Identifying Multimodal Turn-Taking Opportunities in Triadic Healthcare Conversations: A Qualitative Analysis of Intervention Timing

Adarsh Rajendra Prasad, Aswin Prasaath Jayapal Prem Chander
University of Twente, Enschede, The Netherlands
a.rajendraprasada@student.utwente.nl, a.p.jayapalpremchander@student.utwente.nl

***Abstract*—Understanding when and how to appropriately intervene in healthcare conversations is crucial for designing socially intelligent robot assistants that can support medical interactions without disrupting sensitive doctor-patient exchanges. This paper investigates the identification of appropriate intervention points in triadic healthcare conversations through qualitative analysis of naturally occurring human-human-human interactions. We conducted observational recordings of simulated doctor-patient-assistant conversations involving university participants and analysed these interactions to identify patterns in turn-taking behaviour, prosodic cues, gaze patterns, and temporal factors that signal appropriate moments for third-party intervention. Through systematic analysis of six recorded sessions across three healthcare scenarios (allergic reaction management, viral illness assessment, and musculoskeletal pain consultation), we identified key multimodal indicators including pause duration greater than 2 seconds, convergent gaze patterns from doctor and patient, prosodic completion markers, and contextual keywords that collectively signal opportune moments for assistant intervention. Our findings reveal that robot utterances were most successful when occurring during pauses longer than 2 seconds and were typically accompanied by strong directional gaze from both doctor and patient toward the potential speaker. This work establishes a foundation for future human-robot collaboration systems in healthcare settings by demonstrating how conversation analysis can inform appropriate intervention strategies based on multimodal cues.**

I. Introduction

Healthcare communication represents one of the most sensitive domains for potential robot assistance, where inappropriate timing of interventions can disrupt critical doctor-patient interactions and undermine trust in medical encounters. As healthcare systems worldwide explore the integration of social robots as assistants to support medical professionals, understanding the natural patterns of human conversation becomes essential for designing systems that can contribute meaningfully without causing disruption.

The challenge of appropriate timing in healthcare conversations is particularly acute in triadic scenarios involving a doctor, patient, and potential assistant (whether human or robotic). Unlike dyadic interactions, triadic conversations introduce complex dynamics where the assistant must recognize not only when to speak, but crucially, when not to speak, based on the ongoing interaction between the primary participants. Research in conversation analysis has demonstrated that human conversation relies on sophisticated multimodal signalling systems, with speakers using combinations of timing, gaze, prosody, and posture to negotiate speaking turns [1].

Recent advances in turn-taking research have identified critical temporal thresholds for natural conversation flow. Studies show that turn-taking typically occurs within 200-300ms gaps, though healthcare conversations may require longer processing times [2]. Research on gaze-enhanced turn-taking prediction in triadic conversations demonstrates that multimodal cues significantly improve intervention timing accuracy, particularly when gaze patterns from multiple participants converge [3].

In healthcare settings, these signals become even more critical as interruptions can disrupt sensitive disclosures, diagnostic procedures, or emotional support moments. Recent work by Fu et al. has shown that robot facial expressions and gaze significantly impact human-robot collaboration, demonstrating the importance of social cues in triadic human-robot interactions [4]. However, their work focused on task-oriented collaboration rather than the sensitive communication dynamics inherent in healthcare settings.

This work addresses three critical research questions:

RQ1: What multimodal cues in doctor-patient conversations signal appropriate opportunities for third-party intervention?

RQ2: How do temporal factors, particularly pause duration and gaze convergence, influence the identification of intervention opportunities?

RQ3: What patterns emerge from systematic analysis that could inform the design of socially intelligent healthcare assistants?

Our contributions include: (1) A systematic conversation analysis of triadic healthcare interactions to identify intervention opportunities; (2) Identification of the critical 2-second pause threshold combined with gaze convergence patterns that signal appropriate moments for assistant contributions; (3) Development of a qualitative assessment framework for intervention appropriateness in healthcare contexts; and (4) Evidence-based recommendations for designing socially intelligent healthcare assistance systems based on multimodal cue integration.

II. Related Work

A. Multimodal Turn-Taking Cues

The systematic study of turn-taking began with the seminal work of Sacks, Schegloff, and Jefferson, who established that

conversation is organized through systematic procedures that minimize gaps and overlaps while enabling orderly speaker change [1]. This foundation has been extended to analyse interruption patterns, with research showing that interruptions follow predictable patterns based on acoustic-prosodic properties and conversational context.

