify how much of the current action uncertainty is associated with unresolved uncertainty over the task intent. Let A_i denote the next-action random variable under the mixed policy $\pi(\cdot \mid x, h_i)$, with entropy $H(A_i \mid x, h_i) = -\sum_{a \in \mathcal{A}} \pi(a \mid x, h_i) \log \pi(a \mid x, h_i)$. Conditioning on the task intent leaves the residual action uncertainty $H(A_i \mid Z, x, h_i) = \sum_{z \in \mathcal{Z}} P(z \mid x, h_i) H(A_i \mid z, h_i)$, where $H(A_i \mid z, h_i)$ is the entropy of the intent-conditioned policy $\pi(\cdot \mid z, h_i)$. We define *speech-induced action uncertainty* as the conditional mutual information between task intent Z and the next action A_i:

$$\mathcal{U}_i^{\mathrm{SA}} = I(Z; A_i \mid x, h_i) = H(A_i \mid x, h_i) - H(A_i \mid Z, x, h_i). \tag{3}$$

Equivalently, $\mathcal{U}_i^{\mathrm{SA}} = \sum_{z \in \mathcal{Z}} P(z \mid x, h_i) D_{\mathrm{KL}}(\pi(\cdot \mid z, h_i) \| \pi(\cdot \mid x, h_i))$, which captures the belief-weighted disagreement among the action distributions induced by different intents. A larger $\mathcal{U}_i^{\mathrm{SA}}$ therefore indicates greater speech-induced action uncertainty, reflecting stronger disagreement among plausible intents in the current action decision.

In practice, we treat $\mathcal{U}_i^{\mathrm{SA}} > \tau_U$ as indicating that speech-induced action uncertainty has resulted in substantial action disagreement. The threshold τ_U is calibrated using a held-out set of unambiguous spoken instructions as a reference level. As illustrated in Figure 1, this criterion allows the agent to trigger intent disambiguation when speech ambiguity materially affects action selection; without such detection, the agent may commit to an action inconsistent with the intended task, leading to task failure. The detailed calibration procedure is provided in Appendix A.

4 World-Constraint Alignment through Embodied Exploration for Reducing Speech-Induced Uncertainty

When competing speech interpretations induce different actions, resolving the ambiguity requires identifying the environmental information that distinguishes them. We propose a world-constraint alignment framework based on embodied exploration, which localizes decision-relevant intent differences in a shared symbolic space and uses them to guide targeted exploration. Interaction feedback is converted into world constraints that refine subsequent speech interpretation and planning, progressively reducing speech-induced action uncertainty while enabling reliable task execution.

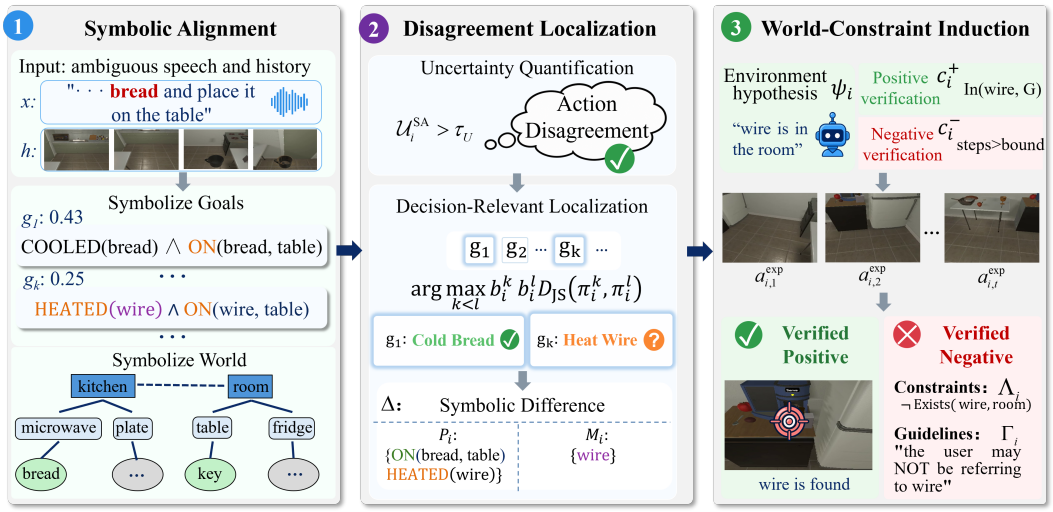


Figure 2: Overview of our method. The framework aligns candidate intents with the observed world and localizes decision-relevant symbolic differences under action disagreement. These differences guide verifiable exploration, where positive outcomes provide valuable world information and negative outcomes induce constraints for subsequent speech reinterpretation.

4.1 Symbolic Intent-World Alignment

Given the spoken instruction x and the current observation history $h_i^o = (o_0, o_1, \ldots, o_i)$, we use a speech-capable LLM to maintain up to K candidate task intents with normalized beliefs, denoted by $\mathcal{B}_i = \{(z_k, b_i^k)\}_{k=1}^{K_i}$, where $K_i \le K$ and $\sum_{k=1}^{K_i} b_i^k = 1$. To make candidate intents directly comparable with the observed world, we map both into a shared symbolic space of typed entities and predicates. Each intent z_k is represented as a symbolic goal $g_k = \bigwedge_{j=1}^{m_k} p_j(v_{j1}, \ldots, v_{jn_j})$, where each predicate specifies a required entity state or relation. For example, "put the heated bread on the plate" becomes $\textsc{Heated}(bread) \wedge \textsc{On}(bread, plate)$. We represent the observed world as $G_i = (V_i, F_i)$, where each $v \in V_i$ denotes a concrete observed entity and F_i contains all symbolic facts extracted from observations available up to step i, including entity states and relations. For example, "a cooled bread is in the fridge" yields $\textsc{Cooled}(bread_1)$ and $\textsc{In}(bread_1, fridge_1)$ in F_i. The representation is initialized as $G_0 = \mathrm{Build}(o_0)$ and incrementally updated as $G_{i+1} = \mathrm{Update}(G_i, o_{i+1})$, where $\mathrm{Build}(\cdot)$ and $\mathrm{Update}(\cdot,\cdot)$ use an LLM-based symbolic parser to extract newly observed entities and symbolic facts.

By aligning each symbolic goal g_k with the observed world G_i, we characterize its current grounding and completion status as

$$\Phi(g_k, G_i) = \big(M_i(g_k), P_i(g_k)\big). \tag{4}$$

The unresolved entity set $M_i(g_k) = \{v \in g_k \mid v \notin V_i\}$ contains goal entities not yet grounded to any observed entity, where $v \notin V_i$ denotes the absence of a matching entity in the observation. The unverified goal-condition set $P_i(g_k) = \{p \in g_k \mid p \notin F_i\}$ contains goal conditions not yet verified by the observed world, where $p \notin F_i$ indicates that the corresponding grounded predicate is not supported by the current facts. Together, these residual sets summarize what remains unresolved for each candidate goal and provide the symbolic basis for subsequent disagreement localization.

