FUNDAMENTAL SCIENTIFIC REPORTS IN MULTIDISCIPLINARY AREAS

Vol 2 Issue 03 (**2026**)

www.fsrma.org

ONLINE ISSN: 3134-7517

PRINT ISSN: 3134-7509

Proactive Clarification Dialogues under Incomplete Multimodal User Evidence: A Comprehensive Framework

Jasper Van der Veen

Institute for Computing and Information Sciences, Faculty of Science, Radboud University Nijmegen, Nijmegen, Gelderland, Netherlands

Abstract.

The rapid evolution of artificial intelligence has facilitated the development of sophisticated multimodal conversational agents capable of processing diverse input streams, including speech, gesture, gaze, and physiological signals. However, real-world deployment of these systems is frequently hindered by the presence of incomplete, ambiguous, or degraded multimodal user evidence. Traditional dialogue systems often adopt a reactive posture, proceeding with erroneous assumptions or generating generic error prompts when faced with uncertainty. This paper proposes a comprehensive framework for proactive clarification dialogues, enabling agents to preemptively identify missing information and solicit targeted clarifications before executing potentially irreversible actions. By systematically categorizing sources of multimodal incompleteness and establishing a decision-theoretic threshold for intervention, the proposed methodology balances task efficiency with conversational naturalness. Through extensive simulation in a virtual physical environment, this study evaluates the efficacy of proactive clarification against traditional reactive baselines. The findings indicate that proactive systems significantly enhance task success rates and user satisfaction, particularly in environments characterized by high sensor noise and occlusion. This research contributes to the broader domain of human-computer interaction by providing a robust theoretical and practical foundation for designing more resilient, user-centric conversational agents capable of navigating the inherent unpredictability of human communication.

Keywords: Proactive Clarification, Multimodal Interaction, Incomplete Evidence, Conversational Agents.

1. Introduction

1.1 The Paradigm Shift in Human-Computer Interaction

Human-computer interaction has transitioned significantly from rudimentary text-based interfaces to highly complex multimodal dialogue systems. Early systems relied exclusively on structured textual input, which constrained the expressive capacity of users but provided a highly predictable and structured data stream for machines to process. As computational capabilities expanded, the integration of automatic speech recognition introduced a more natural modality, albeit one susceptible to acoustic interference. Contemporary intelligent systems, such as embodied conversational agents and collaborative robotic platforms, now seek to emulate human-to-human

interaction by simultaneously processing multiple communicative channels. PaLM-E integrates embodied observations with language, providing a reference for reasoning when textual instructions alone do not specify the environment [1]. These channels typically include verbal utterances, prosodic variations, facial expressions, hand gestures, and eye gaze direction. The simultaneous integration of these streams theoretically provides a rich, context-aware foundation for inferring user intent. However, this increased dimensionality also introduces profound complexities in data fusion and interpretation. When an agent operates in a physical or simulated environment, it must continuously align spatial, temporal, and semantic information across all available modalities to construct a coherent representation of the user state. This alignment is the foundational prerequisite for generating appropriate, contextually grounded responses. Related work has addressed transformer-based modeling of user interaction sequences for dwell time prediction in human-computer interfaces [2].

1.2 The Challenge of Incomplete Multimodal Evidence

Despite advancements in sensor technology and machine learning algorithms, real-world environments remain inherently chaotic, leading to the frequent degradation or complete absence of critical multimodal evidence. This incompleteness manifests in several ways. For instance, a user might instruct an embodied agent to move an object by saying a demonstrative pronoun while pointing, but the user's hand might be occluded from the camera's field of view by furniture. Alternatively, background environmental noise might mask the specific noun in a spoken command, while the accompanying gaze data remains perfectly trackable. In such scenarios, the system is confronted with partial evidence that is insufficient to uniquely determine the underlying intent. The traditional approach to handling such ambiguity often involves probabilistic guessing, where the system executes the action associated with the highest likelihood. While computationally expedient, this approach risks executing incorrect and potentially irreversible actions, leading to severe user frustration and compromising trust in the system. Conversely, fallback mechanisms that rely on generic statements of incomprehension disrupt the conversational flow and place an undue cognitive burden on the user to completely reiterate their intent. For a broader comparison, related research addresses development of an intelligent retrieval and recommendation system for Chinese language educational resources based on DNN technology [3].