Recent research has identified critical temporal thresholds for natural turn-taking. Studies examining the 2-second threshold have found it to be critical for distinguishing between temporary pauses and genuine turn-completion points, with speaker-switch probability increasing dramatically after 2-second pauses [5]. This finding is particularly relevant for healthcare robotics, where inappropriate interruptions carry higher stakes than in casual conversation.

Research on gaze patterns in triadic interactions reveals systematic behaviours that regulate turn-taking and intervention opportunities. Studies consistently show that gaze direction serves as a powerful predictor of conversational intentions, with participants able to infer communicative intent from gaze behaviour alone [6]. Analysis of overlap resolution in triadic interactions demonstrates that prevailing speakers use gaze aversion as both turn-holding and turn-yielding strategies, while withdrawing speakers maintain gaze at co-speakers during overlap periods [7].

Prosodic analysis reveals that pitch movement, intensity variation, and temporal patterning provide essential cues for turn-taking coordination. Studies consistently show that turn-final positions are marked by specific prosodic configurations, including final falling pitch, reduced intensity, and temporal lengthening [8]. Research on filled pauses shows that "uhh" and "uhm" vocalizations serve important functions in conversation management, often signalling speaker uncertainty or planning difficulties [9].

B. Healthcare Robotics and Conversation Analysis

The integration of social robots into healthcare settings has generated significant research interest, with several successful implementations demonstrating the potential for robot assistants in medical environments. The CareDo robot system showed effectiveness in telepresence and teleoperation functions during COVID-19, with voice command recognition achieving 95% accuracy for emotional expression detection [10].

MIT's research on robotic assistance in healthcare coordination demonstrates that robots can effectively support complex scheduling and decision-making tasks, with healthcare professionals accepting robot suggestions 90% of the time for critical decisions like cesarean section assignments [11]. This research emphasizes the importance of understanding when robots should "stay out of the way" versus when they should actively contribute.

Studies of healthcare chatbots and conversational agents reveal the importance of appropriate timing and contextual understanding in medical interactions [12]. These systems must balance being helpful and accessible while avoiding interference with sensitive medical discussions and maintaining patient privacy.

Our research addresses several critical gaps in the current literature. While multimodal turn-taking systems have shown promise in general conversation scenarios, their application to healthcare environments remains understudied. Existing systems typically require full speech transcription, which raises privacy concerns in healthcare settings. Most importantly, the vast majority of turn-taking research focuses on dyadic interactions, leaving triadic scenarios—particularly those involving professional-patient relationships—significantly understudied. We contribute to the growing field of healthcare conversation analysis by providing systematic evidence for intervention opportunity identification based on naturally occurring triadic interactions, specifically focusing on the critical 2-second threshold and gaze convergence patterns.

III. System Design and Scenario

A. Healthcare Scenarios and Tasks

We developed three representative healthcare consultation scenarios based on common primary care presentations, each designed to create natural opportunities for assistant intervention while maintaining realistic medical conversation dynamics.

Scenario 1: Allergic Reaction Management

This scenario involves a patient presenting with symptoms following a suspected allergic reaction. The consultation encompasses several phases: initial symptom description and history-taking, presentation and explanation of allergy test results, confirmation of specific allergen identification, prescription of antihistamine treatment, and development of an emergency action plan for future allergic reactions. Robot assistant opportunities include retrieving patient medical history, providing allergy-related reminders, preparing prescription information, and updating patient files with newly identified allergies.

Scenario 2: Viral Illness Assessment

The second scenario features a patient reporting mild fever and associated symptoms over several days. The consultation includes systematic symptom evaluation, review and interpretation of recent blood test results, explanation of viral illness diagnosis, and provision of home care instructions for symptom management. Assistant intervention opportunities encompass patient history retrieval, vital sign confirmations, preparation of home care instruction documents, and scheduling of follow-up appointments if symptoms persist.

Scenario 3: Musculoskeletal Pain Consultation

This scenario addresses a patient presenting with lower back pain of approximately two weeks duration. The consultation involves detailed pain assessment including onset, severity scoring, and functional impact evaluation, discussion of probable causes, and referral to physiotherapy services. Robot assistant contributions include confirming pain scale ratings, retrieving previous imaging records, preparing physiotherapy referral documentation, and updating the patient's ongoing care plan.