4.2 Disagreement Localization and World-Constraint Acquisition

Action Disagreement Detection. For each candidate intent z_k, the planning LLM produces an intent-conditioned next-action distribution based on its symbolic goal and current alignment with the observed world:

$$\pi_i^k(a) = \pi_\theta^{\mathrm{act}}\left(a \mid h_i, g_k, M_i(g_k), P_i(g_k)\right). \tag{5}$$

Using $\{\pi_i^k\}_{k=1}^{K_i}$ and the intent beliefs $\{b_i^k\}_{k=1}^{K_i}$, we compute the speech-induced action uncertainty $\mathcal{U}_i^{\mathrm{SA}}$ defined in Eq. equation 3. When $\mathcal{U}_i^{\mathrm{SA}} \le \tau_U$, intent uncertainty has limited effect on the current

action decision and execution continues normally; otherwise, substantial action disagreement is detected, prompting the agent to localize decision-relevant symbolic differences and explore the environment for discriminative information.

Decision-Relevant Symbolic Localization. When action disagreement is detected, we first identify the pair of candidate goals that is both highly probable and divergent in its induced action distributions as $(k^*, l^*) = \arg\max_{k<l} b_i^k b_i^l D_{\mathrm{JS}}(\pi_i^k, \pi_i^l)$, where D_{JS} denotes the Jensen–Shannon divergence between the two intent-conditioned action distributions (Lin, 1991). We then localize their unresolved symbolic differences as

$$\Delta_i = \big(M_i(g_{k^*}) \triangle M_i(g_{l^*})\big) \cup \big(P_i(g_{k^*}) \triangle P_i(g_{l^*})\big), \tag{6}$$

where $\triangle$ denotes symmetric difference. This removes unresolved entities and conditions shared by both goals and retains the symbolic distinctions between them. The resulting Δ_i defines the symbolic distinctions to be resolved through subsequent environmental exploration.

Disagreement-Guided Exploration and World-Constraint Acquisition. Given the localized symbolic differences Δ_i, we perform targeted exploration to acquire discriminative environmental information and convert negative exploration outcomes into world constraints. Specifically, the agent first constructs an environmental hypothesis with positive and negative verification conditions:

$$(\psi_i, c_i^+, c_i^-) \sim P_\theta(\cdot \mid h_i, \Delta_i)\,, \tag{7}$$

where ψ_i denotes the exploration hypothesis, and c_i^+ and c_i^- specify the conditions under which the hypothesis is supported or contradicted, respectively. For example, for the hypothesis "a loaf of bread exists in the kitchen," c_i^+ is satisfied once a matching bread is observed, whereas c_i^- is satisfied after all relevant kitchen locations have been inspected without finding one. The agent then interacts with the environment to verify ψ_i. At exploration step t, it selects $a_{i,t}^{\mathrm{exp}} \sim \pi_\theta^{\mathrm{exp}}(\cdot \mid \psi_i, h_{i,t})$, executes the action to obtain $o_{i,t+1}$, and then incrementally updates the symbolic world as $G_{i,t+1} = \mathrm{Update}(G_{i,t}, o_{i,t+1})$. The updated world $G_{i,t+1}$ is evaluated against c_i^+ and c_i^-, and exploration continues until either condition is satisfied or the maximum exploration budget is reached.

When c_i^+ is satisfied, the agent acquires valuable information for the next round of intent interpretation. When c_i^- is satisfied, the negative outcome is converted into two complementary constraints: a symbolic world constraint λ_i and a language-level interpretation constraint γ_i. For example, for the hypothesis "a loaf of bread exists in the kitchen," if the kitchen has been sufficiently explored without finding bread, we obtain $\lambda_i : \neg\,\mathrm{Exists}(\mathit{bread}, \mathit{kitchen})$ and the corresponding interpretation guideline γ_i : "no bread was found in the kitchen, so the user may not be referring to bread." We accumulate the two types of constraints as

$$\Lambda_i^+ = \Lambda_i \cup \{\lambda_i\}, \qquad \Gamma_i^+ = \Gamma_i \cup \{\gamma_i\}, \tag{8}$$

where Λ_i and Γ_i denote the accumulated symbolic world constraints and language-level interpretation constraints, respectively.

4.3 Constraint-Guided Speech Reinterpretation

After acquiring new observations or constraints through exploration, we revisit the original spoken instruction x and update its intent interpretation accordingly. Conditioned on the original instruction x, intent belief $\mathcal{B}_i$, interaction history h_i, and the accumulated constraints Γ_i^+ and Λ_i^+, we obtain

$$\widetilde{\mathcal{B}}_i = \mathcal{R}_\theta\left(x, \mathcal{B}_i, h_i, \Gamma_i^+\right) = \left\{(\tilde{z}_k, \tilde{b}_i^k)\right\}_{k=1}^{\widetilde{K}_i}, \qquad \mathcal{B}_i' = \mathcal{C}_{\Lambda_i^+}\left(\widetilde{\mathcal{B}}_i\right). \tag{9}$$

Here, $\mathcal{R}_\theta$ is a speech-capable LLM that regenerates candidate intents under the guidance of Γ_i^+. The symbolic constraints Λ_i^+ are then enforced through $\mathcal{C}_{\Lambda_i^+}$ to remove candidates inconsistent with the explored world and renormalize the remaining belief weights.

Overall, at each decision step, the agent evaluates the speech-induced action uncertainty $\mathcal{U}_i^{\mathrm{SA}}$ in Eq. (3). When $\mathcal{U}_i^{\mathrm{SA}} \le \tau_U$, the agent proceeds with task execution. Otherwise, it localizes decision-relevant symbolic differences between competing intents, performs targeted exploration, and converts the interaction feedback into constraints for speech reinterpretation. The updated intent belief is then used for subsequent action selection. The procedure terminates upon task completion or exhaustion of the interaction budget. Algorithm 1 in Appendix B summarizes the complete framework.

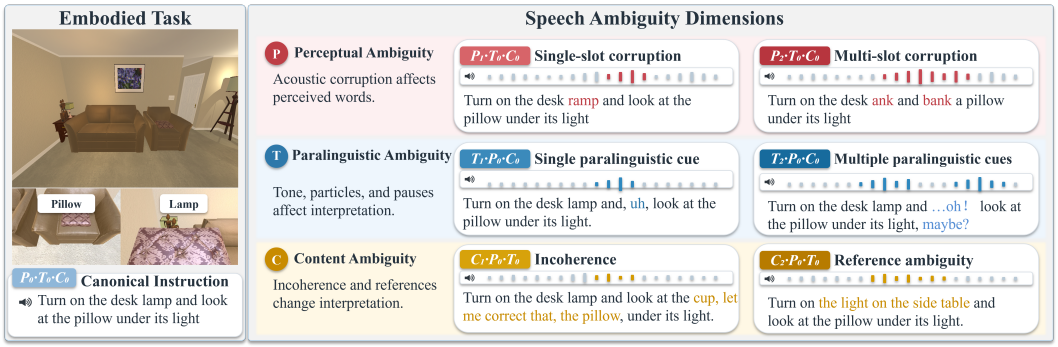


Figure 3: Speech ambiguity levels. Each of the perceptual, paralinguistic, and content dimensions has three levels, yielding $3 \times 3 \times 3 = 27$ speech conditions.