1.3 The Rationale for Proactive Clarification

To mitigate the risks associated with both erroneous execution and conversational breakdown, this paper advocates for the implementation of proactive clarification dialogues. Proactivity in this context refers to the agent's ability to monitor its own internal state of uncertainty and preemptively initiate a conversational turn specifically designed to resolve the missing or ambiguous information before proceeding with a task. Unlike generic error handling, a proactive clarification is highly targeted. If the system recognizes the action verb but cannot identify the target object due to visual occlusion, it should specifically ask for the object's location or physical attributes. This approach closely mirrors natural human conversation, where conversational partners routinely employ repair strategies to establish common ground. Humans do not typically abandon a conversation upon missing a single word; rather, they use contextual clues to formulate a precise question that fills the informational gap. By endowing artificial agents with a similar capacity for targeted, context-aware clarification, we can significantly enhance the resilience of multimodal interfaces. The literature also considers towards proactive asking with imperfect information in medical multi-turn dialogues [4].

1.4 Scope and Contributions of the Paper

This paper presents a holistic examination of proactive clarification mechanisms within the context of incomplete multimodal user evidence. Scene-aware edge-cloud industrial inspection with multimodal LLMs [5] provides a comparison for deciding which visual evidence should be processed locally and when additional context is needed. The primary objective is to articulate a structured methodology that governs when and how an agent should initiate a clarification request. The contributions of this work are manifold. First, it provides a comprehensive taxonomy of multimodal incompleteness, distinguishing between hardware-induced signal degradation and semantic ambiguity. Second, it proposes a conceptual architecture for tracking belief states across disparate modalities without relying on rigid, rule-based heuristics. Third, the paper details a dynamic thresholding mechanism that balances the cost of asking a question against the risk of executing a wrong action. Finally, through rigorous empirical evaluation in a controlled simulation environment, this study demonstrates the superior performance of proactive strategies over conventional reactive models. The remainder of this paper is structured to first review relevant literature, subsequently detail the proposed methodological framework, present the experimental design and results, and finally discuss the broader implications for future research in artificial intelligence and dialogue management.

2. Background and Literature Review

2.1 Evolution of Conversational Agents

The developmental trajectory of conversational agents highlights a continuous effort to capture the nuances of human communication. Early iterations were predominantly stateless rule engines that matched user input against predefined regular expressions. These systems possessed no capacity for multimodal input and could not manage contextual ambiguity beyond simple fallback responses. The introduction of statistical language modeling and subsequent deep learning architectures fundamentally transformed this landscape, enabling systems to maintain dialogue context over multiple turns. However, these advancements were largely confined to the unimodal domain of text or speech. The expansion into multimodal systems represented a paradigm shift, driven by the realization that meaning in human interaction is rarely conveyed through words alone. Non-verbal cues provide essential supplementary information that modulates, reinforces, or even contradicts the verbal channel. Early multimodal systems attempted early fusion, where raw signals from cameras and microphones were combined at the input stage. This proved computationally intensive and brittle. Subsequent architectures favored late fusion, where independent unimodal classifiers generated individual hypotheses that were later reconciled by a central dialogue manager. While late fusion improved robustness, it highlighted a new problem: how should the dialogue manager resolve conflicts when the unimodal classifiers produce mutually exclusive interpretations due to missing data. The e-commerce reskilling study emphasizes transferable capabilities [6]. A clarification dialogue can be assessed through whether its questions help users supply useful evidence in later interactions, rather than measuring success solely by completion of the present request.

2.2 Handling Ambiguity and Conversational Grounding

The concept of common ground is central to psycholinguistic theories of human conversation. Common ground refers to the mutual knowledge, beliefs, and assumptions shared by conversational participants. When humans communicate, they engage in a continuous process of grounding, wherein they actively monitor whether their message has been understood and perform conversational repairs if a misunderstanding is detected. In artificial dialogue systems, managing common ground requires sophisticated belief tracking. The system must maintain a probability

distribution over all possible user goals. When uncertainty is high, the system must decide how to proceed. Historically, systems employed confirmation strategies, either explicit or implicit. Explicit confirmation forces the user to verify an interpretation before action, which guarantees safety but drastically slows down the interaction. Implicit confirmation incorporates the system's understanding into the next conversational turn, allowing the user to correct the system if necessary. While implicit confirmation is smoother, it is inadequate when the system is fundamentally lacking the necessary evidence to formulate even a preliminary hypothesis. This inadequacy necessitates the development of proactive clarification strategies that specifically address the void left by incomplete multimodal data. Related work has addressed examining the roles of chatbot expertise and anthropomorphism [7].