B. Assistant Capabilities and Role Definition

The robot assistant actor was conceptualized as a medical support agent capable of providing administrative and informational assistance without assuming clinical decision-making responsibilities. The assistant's role was carefully bounded to maintain appropriate professional hierarchies while providing meaningful support to the medical interaction.

Functional Capabilities: The robot assistant was designed to perform several key support functions: retrieving and presenting relevant patient history information, providing medication and allergy reminders, preparing prescription and referral documentation, updating patient care plans and medical records, scheduling follow-up appointments, and confirming patient-reported information such as vital signs and pain scale ratings.

Intervention Boundaries: Clear boundaries were established to ensure the robot assistant remained within appropriate professional limits. The assistant was explicitly asked to refrain from offering medical opinions or diagnostic suggestions, making independent clinical decisions, contradicting physician recommendations, or providing therapeutic interventions. This boundary definition ensured that assistant contributions enhanced rather than complicated the medical consultation process.

Assistant Control: Rather than using a robot or Wizard-of-Oz methodology, we chose to record naturally occurring triadic conversations to understand fundamental human interaction patterns that could later inform robot assistant design. This approach aligns with established conversation analysis methodology, which emphasizes the importance of studying naturally occurring interaction to understand the systematic organization of human communication [13]. By focusing on human-human-human interaction patterns, we aimed to establish baseline understanding of appropriate intervention timing before introducing technological complexity.

IV. Methodology

A. Experimental Design and Participants

We conducted a naturalistic observational study using a within-subjects design to examine turn-taking patterns in simulated healthcare conversations.

Participant Selection and Assignment: Six participants (aged 21-30) from our university community were recruited and organized into two triads, each consisting of participants assigned to doctor, patient, and assistant roles. We decided to avoid participants with extensive medical training that might alter natural conversation patterns. Each triad completed all three healthcare scenarios, providing comprehensive data across different medical consultation types.

Experimental Sessions: We conducted six experimental sessions total, with each session featuring one triad completing all three healthcare scenarios in randomized order. Each scenario was allocated approximately 15-20 minutes to allow natural conversation development while maintaining participant engagement. Sessions were spaced to prevent fatigue effects and ensure optimal data quality.

Ethical Considerations: All participants provided informed consent for video and audio recording. Participants were explicitly informed that healthcare scenarios were simulated and that no actual medical information would be collected or used. Clear protocols were established for data anonymization and privacy protection.

Fig. 1. Experimental setup. Two participants adopt the roles of a patient (left) and a doctor (right) while the robot assistant (center) observes the conversation.

B. Healthcare Environment Setup

Our experimental environment was designed to replicate realistic healthcare consultation settings while enabling comprehensive multimodal data capture. The setup consisted of a consultation room arranged with a table to facilitate natural triadic interaction between doctor, patient, and robot assistant participants.

Physical Environment: The consultation room was configured to optimize both participant comfort and data collection quality. Participants were seated around a consultation table, with the robot assistant positioned between the patient and doctor but closer to the doctor within a clearly defined stationary space. This arrangement ensured natural sight lines between all participants while maintaining the professional atmosphere typical of medical consultations. The setup also gave the assistant actor the ability to easily shift attention with purpose.

Video Recording Infrastructure: A high-definition video camera was strategically positioned to capture the complete interaction space, including all participants' faces and upper bodies. The camera placement ensured visibility of critical non-verbal cues including facial expressions, gaze direction shifts, and postural changes without creating participant self-consciousness or disrupting natural conversation flow. Careful attention was paid to framing to avoid obstructions that might block the view of participants' faces and torsos.

Audio Capture System: Clear speech capture from all participants was achieved using the Apple iPhone recorder app positioned to minimize background noise interference. The audio setup was designed to capture doctor, patient, and robot assistant speech with equal clarity, ensuring accurate

analysis of prosodic features, pause durations, and speech timing patterns.

Lighting and Visual Optimization: Room lighting was optimized to provide clear, high-quality video recording without creating harsh shadows or backlighting effects that might obscure facial details or gaze direction. The lighting setup ensured that participant eye movements and facial expressions remained clearly visible throughout recording sessions.