5 SPEAR-BENCH: A BENCHMARK FOR EMBODIED CONTROL UNDER AMBIGUOUS SPEECH

To systematically study how speech ambiguity affects task understanding and embodied control, we construct SPEAR-BENCH along three complementary dimensions of spoken instructions: perceptual ambiguity (P), paralinguistic ambiguity (T), and content ambiguity (C). Each dimension is instantiated at three controlled levels, as illustrated in Figure 3. We build SPEAR-BENCH from ALFWorld tasks by varying the spoken instruction while holding the environment, underlying task goal, and intended task fixed. This matched $3 \times 3 \times 3$ factorial design yields 27 speech conditions for each task and 972 task instances in total.

Perceptual ambiguity (P). Acoustic factors such as environmental noise, accents, and pronunciation variations can make parts of a spoken instruction difficult to perceive reliably. For example, "put the cup in the bin" may be perceived as "pick the cap from the bin." We characterize perceptual ambiguity by the extent of localized corruption in the perceived instruction and define three controlled levels: P_0 denotes clear speech with no localized corruption, P_1 contains corruption in one localized speech segment, and P_2 contains corruption in multiple segments.

Paralinguistic ambiguity (T). Even when the task-relevant content is correctly perceived, paralinguistic cues in speaking style can affect how confidently the intended task is interpreted. We decompose the cues that give rise to such ambiguity into three types: speaking tone, discourse particles, and pauses. Based on these cues, we define three controlled levels: T_0 contains no marked cue, T_1 contains one cue type, and T_2 combines two or more cue types.

Content ambiguity (C). Content ambiguity arises from the flexible and incremental nature of spoken language, where interruptions, revisions, or self-corrections can disrupt discourse coherence, while references and omissions can leave task-relevant information dependent on context. For example, "put that on the table... oh, on the plate" contains both an interrupted revision and a context-dependent reference that must be resolved to recover the intended task. We define three controlled levels according to the source of content ambiguity: C_0 contains direct and coherent instructions, C_1 contains discourse-level incoherence such as interruptions or revisions, and C_2 contains context-dependent expressions such as references or omissions.

Starting from 36 ALFWorld tasks (Shridhar et al., 2021), we construct content variants at different C levels using ChatGPT, synthesize their spoken realizations with controlled paralinguistic cues at different T levels using Qwen-Audio-3.0-TTS (Xiang et al., 2026a), and introduce localized perceptual corruption according to the specified P level. Human annotators further verify that the generated samples preserve the original task intent and conform to their designated ambiguity conditions. The resulting benchmark enables controlled evaluation of how different sources of speech ambiguity affect intent understanding, task execution, and interaction cost. Detailed generation and validation procedures are provided in Appendix C.

6 EXPERIMENTS

We conduct experiments to answer three questions: (1) how our method compares with existing methods under ambiguous speech; (2) how the speech-induced action uncertainty threshold τ_U affects ambiguity resolution; and (3) how the probabilistic uncertainty modeling and symbolic disambiguation mechanisms individually contribute to the overall performance.

6.1 EXPERIMENTAL SETUP

We evaluate our method on SPEAR-Bench and the public REI-Bench (Jiang et al., 2026) against ReAct (Yao et al., 2023), ReflAct (Kim et al., 2025), and WALL-E (Zhou et al., 2025). On SPEAR-Bench, we use Qwen3.8-Omni-Flash (Qwen Team, 2026) for direct speech input and DeepSeek-V4.1-Flash (DeepSeek-AI, 2026) with transcripts from a fixed ASR model. Since REI-Bench provides textual instructions only, we evaluate it with DeepSeek-V4.1-Flash. We report Task Success Rate (Success, ↑), interaction steps (Steps, ↓), and Final Intent Accuracy (Intent Acc., ↑). Further details are provided in Appendix D.

6.2 EVALUATION ACROSS AMBIGUITY BENCHMARKS

Overall performance. Table 1 summarizes the main results on SPEAR-Bench and REI-Bench. On SPEAR-Bench, our method outperforms all baselines in task success under both backbone settings, achieving Success rates of 72.99 and 83.11 with DeepSeek-V4.1-Flash and Qwen3.8-Omni-Flash, respectively, while also requiring the fewest interaction steps. The corresponding Final Intent Accuracy reaches 81.23 and 92.42. On REI-Bench, our method again achieves the highest task success at 80.77 and a Final Intent Accuracy of 85.18, while requiring only 20.79 interaction steps on average. Overall, these results demonstrate that our framework consistently improves embodied planning under ambiguity-induced uncertainty, across both ambiguous speech and textual referring expressions. Fine-grained comparisons across ambiguity conditions are provided in Appendix D.

Table 1: Main results across ambiguity benchmarks. Intent Acc. for our method is reported in parentheses. REI-Bench contains textual instructions only and is therefore evaluated with DeepSeek-V4.1-Flash. Higher Success and lower Steps are better.

(a) SPEAR-Bench

	DeepSeek-V4.1-Flash		Qwen3.8-Omni-Flash	
Method	Success ↑	Steps ↓	Success ↑	Steps ↓
ReAct	46.60	29.60	39.20	32.35
ReflAct	58.95	30.65	47.84	33.55
WALL-E	45.99	29.61	46.91	31.79
Ours	**72.99 (81.23)**	**24.98**	**83.11 (92.42)**	**21.95**

(b) REI-Bench

	DeepSeek-V4.1-Flash	
Method	Success ↑	Steps ↓
ReAct	50.74	31.57
ReflAct	60.00	**16.56**
WALL-E	47.04	30.35
Ours	**80.77 (85.18)**	20.79

Performance across ambiguity levels. We further analyze how different sources of speech ambiguity affect embodied control using the controlled structure of SPEAR-Bench. For each ambiguity dimension, performance at each level is averaged over all nine combinations of the other two dimensions. Table 2 shows that the three ambiguity sources lead to distinct performance patterns. Perceptual ambiguity produces a gradual degradation as its level increases, whereas paralinguistic ambiguity has a relatively limited effect on task success despite larger changes in intent accuracy and interaction cost. Content ambiguity exhibits the strongest variation across levels, with C_2 being substantially more challenging than C_0 and C_1. These results indicate that different sources of speech ambiguity propagate differently to intent understanding and embodied execution, rather than inducing a uniform degradation in performance. Complete results for all 27 P_i-T_j-C_k conditions are provided in Appendix D.

6.3 SENSITIVITY TO THE UNCERTAINTY THRESHOLD

We analyze the sensitivity of our method to the speech-induced action uncertainty threshold τ_U. A smaller τ_U triggers ambiguity resolution more frequently, whereas a larger threshold allows execu-

Table 2: Performance across ambiguity levels on SPEAR-Bench. L_0, L_1, and L_2 indicate increasing ambiguity levels. Results are marginalized over the other two dimensions for each ambiguity type.

	Success ↑			Intent Acc. ↑			Steps ↓		
Ambiguity	L_0	L_1	L_2	L_0	L_1	L_2	L_0	L_1	L_2
Perceptual (P)	88.89	86.11	83.33	100.00	100.00	91.67	18.53	20.69	21.67
Paralinguistic (T)	88.89	86.11	86.11	100.00	97.00	91.00	18.53	19.75	21.44
Content (C)	88.89	91.43	60.00	100.00	100.00	62.86	18.53	20.66	31.63

tion to continue under greater action disagreement. As shown in Figure 4, increasing τ_U consistently reduces interaction cost, but its effect on task success is non-monotonic. Small thresholds lead to frequent ambiguity resolution and substantially more interaction steps without corresponding gains in task success. In contrast, an excessively large threshold reduces interaction cost but degrades task success by allowing unresolved action disagreement to persist. The intermediate setting $\tau_U = 0.5$ achieves the highest task success while maintaining relatively low interaction cost, providing a favorable balance between ambiguity resolution and execution efficiency. Full results and trigger statistics are provided in Appendix D.