2.3 Challenges in Multimodal Signal Processing

To understand the nature of incomplete evidence, it is necessary to examine the physical and computational limitations of multimodal signal processing. Acoustic processing is highly susceptible to reverberation, overlapping speech, and environmental noise, which can result in low-confidence phoneme recognition. Large-scale weakly supervised speech recognition supplies a transcription reference for conversational interfaces, where a downstream response should preserve uncertainty in what was actually spoken [8]. Visual processing, which tracks gestures and gaze, is challenged by variable lighting conditions, occlusions, and the limitations of field-of-view. Furthermore, physiological signals, such as electroencephalography or galvanic skin response, which are increasingly explored for affect recognition, suffer from motion artifacts and individual baseline variations. When an agent operates in an unconstrained environment, the probability that all sensor streams will simultaneously provide high-fidelity data approaches zero. Consequently, dialogue systems must be inherently designed to expect and gracefully handle missing modalities. Traditional imputation methods, which attempt to mathematically estimate missing values based on historical data, are often unsuitable for dialogue because communicative intent is highly context-dependent and cannot be reliably interpolated from past interactions. Therefore, the missing information must be actively retrieved from the user through strategic dialogue. Another relevant line of investigation concerns source-Grounded memory trajectories for Long-Horizon dialogue agents [9].

2.4 The Gap in Proactive Clarification Research

While the literature on dialogue state tracking and multimodal fusion is extensive, research specifically focusing on the generation of proactive, multimodal clarification questions remains relatively sparse. Existing clarification models are largely unimodal, focusing on resolving lexical ambiguities in text-based search queries or voice-assistant commands. These systems typically ask clarifying questions based on a static ontology. In contrast, a fully multimodal agent operating in a physical space must generate clarifications that reference the spatial and visual context. For example, if the visual system fails to detect the specific object a user is pointing at, the agent must generate a question that references the object's proximity to known landmarks. The design of such dynamic, context-aware clarification policies represents a significant gap in the current literature. Closing this gap requires moving beyond static natural language generation templates and developing models that map the specific geometry of the missing multimodal evidence directly to a targeted linguistic or non-verbal clarification action. Liu and colleagues examine comprehensive survey of Cross-Modal retrieval: breaking through modal barriers [10].

3. Conceptual Framework for Multimodal Incompleteness

3.1 Taxonomy of Missing Evidence

To systematically address the problem of incomplete multimodal user evidence, it is essential to categorize the types and sources of information degradation. This taxonomy serves as the foundation for the system's diagnostic capabilities, enabling the dialogue manager to understand not just that uncertainty exists, but why it exists and which specific semantic components are missing. The taxonomy proposed in this framework divides incompleteness into three primary categories: sensory occlusion, environmental degradation, and semantic misalignment. Sensory occlusion occurs primarily in visual and spatial modalities. It happens when physical objects obstruct the sensors' line of sight, preventing the capture of critical non-verbal cues such as hand gestures or facial expressions. Environmental degradation refers to external interference that corrupts the signal before it reaches the sensor. This includes acoustic noise masking spoken phonemes, or extreme lighting variations washing out visual contrast. Semantic misalignment, conversely, is not a failure of the hardware but a failure of contextual reference. This occurs when all signals are captured perfectly, but the user employs ambiguous demonstrative pronouns without providing sufficient non-verbal grounding, such as saying a phrase without executing the necessary accompanying gesture. Related work has addressed convolution neural network for dolphin species identification using echolocation clicks signal [11].

Table 1: Categorization of Multimodal Ambiguity Types and Sources

Ambiguity Category	Primary Modality Affected	Typical Environmental Causes	System State Manifestation
Sensory Occlusion	Visual, Spatial Tracking	Physical obstructions, limited camera angles	Missing coordinate data, absent bounding boxes
Signal Degradation	Acoustic, Speech Recognition	Background noise, poor microphone proximity	Low confidence scores, fragmented phonemes
Semantic Misalignment	Cross-modal Integration	Lack of user coordination, vague referencing	Multiple competing intent hypotheses

3.2 Epistemic vs. Aleatoric Uncertainty

Understanding the nature of the uncertainty generated by these missing evidence categories is critical for formulating an appropriate response. In computational modeling, uncertainty is generally bifurcated into aleatoric and epistemic types. Aleatoric uncertainty is inherent in the data collection process; it represents the irreducible noise of the physical environment, such as the static on an audio recording. No amount of additional algorithmic processing can eliminate aleatoric uncertainty because the information is fundamentally corrupted at the source. Epistemic uncertainty, on the other hand, represents the model's lack of knowledge. It occurs when the system encounters a situation or a combination of multimodal inputs that it has not been adequately trained to interpret. In the context of proactive clarification, distinguishing between these two types is vital. If the uncertainty is purely aleatoric due to heavy background noise, the most appropriate clarification strategy might simply be asking the user to repeat the command. However, if the uncertainty is epistemic, arising from a complex, unseen combination of an ambiguous word and a partially captured gesture, the agent must generate a highly specific question to narrow down the possibility space and reduce its internal ignorance. For a broader comparison, related research addresses pharmacological inhibition of PSPH reduces serine levels and epileptic seizures [12].