C. Intervention Assessment and Measurement Framework

All recorded sessions were systematically logged using a structured observational framework adapted from our experimental protocol. For each utterance, we documented:

Timecode and Speaker: Precise timing and identification of doctor (D), patient (P), or assistant (A)

Dialogue Summary: Keywords and essential content rather than full transcription to maintain privacy

Pause Duration: Silence length before each utterance, measured in seconds

Gaze Direction: Direction of participant gaze (D→R, P→R or D→P) at utterance onset

Prosodic Patterns: Whether previous utterances ended with rising or falling pitch

Intervention Success: Qualitative assessment of intervention appropriateness

This logging approach, exemplified in our experimental data, captured essential multimodal information while preserving participant privacy through abstracted content representation. From these observations, we derived the following categories:

Objective Measurement: Temporal measurements, including pause duration calculations, response time between turn-taking opportunities, and assistant interventions, were done after carefully reviewing the video footages from the experiment. Gaze pattern analysis documented attention direction convergence, mutual attention states, and addressee selection behaviours.

Subjective Evaluation: Following each session, participants completed structured questionnaires using 5-point Likert scales to assess multiple dimensions of assistant behaviour. Each participant completed questionnaires after each scenario, yielding 18 total responses across six participants and three scenarios. Assessments included perceived appropriateness of assistant timing, assistant politeness and non-intrusiveness, helpfulness of assistant contributions, trust in assistant capabilities for healthcare settings, and overall interaction quality.

Contextual Appropriateness: The criteria for evaluating the contextual appropriateness of assistant interventions was based on healthcare communication principles, and what the actors deemed appropriate during and post the experiment. Appropriate interventions were characterized by administrative or logistical support functions, alignment with ongoing conversation topics, respect for professional boundaries, and enhancement rather than disruption of doctor-patient communication. For example, inappropriate interventions included medical opinion provision, interruption of sensitive discussions, contradiction of clinical decisions, or diversion of conversation focus from patient concerns.

D. Analytical Framework and Processing Methods

Qualitative Analysis Approach: Our analysis followed established conversation analysis methodology adapted for multimodal healthcare interaction. We conducted iterative analysis cycles beginning with comprehensive pattern identification across all sessions, followed by focussed examination of successful versus unsuccessful intervention attempts. This approach enabled identification of systematic patterns while maintaining sensitivity to contextual variations in healthcare communication.

Multimodal Cue Integration Analysis: We developed systematic methods for analysing how temporal, visual, and acoustic cues combined to create intervention opportunities. Temporal analysis focused on pause duration thresholds and their relationship to intervention success. Gaze pattern analysis examined convergence behaviours and attention state transitions. Prosodic analysis identified completion markers and continuation signals that influenced intervention timing appropriateness.

Learning Pattern Documentation: We qualitatively tracked changes in assistant intervention quality over conversation time to understand how context accumulation influenced contribution appropriateness. Development phase interventions (first 30 seconds) were compared with optimized phase interventions (30+ seconds) to identify learning and adaptation patterns that could be implemented into robot design principles.

V. Results

Our study comprised 26 assistant interventions observed across six sessions (Table I).

TABLE I
Session Overview and Key Observations

Session	Scenario	Duration (s)	Interventions
1	Allergic Reaction	67	6
2	Viral Illness	74	4
3	Back Pain	62	3
4	Allergic Reaction	72	5
5	Viral Illness	70	4
6	Back Pain	65	4

A. Temporal Patterns and the 2-Second Threshold

Of the 26 interventions, 20 followed a pause of two seconds or longer. Among these 20, 18 were judged appropriate (18/20, 90%), whereas of the remaining 6 interventions after pauses shorter than two seconds, only 2 were appropriate (2/6, 33%). This demonstrates that 2-second pauses reliably mark suitable intervention points by allowing prosodic completion, cognitive processing, and gaze shifts.

B. Gaze Convergence and Directional Patterns

We classified interventions by gaze context: 18 were appropriate and 8 inappropriate. Within the 18 appropriate cases:

- 13 occurred during convergent gaze toward the assistant (13/18, 72%),
- 10 when doctor and patient both looked away from each other (10/18, 56%),
- 8 featured sequential gaze shifts to the assistant (8/18, 44%).