Table 3: Ablation of the two core mechanisms, speech-induced action uncertainty and symbolic world-constraint alignment, on SPEAR-Bench with Qwen3.8-Omni-Flash.

Method	Success ↑	Intent Acc. ↑	Steps ↓
w/o $\mathcal{U}^{\mathrm{SA}}$	58.71	71.96	31.08
w/o Sym.	75.93	–	16.23
Ours	**83.11**	**92.42**	**21.95**

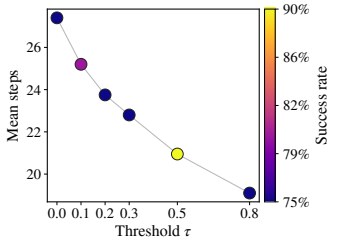


Figure 4: Sensitivity to the uncertainty threshold choice.

6.4 Ablation of Core Mechanisms

We ablate the two core mechanisms of our framework: speech-induced action uncertainty $\mathcal{U}^{\mathrm{SA}}$ and world-constraint alignment. In *w/o* $\mathcal{U}^{\mathrm{SA}}$, we replace the uncertainty criterion $\mathcal{U}_i^{\mathrm{SA}} > \tau_U$ with direct LLM judgment of whether the current intent ambiguity requires resolution, while retaining the subsequent symbolic disambiguation procedure. In *w/o Sym.*, we retain $\mathcal{U}^{\mathrm{SA}}$ for identifying decision-relevant ambiguity, but replace symbolic disagreement localization and world-constraint alignment with direct LLM-based disambiguation from the spoken instruction and interaction history. All variants use Qwen3.8-Omni-Flash on SPEAR-Bench with the remaining settings fixed.

As shown in Table 3, removing either $\mathcal{U}^{\mathrm{SA}}$ or the symbolic world-constraint mechanism substantially degrades task success, validating both components of our framework. These results show that ambiguity handling requires not only identifying when intent uncertainty becomes decision-relevant, but also a structured mechanism for resolving it through interaction. Final Intent Accuracy is not reported for *w/o Sym.*, as this variant lacks the symbolic intent representation required for alignment with the ALFWorld PDDL task specification.

7 Conclusion

We study embodied control under ambiguous spoken instructions and introduce speech-induced action uncertainty to quantify when intent uncertainty affects action decisions. We further develop world-constraint alignment, which uses embodied exploration and interaction-derived constraints to refine speech interpretation and planning. We also introduce SPEAR-Bench for evaluating perceptual, paralinguistic, and content ambiguity. Experiments on SPEAR-Bench and REI-Bench show improved task success and final intent recognition with lower interaction cost. Overall, our results suggest that speech ambiguity need not be fully resolved before acting, but should be addressed when it becomes consequential for embodied decision making.

REFERENCES

Wonje Choi, Jinwoo Park, Sanghyun Ahn, Daehee Lee, and Honguk Woo. NeSyC: A neuro-symbolic continual learner for complex embodied tasks in open domains. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d569e4655e2835e3d38310b67c5ad646-Abstract-Conference.html.

DeepSeek-AI. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression. Technical report, DeepSeek-AI, 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash/blob/main/DeepSeek_V41_Tech_Report.pdf.

Xin Ding, Jianyu Wei, Yifan Yang, Shiqi Jiang, Qianxi Zhang, Hao Wu, Fucheng Jia, Liang Mi, Yuxuan Yan, Weijun Wang, Yunxin Liu, Zhibo Chen, and Ting Cao. AdaNav: Adaptive reasoning with uncertainty for vision-language navigation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2509.24387.

Weixi Feng, Tsu-Jui Fu, Yujie Lu, and William Yang Wang. ULN: Towards underspecified vision-and-language navigation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6394–6412. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.429. URL https://aclanthology.org/2022.emnlp-main.429/.

Helong Huang, Min Cen, Kai Tan, Xingyue Quan, Guowei Huang, and Hong Zhang. GraphCoT-VLA: A 3D spatial-aware reasoning vision-language-action model for robotic manipulation with ambiguous instructions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18324–18332, 2026. doi: 10.1609/aaai.v40i22.38896. URL https://ojs.aaai.org/index.php/AAAI/article/view/38896.

Anastasiia Ivanova, Eva Bakaeva, Zoya Volovikova, Alexey K. Kovalev, and Aleksandr I. Panov. AmbiK: Dataset of ambiguous tasks in kitchen environment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33216–33241. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1593. URL https://aclanthology.org/2025.acl-long.1593/.

Chenxi Jiang, Chuhao Zhou, and Jianfei Yang. REI-Bench: Can embodied agents understand vague human instructions in task planning? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=vmBIF25KLf.

Jeonghye Kim, Sojeong Rhee, Minbeom Kim, Dohyung Kim, Sangmook Lee, Youngchul Sung, and Kyomin Jung. ReflAct: World-grounded decision making in LLM agents via goal-state reflection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33433–33465, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1697. URL https://aclanthology.org/2025.emnlp-main.1697/.

Pengqi Li, Lizhong Ding, Zhehao Zhou, Chunhui Zhang, Jiarun Fu, Hao Li, Ye Yuan, and Guoren Wang. Demystifying uncertainty in LLMs: Active calibration between concepts and human evaluations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5726–5759, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.259. URL https://aclanthology.org/2026.acl-long.259/.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Huaping Liu, Di Guo, and Angelo Cangelosi. Embodied intelligence: A synergy of morphology, action, perception and learning. *ACM Computing Surveys*, 57(7), 2025. doi: 10.1145/3717059. URL https://doi.org/10.1145/3717059.

Changdae Oh, Seongheon Park, To Eun Kim, Jiatong Li, Wendi Li, Samuel Yeh, Sean Du, Hamed Hassani, Paul Bogdan, Dawn Song, and Sharon Li. Uncertainty quantification in LLM agents:

Foundations, emerging challenges, and opportunities. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16219–16250, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.738. URL https://aclanthology.org/2026.acl-long.738/.

Qwen Team. Qwen3.8-Omni: Towards native omni-modal agents. *arXiv preprint arXiv:2609.25611*, 2026. URL https://arxiv.org/abs/2609.25611.

Allen Z. Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, Zhenjia Xu, Dorsa Sadigh, Andy Zeng, and Anirudha Majumdar. Robots that ask for help: Uncertainty alignment for large language model planners. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 661–682, 2023.

David Sasu, Benedict Quartey, Kweku Andoh Yamoah, and Natalie Schluter. Enhancing speech instruction understanding and disambiguation in robotics via speech prosody. In *Interspeech 2025*, pp. 1958–1962, 2025. doi: 10.21437/Interspeech.2025-1961.

SeungWon Seo, SooBin Lim, SeongRae Noh, Haneul Kim, and HyeongYeop Kang. From assumptions to actions: Turning LLM reasoning into uncertainty-aware planning for embodied agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=GODFBZhFcX.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021.