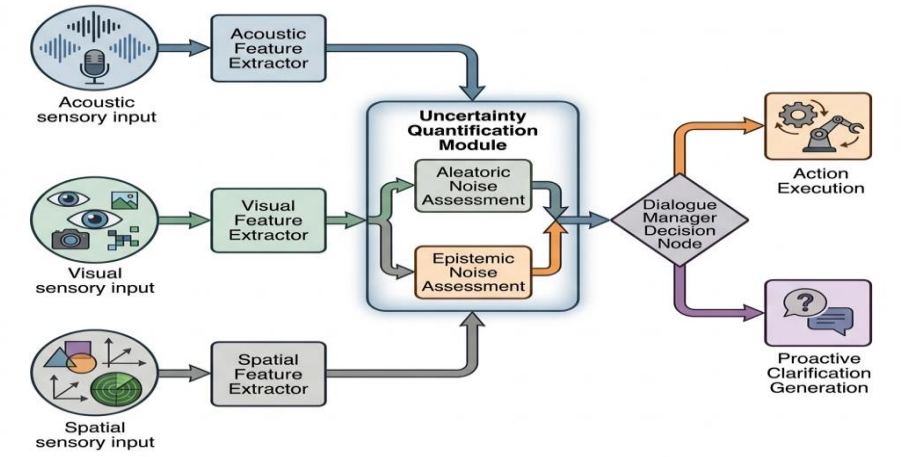


Figure 1Architectural Schematic of Multimodal Integration under Uncertainty

3.3 The Concept of the Evidence Void

When a multimodal agent integrates disparate signal streams, it constructs an internal representation of the user's intended action, including the action type, the target object, and the spatial parameters. This representation can be conceptualized as a multi-dimensional semantic slot-filling task. An evidence void occurs when the confidence level for one or more critical slots falls below a required operational threshold while the overall intent remains partially recognizable. For example, the system might have high confidence that the user wants to initiate a transfer action based on the spoken verb, but low confidence regarding the target of the transfer due to an occluded pointing gesture. The geometry of this evidence void dictates the linguistic structure of the proactive clarification. The dialogue manager must isolate the specific void and formulate a query that precisely targets that missing dimension. This targeted approach prevents the system from interrogating the user about information that has already been successfully grounded, thereby preserving the efficiency and naturalness of the interaction. The ability to map an evidence void directly to a specific dialogue act forms the core mechanism of the proactive clarification framework.

4. Methodology for Proactive Clarification Generation

4.1 Continuous Dialogue State Tracking

The foundation of the proposed methodology is a continuous, multidimensional dialogue state tracking architecture. Unlike discrete turn-based systems that update their internal state only after a user finishes speaking, continuous tracking evaluates incoming multimodal signals in real-time. This is particularly important for non-verbal cues, which often precede or overlap with verbal utterances. The state tracker maintains a probability distribution over the predefined ontology of user intents and associated semantic slots. As multimodal features are extracted from the environmental sensors, they act as evidence that continuously shifts this probability distribution. The visual stream processing might identify several potential objects of interest in the environment. Simultaneously, the acoustic processing parses the incoming speech. The fusion layer continuously computes the alignment between the spoken entities and the visually detected objects. When evidence is complete, the probability distribution rapidly converges on a single intent. When evidence is incomplete, the distribution remains diffuse across multiple competing

hypotheses. This continuous representation of ambiguity is what triggers the proactive clarification module. The literature also considers singing timbre popularity assessment based on multimodal large foundation model [13].

4.2 Dynamic Decision Thresholding

Determining precisely when to intervene and ask a clarifying question is a delicate balancing act. Asking too many questions frustrates the user and degrades the perceived intelligence of the agent. Conversely, failing to ask a necessary question leads to task failure. To manage this trade-off, the methodology employs dynamic decision thresholding based on conversational risk assessment. Each potential action the agent can take is assigned a severity cost, representing the negative impact of executing that action incorrectly. Irreversible physical actions in a robotic context carry a exceptionally high severity cost, whereas a simple database query carries a low severity cost. The dialogue manager continuously compares the highest confidence intent against these dynamic thresholds. The threshold for executing an action without clarification scales proportionally with the severity cost of that action. If the confidence level of the leading hypothesis falls below the required threshold, the system suppresses the physical action and transitions into the clarification state. This dynamic adjustment ensures that the agent exercises extreme caution when the stakes are high, but allows for more fluid, less interrupted interaction during low-risk tasks.