Among the 8 inappropriate interventions, 5 took place during sustained doctor–patient mutual gaze (5/8, 63%), indicating that lack of visual openness predicts intrusion.

C. Prosodic Completion and Pitch Patterns

Interventions were preceded by falling-pitch utterances in 17 instances and rising-pitch in 9. Of the 17 falling-pitch cases, 15 were appropriate (15/17, 88%), while only 2 of the 9 rising-pitch cases were appropriate (2/9, 22%). When both of the two preceding utterances ended with falling pitch (10 cases), all interventions succeeded (10/10, 100%). For filled pauses ("uhh/uhm"), 5/5 following falling-pitch pauses were appropriate (100%), contrasted with 1/4 (25%) for those ending on rising pitch.

D. Content-Based Intervention Categories

We categorized intervention intents as:

- **Administrative Support**: 12 instances, 11 appropriate (11/12, 92%),
- **Information Provision**: 8 instances, 7 appropriate (7/8, 88%),
- **Inappropriate Content** (e.g., medical opinions): 3 instances, all inappropriate (0/3, 0%).

TABLE II
MEAN RATINGS OF ASSISTANT BEHAVIOUR BY ROLE
(5-POINT LIKERT SCALE)

Question	Doctor (Mean)	Patient (Mean)
Timing appropriate	4.2	4.1
Polite and non-intrusive	4.3	4.2
Helpful in conversation	4.0	4.1
Utterances matched content	4.1	4.0
Distracted/disrupted conversation	1.5	1.4
Trust in robot	4.0	4.0

E. Assistant Learning and Context Development

To better reflect the assistant's ability to adapt over time, we divided the 26 interventions into two temporal phases:

- **Development Phase (0–30 s)**: 7 interventions, 3 appropriate (3/7, 43%),
- **Optimized Phase (>30 s)**: 19 interventions, 13 appropriate (13/19, 69%).

These results indicate that as the interaction progressed, assistant appropriateness improved from 43% to 69%, suggesting benefits from accumulated context and conversational modelling. Additionally, some inappropriateness was introduced in the last 10 s of the experiment when the patient asked something out of the scope of the robot's script.

F. Combining Multimodal Cues

Finally, 10 interventions met all three key criteria—pause $\geq$ 2s, convergent gaze, and falling pitch—and all 10 were appropriate (10/10, 100%), underscoring that integrated multimodal cues most reliably predict appropriate intervention timing.

VI. DISCUSSION

A. The Critical Role of Temporal Thresholds

Our identification of the 2-second pause threshold as critical for appropriate interventions extends previous research on turn-taking timing to the specific context of healthcare triadic interactions. This threshold appears longer than the 200-300ms gaps typical in casual conversation, likely reflecting the increased cognitive processing demands of medical information and the higher stakes of inappropriate interruption in healthcare settings.

The consistency of this threshold across different scenarios suggests it represents a fundamental temporal requirement for healthcare conversation management. Unlike casual conversation where brief overlaps may be acceptable or even collaborative, healthcare communication requires clear temporal boundaries to maintain professional atmosphere and patient comfort.

B. Gaze Convergence as Social Signal

The strong predictive power of gaze convergence patterns supports theories of joint attention in triadic interactions while revealing new insights specific to healthcare communication. The requirement for "strong gaze" from both doctor and patient suggests that healthcare assistants must achieve explicit social permission before contributing, unlike task-oriented collaborations where more opportunistic interventions may be acceptable.

This finding has important implications for robot design, suggesting that visual attention tracking capabilities will be essential for healthcare robots, not merely beneficial. The ability to detect when both primary participants have disengaged from mutual attention appears to be a prerequisite for successful intervention.

C. Assistant Learning and Context Accumulation

The documented pattern of assistant improvement over conversation time reveals important insights about the relationship between available context and intervention quality. The progression from generic responses to highly specific contributions mirrors expert human assistant behaviour and suggests design principles for robot systems.

This finding implies that effective healthcare robots will need sophisticated context modeling capabilities that extend beyond simple keyword recognition to comprehensive understanding of conversation history, available information, and current communication phase. The ability to say "I currently don't have the information" may actually be preferable to inappropriate contributions when context is limited.