Manan Suri, Puneet Mathur, Nedim Lipka, Franck Dernoncourt, Ryan A. Rossi, and Dinesh Manocha. Structured uncertainty guided clarification for LLM agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 40811–40838, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.2028. URL https://aclanthology.org/2026.findings-acl.2028/.

Bajian Xiang, Cheng Wen, Han Zhao, Hao Wang, Haoxu Wang, Jiawei Jin, Jiayan Cui, Jie Chen, Mengxi Nie, Tianyu Zhao, Weiqin Li, Xiang Lv, Xiangang Li, Yang Xiang, and Yang Zhou. Qwen-Audio-3.0-TTS: Freely controllable and highly robust speech synthesis with multi-stage training paradigm. *arXiv preprint arXiv:2607.23938*, 2026a.

Kun Xiang, Terry Jingchen Zhang, Yinya Huang, Jixi He, Zirong Liu, Yueling Tang, Ruizhe Zhou, Lijing Luo, Youpeng Wen, Xiuwei Chen, Bingqian Lin, Jianhua Han, Hang Xu, Hanhui Li, Bin Dong, and Xiaodan Liang. Aligning perception, reasoning, modeling and interaction: A survey on physical AI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026b. doi: 10.1109/TPAMI.2026.3703038. URL https://doi.org/10.1109/TPAMI.2026.3703038. Early Access.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Lance Ying, Xinyi Li, Shivam Aarya, Yizirui Fang, Yifan Yin, Jason Xinyu Liu, Stefanie Tellex, Joshua B. Tenenbaum, and Tianmin Shu. Pragmatic embodied spoken instruction following in human-robot collaboration with theory of mind. In *2026 IEEE International Conference on Robotics and Automation (ICRA)*, 2026. URL https://arxiv.org/abs/2409.10849.

Siyu Zhou, Tianyi Zhou, Yijun Yang, Guodong Long, Deheng Ye, Jing Jiang, and Chengqi Zhang. WALL-E 2.0: World alignment by NeuroSymbolic learning improves world model-based LLM agents. *arXiv preprint arXiv:2504.15785*, 2025. URL https://arxiv.org/abs/2504.15785.

A Calibration of the Speech-Induced Action Uncertainty Threshold

The threshold τ_U determines when speech-induced action uncertainty is sufficiently large to trigger ambiguity resolution. A key question is how to distinguish decision-relevant disagreement from the background disagreement that can arise during normal inference. Even for an unambiguous spoken instruction, the model may assign probability mass to multiple intent interpretations, resulting in nonzero U_i^{SA}. We therefore calibrate τ_U against the uncertainty observed under unambiguous speech, using it as a reference for the disagreement naturally produced by the inference pipeline.

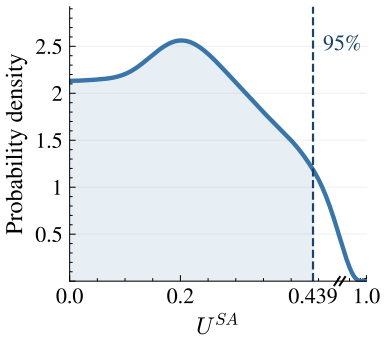


Figure 5: Empirical distribution of speech-induced action uncertainty under unambiguous speech. The dashed line denotes the 95th percentile ($\tau_U = 0.439$).

To estimate the uncertainty that arises in the absence of speech ambiguity, we construct a calibration set using $N = 120$ held-out ALFWorld tasks. For each task, we synthesize a clear spoken instruction from its original unambiguous instruction using Qwen-Audio-3.0-TTS, following the same speech synthesis pipeline as SPEAR-Bench (Appendix C), but without introducing perceptual, paralinguistic, or content ambiguity. We then apply the inference procedure in Section 4 and compute $U_i^{\mathrm{SA}} = I(Z; A_i \mid x, h_i)$ at every decision step along each task trajectory. This yields a collection $\mathcal{U}_{\mathrm{clean}}$ of uncertainty measurements across tasks and interaction stages, which serves as the reference distribution under unambiguous speech.

Across the 120 calibration trajectories, we record the speech-induced action uncertainty U_i^{SA} at every decision step and plot its empirical distribution in Figure 5. We observe that, under unambiguous speech, most uncertainty values are concentrated below 0.4, indicating that a moderate amount of disagreement can arise naturally from the inference process even without speech ambiguity. We therefore place τ_U near the upper tail of this clean-speech distribution and regard 0.4–0.5 as a reasonable operating range. Specifically, we use its 95th percentile, $\tau_U = \widehat{Q}_{0.95}(\mathcal{U}_{\mathrm{clean}}) = 0.439$, as the default threshold for all main experiments. The sensitivity to this choice is further analyzed in Section 6.3.

B Inference Procedure for World-Constraint Alignment

Algorithm 1 summarizes the complete inference procedure. At each interaction step, the agent evaluates speech-induced action uncertainty and either executes the current task policy or initiates targeted exploration. When ambiguity resolution is triggered, the agent localizes the symbolic differences between competing intents, selects a discriminative and verifiable difference, and explores the environment to test the corresponding hypothesis. Observations acquired during exploration continuously update the world representation, while intent inference is resumed after verification terminates. Verified hypotheses are incorporated into the world state, whereas unsupported hypotheses induce task-specific constraints that guide subsequent speech reinterpretation and action selection. All actions count toward a global budget of $B = 50$, with each exploration limited to $T_{\mathrm{exp}} = 20$ actions.

C SPEAR-Bench Construction Details

This appendix details the construction and validation of SPEAR-Bench. For each selected ALFWorld task (Shridhar et al., 2021), we preserve the environment and intended task goal while sequentially constructing content, paralinguistic, and perceptual variants and verifying that they satisfy their designated ambiguity conditions. We describe the source task selection, ambiguity construction and verification procedures, and final benchmark statistics.

Algorithm 1 World-Constraint Alignment through Embodied Exploration

Require: Speech x, initial observation o_0, threshold τ_U, global budget $B = 50$, exploration budget $T_{\text{exp}} = 20$