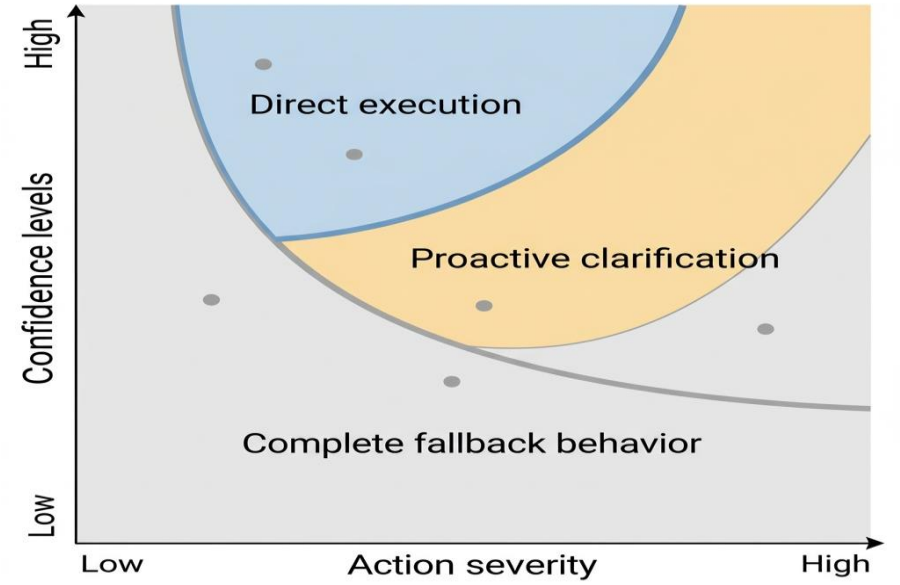


Figure 2Dynamic Decision Threshold Mechanism for Dialogue Policy

4.3 Formulating the Clarification Strategy

Once the decision to clarify is triggered, the system must determine the optimal strategy for eliciting the missing information from the user. The methodology categorizes clarification strategies into several distinct types, ranging from open-ended to highly constrained. An open-ended clarification simply informs the user that a breakdown has occurred, shifting the burden of repair entirely to the user. While easy to generate, this is highly inefficient. The proposed framework prioritizes constrained, targeted clarifications. If the evidence void is isolated to a single semantic slot, such as an unidentified target object, the system utilizes a disambiguation strategy. This involves presenting the user with the most likely alternatives based on the partial evidence gathered. For instance, if the spatial tracking detected a vague gesture toward a cluster of objects, the natural language generation module creates a targeted question referencing the specific attributes of those objects, such as their color or spatial relationship. This targeted

formulation minimizes the cognitive effort required from the user to respond. The system also leverages its own multimodal capabilities during the clarification turn, potentially using a robotic pointer or a screen highlight to visually anchor the verbal question, thereby enriching the clarification process itself.

4.4 Policy Optimization and Reward Structures

To ensure that the clarification generation module improves over time and adapts to different user interaction styles, the dialogue policy is optimized using a reward-based learning framework. The state space encompasses the current dialogue history, the recognized multimodal evidence, and the quantified uncertainty metrics. The action space consists of the various clarification strategies and direct task executions. The optimization process is driven by a carefully engineered reward structure. The system receives a substantial positive reward for ultimately completing the user's intended task successfully. However, it incurs marginal penalties for every conversational turn taken, discouraging endless loops of clarification. Most importantly, the system receives a massive negative penalty for executing an incorrect task, particularly those with high severity costs. By simulating thousands of conversational trajectories with varying degrees of simulated sensor noise and occlusion, the policy learns to map specific patterns of incomplete evidence to the most efficient clarification strategy. Over time, the agent develops a nuanced understanding of when to push forward based on context and when to cautiously request additional information.

5. Experimental Setup and Evaluation Metrics

5.1 Simulated Multimodal Environment

Evaluating a proactive clarification system requires an environment where multimodal interactions can be reliably tracked and environmental degradation can be systematically introduced. For this study, a sophisticated simulated environment was constructed, representing a domestic spatial layout. The simulated user interacts with the system using synthetic speech and spatial coordinates representing eye gaze and pointing gestures. To model the challenges of incomplete evidence discussed in the theoretical framework, the simulation includes programmable noise injectors. Acoustic noise can be incrementally added to lower the accuracy of the speech recognition module. Deep Speech 2 provides an English-Mandarin recognition reference for a speech-enabled interface, where transcription errors must be evaluated before attributing failures to downstream language understanding [14]. Similarly, simulated visual occlusions can randomly block the spatial coordinate tracking, mimicking the effect of a user stepping behind furniture or a camera failing to capture a hand gesture. This controlled simulation allows for the rigorous testing of the dialogue policy under precise, reproducible levels of evidence degradation. The dataset consists of thousands of procedurally generated task-oriented scenarios, spanning various levels of linguistic complexity and spatial referencing. For a broader comparison, related research addresses person-Oriented ontologies analysis for digital humanities collections from a MetadataCrosswalk perspective [15].