$h_0 \leftarrow (o_0)$, $G_0 \leftarrow \text{Build}(o_0)$, $\Lambda_0 = \Gamma_0 = \emptyset$, $i \leftarrow 0$
$\mathcal{B}_0 \leftarrow \text{Infer}(x, h_0, \Gamma_0)$
while task incomplete and $i < B$ **do**
  Obtain $\{\pi_i^k\}$ from $\mathcal{B}_i, G_i$ and compute U_i^{SA}
  if $U_i^{\text{SA}} \leq \tau_U$ **then**
    $a_i \leftarrow \arg\max_{a \in \mathcal{A}_{\Lambda_i}} \sum_k b_i^k \pi_i^k(a)$
    $(h_{i+1}, G_{i+1}) \leftarrow \text{Step}(a_i, h_i, G_i)$; $i \leftarrow i + 1$
    $\mathcal{B}_i \leftarrow \text{Infer}(x, h_i, \Gamma_i)$
  else
    $(k^*, l^*) \leftarrow \arg\max_{k<l} b_i^k b_i^l D_{\text{JS}}(\pi_i^k, \pi_i^l)$
    $\Delta_i \leftarrow \text{Diff}(z_{k^*}, z_{l^*}, G_i)$
    $(\psi, c) \leftarrow \text{Hypothesize}(h_i, G_i, \Delta_i)$
    $t \leftarrow 0$, $r \leftarrow$ UNRESOLVED
    while $r =$ UNRESOLVED and $t < T_{\text{exp}}$ and $i < B$ **do**
      $a_i \leftarrow \text{Explore}(\psi, c, h_i, G_i, \Lambda_i)$
      $(h_{i+1}, G_{i+1}) \leftarrow \text{Step}(a_i, h_i, G_i)$; $i \leftarrow i + 1$, $t \leftarrow t + 1$
      $r \leftarrow \text{Verify}(\psi, c, h_i, G_i)$
    end while
    if $r \neq$ VERIFIED **then**
      $(\lambda, \gamma) \leftarrow \text{Constrain}(\psi, h_i, G_i)$; $\Lambda_i \leftarrow \Lambda_i \cup \{\lambda\}$, $\Gamma_i \leftarrow \Gamma_i \cup \{\gamma\}$
    end if
    $\widetilde{\mathcal{B}}_i \leftarrow \mathcal{R}_\theta(x, \mathcal{B}_i, h_i, \Gamma_i)$; $\mathcal{B}_i \leftarrow \mathcal{C}_{\Lambda_i}(\widetilde{\mathcal{B}}_i)$
  end if
end while
return task outcome

C.1 Source Dataset and Task Selection

We construct SPEAR-Bench from ALFWorld, where each episode pairs a natural-language household instruction with an executable environment and a well-defined task goal. This pairing allows us to vary the spoken realization of an instruction while keeping the underlying environment and intended task fixed. We select 36 source tasks from six ALFWorld task types, with six tasks from each type: `look at obj in light`, `pick and place simple`, `pick clean then place in recep`, `pick cool then place in recep`, `pick heat then place in recep`, and `pick two obj and place`. These task types cover object inspection, placement, cleaning, cooling, heating, and multi-object manipulation. For each source task, we preserve the original environment and gold task goal across all benchmark variants. We additionally extract task-relevant information from the original instruction, including the target object and destination, which is used to condition subsequent generation and verify that the intended task is preserved.

C.2 Construction of Ambiguity Dimensions

Each benchmark condition is specified by a combination P_i-T_j-C_l, where $i, j, l \in \{0, 1, 2\}$ denote the corresponding levels of perceptual, paralinguistic, and content ambiguity. For each source task and each P_i-T_j-C_l condition, we construct the sample sequentially in the order $C \rightarrow T \rightarrow P$: we first generate the content variant at level C_l, then synthesize its spoken realization with paralinguistic cues at level T_j, and finally introduce perceptual corruption according to level P_i. The underlying environment and intended task goal are kept fixed throughout this process.

Content ambiguity (C). We construct content ambiguity by modifying the linguistic realization of the instruction while keeping the gold task goal fixed. Starting from each original ALFWorld instruction, we use GPT-4o to generate a content variant for the designated C level, conditioned on the original instruction, gold task information, and environment context. The generation prompt constrains the modification to the ambiguity pattern specified by that level while preserving the

intended task. For C_0, we retain the original direct and coherent instruction without introducing additional content ambiguity. For C_1, we introduce discourse-level disruption through interruption, revision, or self-correction. For example, "Put the mug on the plate" may be reformulated as "Put the mug on the table—no, on the plate." For C_2, we introduce context-dependent expressions by replacing explicit task-relevant entities or locations with references or omissions that require the current environment for interpretation. For example, an explicitly named object may instead be referred to as "that one" or by a context-dependent description.

Each generated candidate is first screened by DeepSeek-V4.1 for consistency with the original task. The automatic check verifies that the gold task-relevant information is either explicitly preserved or remains recoverable from the environment context, and that the generated instruction exhibits the ambiguity pattern required by its assigned C level. Human annotators then inspect the original task, environment context, and generated instruction and verify (1) whether the intended task is preserved and (2) whether the instruction conforms to the designated C level. Candidates that change the gold task, make the intended task unrecoverable, or fail to exhibit the required ambiguity pattern are rejected and regenerated.

Paralinguistic ambiguity (T). For each content variant, we synthesize spoken realizations with controlled paralinguistic cues using Qwen-Audio-3.0-TTS (Xiang et al., 2026a). The task-relevant semantic content is kept fixed, while the speech realization is varied according to the T level. We consider three types of paralinguistic cues: speaking tone, discourse particles, and pauses. These include hesitant or uncertain intonation, particles such as "um" or "well", and salient pauses or broken rhythm around task-relevant portions of the utterance. The three levels are defined by the number of marked cue types:

$$T_0 : 0 \text{ marked cue types}, \qquad T_1 : 1 \text{ cue type}, \qquad T_2 :\geq 2 \text{ cue types}. \tag{10}$$

Specifically, T_0 uses a neutral realization without marked paralinguistic cues. For T_1, one cue type is randomly selected and introduced, while for T_2, two or more cue types are randomly combined within the same utterance. This construction varies the paralinguistic realization while preserving the underlying task semantics.

For each T level, we generate multiple candidate realizations using the sampled cue configurations and evaluate them through human listening. Human annotators assess whether the paralinguistic variation changes the intended task conveyed by the original instruction. We retain samples for which the intended task remains unchanged and the designated paralinguistic cues are perceptually identifiable. Samples that alter the perceived task intent, fail to exhibit the designated cues, or substantially degrade intelligibility are rejected and regenerated.

Perceptual ambiguity (P). For each synthesized speech sample, we introduce localized acoustic corruption around task-relevant segments while leaving the remaining speech unchanged. This construction creates perceptual uncertainty in critical parts of the instruction without making the entire utterance unintelligible. We mix environmental noise into selected task-relevant segments and generate candidate corruptions at multiple signal-to-noise ratios (SNRs). SNR is used only as a generation parameter for controlling acoustic strength; the final P level is determined by the resulting localized corruption rather than by a fixed SNR threshold. We use a fixed ASR model, Qwen3.8-Omni-Flash, to transcribe each corrupted candidate and compare task-relevant slots in the transcription with those in the clean instruction. Let N_{err} denote the number of task-relevant localized segments whose content becomes corrupted after acoustic degradation. We define the three levels as

$$P_0 : N_{\text{err}} = 0, \qquad P_1 : N_{\text{err}} = 1, \qquad P_2 : N_{\text{err}} \geq 2. \tag{11}$$

P_0 therefore corresponds to the clean synthesized speech, P_1 contains a single localized perceptual corruption, and P_2 contains corruption in multiple task-relevant segments. Candidates that fail to produce the designated corruption pattern, or that make the utterance globally unintelligible, are discarded and regenerated.

Human annotators further listen to the retained candidates together with their clean references. They verify that the intended task-relevant segment or segments are perceptually degraded, that the remaining speech is intelligible, and that the sample conforms to its assigned P level. Samples that fail to produce the designated localized corruption or substantially impair the intelligibility of the entire instruction are rejected and regenerated.

Table 4: Detailed results across all 27 ambiguity conditions on SPEAR-Bench with Qwen3.8-Omni-Flash. Higher Success and lower Steps are better.