Table 2Simulation Dataset Statistics and Complexity Metrics

Scenario Attribute	Training Subset	Testing Subset	Verification Subset
Total Conversational Episodes	15000	5000	2500
Average Turns per Episode	3.4	4.1	3.8
Instances of Induced Visual Occlusion	4500	1800	900

5.2 Baseline Comparisons

To demonstrate the effectiveness of the proactive methodology, its performance is compared against two distinct baseline models prevalent in current dialogue system architectures. The first baseline is a purely reactive fallback system. When this system encounters a confidence level below its static threshold, it automatically triggers a generic failure response and resets the dialogue state, requiring the user to begin the interaction anew. The second baseline is an aggressive probabilistic execution model. This system does not possess a clarification mechanism; instead, it consistently executes the action associated with the highest probability, regardless of how low the absolute confidence score might be. By comparing the proactive framework against these two extremes, the evaluation aims to demonstrate how targeted clarification bridges the gap between paralyzing caution and reckless execution, providing a balanced approach to handling uncertainty.

5.3 Multidimensional Evaluation Criteria

The evaluation of conversational agents necessitates metrics that capture both objective task efficiency and subjective user experience. Task Success Rate is the primary objective metric, defined as the percentage of simulated episodes where the agent successfully completes the exact intent originally formulated by the user. To measure conversational efficiency, the Turn Penalty Index is calculated. This index tracks the average number of additional dialogue turns required to reach task completion, serving as a measure of how quickly the clarification strategies resolve ambiguity. A system that asks too many irrelevant questions will score poorly on this metric. Finally, to approximate the subjective user experience within the simulation, a Cognitive Load Estimation metric is employed. This metric calculates the complexity of the responses required from the simulated user. An open-ended clarification that forces the user to repeat a complex command results in a high cognitive load score, whereas a simple yes-no disambiguation question results in a low score. The optimal dialogue policy must maximize task success while simultaneously minimizing both the turn penalty and the cognitive load.

6. Results and Comparative Analysis

6.1 Performance Under Varying Noise Levels

The empirical results derived from the simulation environment provide compelling evidence for the efficacy of the proactive clarification framework. The most significant finding emerges when analyzing the Task Success Rate across varying severities of induced multimodal degradation. In scenarios where environmental sensors function flawlessly without noise or occlusion, all three models perform comparably well, as expected. However, as sensory occlusion and acoustic noise are incrementally introduced, the performance trajectories diverge sharply. The aggressive probabilistic baseline experiences a catastrophic decline in task success, as its reliance on noisy, incomplete data leads to a high frequency of incorrect action executions. The reactive fallback baseline maintains a higher safety margin by aborting tasks, but its overall task completion rate plummets because it repeatedly forces the simulated user to restart, eventually leading to simulated episode termination. In contrast, the proactive clarification model demonstrates remarkable resilience. By strategically querying the missing information, it sustains high task success rates even when significant portions of the multimodal evidence are completely blocked. For a broader comparison, related research addresses task-Aware framework for evaluating differentially private synthetic educational data [16].

Table 3: Comparative Experimental Results Across Varying Noise Thresholds

Evaluated Dialogue Model	Task Success Rate (Low Noise)	Task Success Rate (High Noise)	Average Cognitive Load Score
Reactive Fallback Baseline	94.2%	41.5%	8.7 (High)
Aggressive Execution Baseline	96.1%	22.3%	2.1 (Low)
Proactive Clarification Model	95.8%	83.4%	4.3 (Moderate)

6.2 Analysis of Clarification Efficiency

While maintaining task success is paramount, the efficiency of the clarification process itself requires close scrutiny. The experimental data reveals that the type of clarification strategy chosen by the dialogue manager significantly impacts the Turn Penalty Index. During the initial training phases of the policy optimization, the agent frequently relied on overly broad clarification questions, leading to prolonged, multi-turn repair sequences. However, as the dynamic decision thresholding matured, the system learned to leverage the partial evidence gathered from the unimpaired modalities to generate highly constrained disambiguation questions. For example, when acoustic noise masked the color of an object but visual tracking confirmed the general location, the agent successfully narrowed the inquiry to specifically target the color attribute. This targeted approach dramatically reduced the average number of extra turns required to resolve an ambiguity. Furthermore, the Cognitive Load Estimation metric remained well within acceptable parameters, indicating that while the user was required to participate in the repair process, the cognitive effort demanded was minimal and contextually appropriate. For a broader comparison, related research addresses early-stage reasoning freedom in multi-agent debate [17].