	ReAct		ReflAct		WALL-E		**Ours**	
Condition	Succ. ↑	Steps ↓	Succ. ↑	Steps ↓	Succ. ↑	Steps ↓	Succ. ↑	Steps ↓
P_0-T_0-C_0	52.84	45.48	64.98	44.70	65.93	45.53	**100**	**23.31**
P_0-T_0-C_1	54.49	43.66	61.04	44.83	61.93	36.47	**100**	**21.15**
P_0-T_0-C_2	21.47	6.09	27.57	12.57	25.97	5.41	**55.55**	**18.54**
P_0-T_1-C_0	34.67	47.00	43.32	48.38	39.96	45.01	**100**	**20.24**
P_0-T_1-C_1	56.14	45.48	61.04	44.44	63.93	51.87	**100**	**21.15**
P_0-T_1-C_2	28.07	3.28	33.47	9.77	35.96	2.60	**55.55**	**21.15**
P_0-T_2-C_0	41.28	39.47	53.16	44.61	49.94	41.27	**100**	**20.61**
P_0-T_2-C_1	56.14	48.75	66.94	47.02	65.93	42.32	**100**	**22.06**
P_0-T_2-C_2	19.81	2.43	19.69	1.30	23.97	1.96	**55.54**	**20.17**
P_1-T_0-C_0	54.49	45.88	64.98	47.86	65.93	45.53	**100**	**19.70**
P_1-T_0-C_1	51.19	48.26	64.98	48.25	65.92	49.42	**100**	**26.38**
P_1-T_0-C_2	23.12	10.25	23.63	4.07	27.97	15.46	**55.54**	**19.19**
P_1-T_1-C_0	34.67	48.67	47.26	45.58	39.96	46.76	**100**	**25.67**
P_1-T_1-C_1	54.49	47.73	66.94	47.31	63.93	45.12	**100**	**21.15**
P_1-T_1-C_2	29.72	5.68	35.44	3.81	31.96	3.27	**55.54**	**33.18**
P_1-T_2-C_0	44.58	43.16	49.22	45.58	53.94	44.04	**100**	**20.24**
P_1-T_2-C_1	49.53	46.76	66.94	47.31	65.92	42.52	**100**	**22.42**
P_1-T_2-C_2	19.81	2.93	25.60	17.29	23.97	3.65	**44.43**	**20.33**
P_2-T_0-C_0	49.53	45.18	63.01	51.18	61.93	46.09	**100**	**24.76**
P_2-T_0-C_1	51.19	49.78	64.98	45.39	57.93	44.13	**100**	**24.40**
P_2-T_0-C_2	23.12	11.10	29.54	19.76	27.97	11.28	**33.32**	**19.52**
P_2-T_1-C_0	26.42	46.29	33.47	43.27	33.96	47.11	**100**	**18.07**
P_2-T_1-C_1	46.23	45.91	55.14	46.53	55.94	47.81	**100**	**21.86**
P_2-T_1-C_2	31.37	5.53	37.41	4.98	33.96	3.95	**55.54**	**25.05**
P_2-T_2-C_0	36.33	39.94	45.29	41.48	41.95	39.95	**77.76**	**22.07**
P_2-T_2-C_1	47.89	44.13	63.01	43.43	55.94	44.69	**100**	**20.43**
P_2-T_2-C_2	19.81	4.63	23.63	5.15	23.97	5.11	**55.54**	**19.85**

C.3 Dataset Statistics

The final SPEAR-Bench contains 36 underlying ALFWorld tasks, each instantiated under all $3 \times 3 \times 3 = 27$ combinations of perceptual, paralinguistic, and content ambiguity, resulting in 972 benchmark instances. After generation and human verification, the resulting speech corpus contains 5,808 seconds of audio, with an average utterance duration of 6 seconds and an average instruction length of 15 words (78 characters).

D Additional Experimental Details

This appendix provides additional experimental details and fine-grained results omitted from the main text. We first describe the evaluation setup and implementation details, and then report complete results across individual ambiguity conditions on SPEAR-Bench and REI-Bench. These results complement the aggregate analyses in Section 6 by showing the behavior of different methods under specific ambiguity configurations.

D.1 Experimental Setup

We evaluate on SPEAR-Bench and REI-Bench using the same protocol as in the main experiments. We compare against ReAct (Yao et al., 2023), ReflAct (Kim et al., 2025), and WALL-E (Zhou et al., 2025). On SPEAR-Bench, Qwen3.8-Omni-Flash (Qwen Team, 2026) directly processes the spoken instruction, while DeepSeek-V4.1-Flash (DeepSeek-AI, 2026) receives transcripts produced by a fixed ASR model. Since REI-Bench provides textual instructions only, we evaluate it with DeepSeek-V4.1-Flash. All methods operate under a global interaction budget of 50 environment actions. For our method, each episode of ambiguity-resolution is additionally limited to at most

Table 5: Detailed results across all 27 ambiguity conditions on SPEAR-Bench with DeepSeek-V4.1-Flash. Higher Success and lower Steps are better.

	ReAct		ReflAct		WALL-E		**Ours**	
Condition	Succ. ↑	Steps ↓	Succ. ↑	Steps ↓	Succ. ↑	Steps ↓	Succ. ↑	Steps ↓
P_0-T_0-C_0	67.62	44.21	84.61	38.72	68.46	46.23	**79.37**	**25.84**
P_0-T_0-C_1	62.79	41.78	84.61	36.19	61.38	45.36	**90.70**	**24.83**
P_0-T_0-C_2	26.57	3.99	31.72	4.59	30.69	6.98	**56.69**	**23.39**
P_0-T_1-C_0	55.55	43.77	66.09	47.87	51.94	42.32	**79.37**	**24.28**
P_0-T_1-C_1	60.37	43.14	68.74	46.51	59.01	46.23	**79.37**	**22.38**
P_0-T_1-C_2	28.98	3.00	34.37	3.23	33.05	5.09	**56.69**	**27.81**
P_0-T_2-C_0	62.79	46.53	81.96	43.18	63.74	44.14	**90.70**	**21.81**
P_0-T_2-C_1	55.55	47.77	76.68	46.28	47.22	39.64	**90.70**	**22.07**
P_0-T_2-C_2	24.15	1.60	29.09	8.83	23.61	1.58	**56.69**	**27.37**
P_1-T_0-C_0	57.96	37.10	81.96	41.30	59.01	40.70	**79.37**	**35.00**
P_1-T_0-C_1	60.37	43.61	76.68	39.91	59.01	37.86	**79.36**	**26.17**
P_1-T_0-C_2	24.15	3.59	31.72	4.85	25.97	3.94	**56.69**	**20.30**
P_1-T_1-C_0	62.79	47.77	81.96	45.28	63.74	38.69	**79.36**	**22.07**
P_1-T_1-C_1	60.37	42.18	76.68	36.81	54.30	43.39	**90.70**	**22.62**
P_1-T_1-C_2	31.39	3.99	31.72	3.49	28.33	2.29	**68.03**	**24.28**
P_1-T_2-C_0	62.79	37.94	74.03	41.92	61.38	40.67	**79.36**	**23.97**
P_1-T_2-C_1	55.55	41.86	63.46	42.31	49.57	37.94	**79.36**	**23.64**
P_1-T_2-C_2	26.57	3.28	31.72	24.81	23.61	1.58	**56.69**	**24.28**
P_2-T_0-C_0	57.96	42.10	79.32	44.54	63.74	46.03	**79.36**	**32.46**
P_2-T_0-C_1	55.54	43.93	74.03	39.98	51.93	40.86	**79.36**	**21.12**
P_2-T_0-C_2	24.15	2.80	26.44	2.26	23.61	2.76	**45.35**	**33.10**
P_2-T_1-C_0	45.89	38.90	58.17	48.23	42.49	42.28	**68.03**	**22.07**
P_2-T_1-C_1	41.05	44.17	58.17	39.69	44.86	45.24	**90.70**	**24.83**
P_2-T_1-C_2	31.39	3.40	37.02	3.69	28.33	2.96	**45.35**	**24.28**
P_2-T_2-C_0	50.71	41.10	66.09	46.83	51.93	46.98	**68.03**	**24.28**
P_2-T_2-C_1	43.47	45.69	60.81	46.25	49.57	47.73	**100.00**	**26.48**
P_2-T_2-C_2	21.73	0.00	23.80	0.00	21.25	0.00	**45.35**	**23.73**