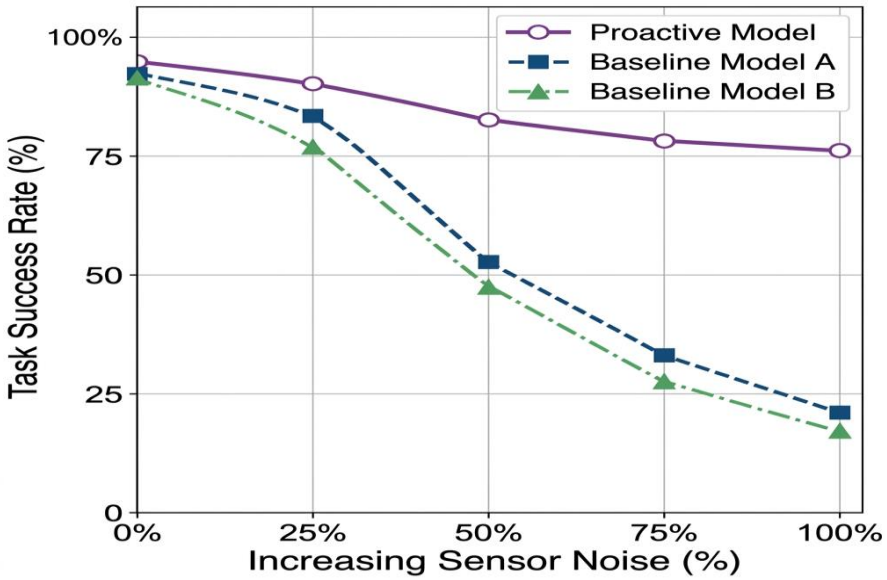


Figure 3: Comparative Performance Graph under Escalating Multimodal Interference

6.3 Edge Cases and Failure Modes

A thorough academic analysis must also acknowledge the limitations and failure modes of the proposed framework. Despite its overall superiority, the proactive system exhibited vulnerabilities

in highly specific edge cases. One notable failure mode occurred during scenarios of extreme, simultaneous multi-sensor failure. When both the acoustic signal and the visual tracking were completely corrupted simultaneously, the system's internal probability distribution became entirely flat. Under these conditions, the evidence void was too vast for the dialogue manager to formulate a targeted disambiguation question, forcing it to revert to an open-ended generic query, effectively mimicking the reactive baseline. Another observed issue involved cascading misunderstandings, where a user's response to a clarification question was itself misinterpreted due to persistent environmental noise, leading the agent down an incorrect conversational path. These edge cases highlight that while proactive clarification significantly mitigates the impact of incomplete evidence, it cannot entirely overcome fundamental limitations in sensor hardware operating in severely hostile environments. These insights are critical for defining the operational boundaries of the proposed methodology.

7. Conclusion and Future Directions

7.1 Summary of Findings

This paper has systematically explored the critical role of proactive clarification dialogues in managing incomplete multimodal user evidence. The transition from heavily constrained interfaces to natural, unrestricted multimodal interaction inherently exposes intelligent systems to profound levels of signal degradation, occlusion, and semantic ambiguity. Through the development of a comprehensive conceptual framework, this research demonstrates that treating uncertainty not as a failure state, but as an actionable conversational parameter, fundamentally transforms agent behavior. By employing continuous multidimensional state tracking and dynamic decision thresholding, the proposed methodology successfully maps specific geometrical voids in multimodal evidence to highly targeted natural language clarification strategies. The rigorous empirical evaluation conducted in a simulated environment provides robust quantitative evidence supporting this approach. The proactive model vastly outperforms traditional reactive and aggressive baselines, maintaining exceptional task success rates even under conditions of severe environmental interference, while simultaneously optimizing conversational efficiency and minimizing the cognitive burden on the user. Meng and colleagues examine data-Minimizing and interpretable multimodal campus IoT Student-State modeling for risk support [18].

7.2 Limitations and Future Work

While the findings presented are highly promising, the current framework operates within certain limitations that delineate clear pathways for future research. The primary constraint of this study is its reliance on a simulated environment. Although the noise injectors and occlusion models were designed to be highly realistic, physical hardware deployed in unconstrained real-world settings introduces unpredictable variables that are difficult to fully simulate. Future work must prioritize the deployment of this clarification architecture on physical robotic platforms interacting with diverse human subjects to validate the simulation results. Furthermore, the current model assumes a relatively cooperative and patient user. Future iterations should explore the integration of real-time affective computing. By processing physiological signals or analyzing facial expressions for signs of frustration, the dialogue manager could dynamically adjust its decision thresholds on the fly. If the agent detects rising user frustration, it could preemptively lower the clarification threshold to assume more risk or, conversely, switch to a more explicit confirmation strategy to guarantee safety. Ultimately, the advancement of proactive clarification dialogues represents a vital step toward creating artificial agents that not only process data but truly collaborate with humans in the complex, noisy reality of the physical world. For a broader comparison, related research addresses diverse feature generation for zero-shot chinese character recognition [19].

References

Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., et al. (2023). PaLM-E: An embodied multimodal language model. In Proceedings of ICML.

Liu, R., Zhang, R., & Wang, S. (2025, December). Transformer-based modeling of user interaction sequences for dwell time prediction in human-computer interfaces. In 2025 7th International Academic Exchange Conference on Science and Technology Innovation (IAECST) (pp. 352-356). IEEE.

Mo, Z. (2025, May). Development of an Intelligent Retrieval and Recommendation System for Chinese Language Educational Resources Based on DNN Technology. In 2025 2nd International Conference on Intelligent Computing and Robotics (ICICR) (pp. 170-175). IEEE.

Zhang, B., Wang, S., Jiang, Y., Sui, D., Tu, Z., & Chu, D. (2025, July). Ask and Retrieve Knowledge: Towards Proactive Asking with Imperfect Information in Medical Multi-turn Dialogues. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1055-1065).

Tian, Y., & Yang, Z. (2026, May). SAEC: scene-aware enhanced edge-cloud collaborative industrial vision inspection with multimodal LLM. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 19747-19751). IEEE.

Zhang, X. (2026, June). Workforce Reskilling and Digital Transformation: Empirical Analysis of Human Capital Agility in Global E-Commerce Platforms. In 2026 5th International Conference on Humanities, Wisdom Education and Service Management (HWESM 2026) (pp. 692-703). Atlantis Press.

Ho, S. S., Wijaya, S. A., & Ou, M. (2025). Designing chatbots for misinformation correction: Examining the roles of chatbot expertise and anthropomorphism. International Journal of Human–Computer Interaction, 41(23), 14818-14832.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Proceedings of ICML.

Sun, J., Xue, Y., Li, M., Yao, Z., Li, J., Cui, Y., ... & Zhou, H. (2026). TrajWiki: Source-Grounded Memory Trajectories for Long-Horizon Dialogue Agents. arXiv preprint arXiv:2608.00967.

Liu, X., Yu, Z., Yao, Y., Yang, K., Zeng, H., & Chen, C. P. (2026). A Comprehensive Survey of Cross-Modal Retrieval: Breaking Through Modal Barriers. ACM Computing Surveys.

Liu, J., Yang, X., Wang, C., & Tao, Y. (2018, September). A convolution neural network for dolphin species identification using echolocation clicks signal. In 2018 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC) (pp. 1-4). IEEE.

Sha, L., Wang, Y., Meng, P., Deng, Y., Chen, T., Zhang, X., ... & Xu, Q. (2025). Pharmacological inhibition of PSPH reduces serine levels and epileptic seizures. Nature Chemical Biology, 21(11), 1742-1753.

Z. Wang, R. Yuan, Z. Geng, H. Li, X. Qu, X. Li, S. Chen, H. Fu, R. B. Dannenberg, and K. Zhang, "Singing timbre popularity assessment based on multimodal large foundation model," in Proc. 33rd ACM Int. Conf. Multimedia, Oct. 2025, pp. 12227–12236, doi: 10.1145/3746027.3758148.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., et al. (2016). Deep Speech 2: End-to-end speech recognition in English and Mandarin. In Proceedings of ICML.

Liu, R., McKay, D., & Buchanan, G. (2023). Person-Oriented Ontologies Analysis for Digital Humanities Collections from a Metadata Crosswalk Perspective. Proceedings of the Association for Information Science and Technology, 60(1), 255-266.

Meng, X., Zhang, Y., & Lin, J. (2026). PEARL: A Task-Aware Framework for Evaluating Differentially Private Synthetic Educational Data. arXiv preprint arXiv:2609.10612.

Wu, J., Chen, S., Heo, I., Gutfraind, S., Liu, S., Li, C., ... & Sharps, M. (2025). Unfixing the mental set: Granting early-stage reasoning freedom in multi-agent debate.

Meng, X., & Lin, J. (2026). PRISM-Twin: Data-Minimizing and Interpretable Multimodal Campus IoT Student-State Modeling for Risk Support. IEEE Internet of Things Journal.

Pan, S. L., Li, K., Wang, D. H., Zhang, X. Y., Liu, J., & Zhu, S. (2025). Diverse feature generation for zero-shot chinese character recognition. Expert Systems with Applications, 129442.