$T_{\text{exp}} = 20$ exploration actions, and the uncertainty threshold is fixed according to the calibration procedure in Appendix A. We report Task Success Rate (Success, ↑), interaction steps (Steps, ↓), and Final Intent Accuracy (Intent Acc., ↑), following the definitions used in the main text. Unless otherwise specified, all remaining settings are kept fixed across methods and ambiguity conditions.

D.2 Detailed Results on SPEAR-Bench

Tables 4 and 5 report results under all 27 combinations of perceptual, paralinguistic, and content ambiguity on SPEAR-Bench. Unlike the marginalized results in the main text, these tables expose performance under each specific P_i-T_j-C_k configuration and allow us to examine how different ambiguity sources interact.

Qwen3.8-Omni-Flash. As shown in Table 4, our method achieves the highest task success in nearly all ambiguity configurations. Performance remains particularly stable across most C_0 and C_1 conditions, even as perceptual and paralinguistic ambiguity increase. In contrast, configurations involving C_2 are substantially more challenging for all methods, and the degradation becomes more pronounced when combined with stronger perceptual or paralinguistic ambiguity. Despite this difficulty, our method retains a clear advantage over the baselines in most of these conditions, indicating that the proposed disambiguation mechanism remains effective under compounded ambiguity.

DeepSeek-V4.1-Flash. Table 5 shows a similar overall pattern with the transcript-based backbone. The advantage of our method is most evident in conditions with stronger content ambiguity, where direct planning baselines degrade substantially. The gains are especially pronounced when content ambiguity is combined with perceptual or paralinguistic degradation, suggesting that the benefit of explicit uncertainty modeling and interaction-based disambiguation becomes larger as multiple

Table 6: Detailed results across all nine ambiguity conditions on REI-Bench with DeepSeek-V4.1-Flash. Higher Success and lower Steps are better.

		ReAct		ReflAct		WALL-E		**Ours**	
Context	RE Type	Succ. ↑	Steps ↓	Succ. ↑	Steps ↓	Succ. ↑	Steps ↓	Succ. ↑	Steps ↓
Standard	Explicit	66.67	34.88	76.67	13.62	66.68	33.68	**89.28**	**17.71**
	Mixed	66.67	32.16	80.00	18.55	56.68	29.30	**88.45**	**16.65**
	Implicit	66.66	31.30	73.33	15.92	63.34	26.47	**85.71**	**20.36**
Noised	Explicit	60.00	32.46	63.33	14.86	46.67	25.09	**78.57**	**22.61**
	Mixed	46.67	31.87	56.67	23.46	43.33	29.43	**76.92**	**21.12**
	Implicit	43.33	26.18	60.00	18.47	40.00	28.08	**82.14**	**19.68**
Short	Explicit	50.00	32.04	46.67	15.70	43.33	34.43	**77.78**	**21.56**
	Mixed	33.33	34.01	50.00	15.57	33.33	37.53	**75.00**	**22.50**
	Implicit	23.33	29.23	33.33	12.89	30.00	29.14	**73.08**	**24.92**

Table 7: Detailed results across all nine ambiguity conditions on REI-Bench with Qwen3.8-Omni-Flash. Higher Success and lower Steps are better.

		ReAct		ReflAct		WALL-E		**Ours**	
Context	RE Type	Succ. ↑	Steps ↓	Succ. ↑	Steps ↓	Succ. ↑	Steps ↓	Succ. ↑	Steps ↓
Standard	Explicit	76.67	36.75	86.67	18.79	63.33	29.88	**77.78**	**11.86**
	Mixed	76.67	33.64	86.67	27.71	70.00	31.98	**88.89**	**15.00**
	Implicit	76.67	38.70	76.67	26.08	73.34	28.53	**77.78**	**13.86**
Noised	Explicit	66.67	33.03	70.00	25.13	66.67	28.87	**66.67**	**10.67**
	Mixed	60.00	39.44	60.00	24.96	63.33	34.45	**66.67**	**9.83**
	Implicit	56.67	42.20	66.67	25.10	60.00	31.69	**77.78**	**15.29**
Short	Explicit	56.67	34.51	43.33	25.05	53.33	31.13	**66.67**	**17.50**
	Mixed	46.67	29.40	53.33	24.51	50.00	29.67	**55.56**	**12.40**
	Implicit	43.33	48.16	40.00	18.77	43.33	40.35	**66.67**	**12.83**

ambiguity sources interact. At the same time, several lower-ambiguity conditions show smaller differences between methods, indicating that the proposed mechanisms are most beneficial when ambiguity becomes consequential for task execution.

Overall, the condition-level results show that the aggregate improvements reported in the main text are not driven by a small subset of cases. Across both backbones, our method remains effective over a broad range of ambiguity combinations, with its largest advantages appearing in the more challenging multi-source settings.

D.3 Detailed Results on REI-Bench

Tables 6 and 7 report detailed results across all nine combinations of context conditions and referring-expression types on REI-Bench. We analyze performance along the two controlled dimensions of the benchmark: context quality and referring-expression type.

DeepSeek-V4.1-Flash. As shown in Table 6, our method achieves the highest task success across all nine ambiguity conditions. Its advantage becomes more pronounced as the available context becomes less informative. Averaged over referring-expression types, Success decreases from 87.81 under Standard context to 79.21 under Noised context and 75.29 under Short context, whereas the baselines exhibit substantially larger degradation. Performance also remains relatively stable across Explicit, Mixed, and Implicit referring expressions, suggesting that the method is robust to different forms of referring-expression ambiguity.

Qwen3.8-Omni-Flash. Table 7 shows a similar trend. Under Standard context, the proposed method is comparable to the strongest baseline, while its advantage becomes clearer under Noised and Short contexts. In particular, our method achieves the highest task success in all three Short-

context conditions, where reduced contextual information makes reference resolution more difficult. Across referring-expression types, performance remains relatively stable, indicating that the benefit is not restricted to a particular expression form.

Across both backbones, the main performance differences emerge as contextual information becomes less reliable or incomplete. While competing methods degrade substantially from Standard to Short contexts, our method retains stronger task success across different referring-expression types. These results complement the aggregate results in the main text and suggest that the proposed ambiguity-resolution framework is particularly beneficial when task interpretation must rely on limited or degraded contextual information.