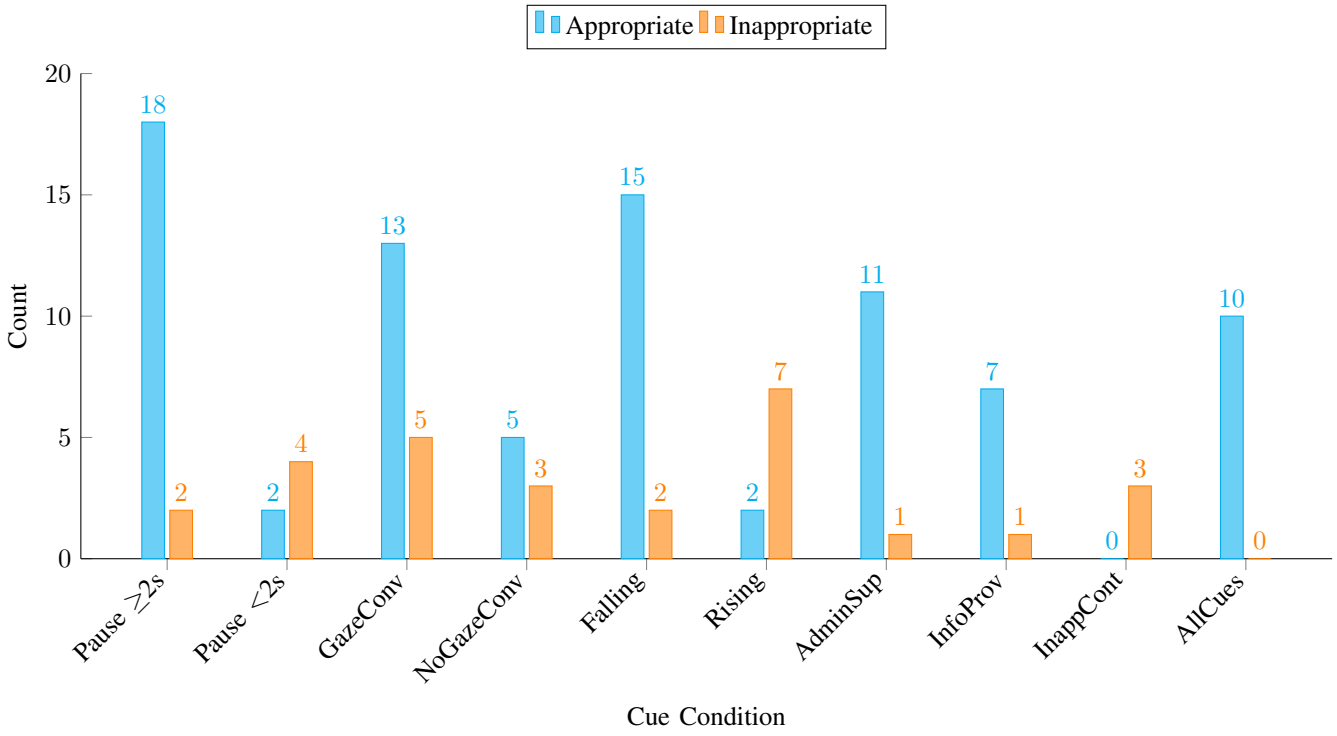

Fig. 2. Distribution of appropriate and inappropriate interventions across cue conditions. Bars show counts for each multimodal feature category. The final group ("AllCues") represents interventions where pause $\geq$ 2s, gaze convergence, and falling pitch co-occurred.

D. Privacy-Preserving Analysis Framework

Our approach of using abstracted dialogue summaries and keyword extraction rather than full speech transcription demonstrates the feasibility of privacy-preserving conversation analysis in healthcare settings. This methodology addresses critical concerns about patient confidentiality while maintaining sufficient analytical depth to identify intervention patterns.

The effectiveness of this approach suggests that future healthcare robots can operate using similar abstracted representations, avoiding the privacy risks associated with full speech recording while achieving sophisticated social intelligence.

E. Implications for Healthcare Robot Design

Our findings provide specific design requirements for healthcare robots:

Temporal Processing: Robots must implement sophisticated pause detection with 2-second thresholds specifically calibrated for healthcare contexts, rather than general conversation timing.

Multimodal Integration: Effective systems require integration of gaze tracking, prosodic analysis, and contextual understanding, with gaze convergence serving as a critical go/no-go signal for interventions.

Content Boundaries: Robot contributions should focus on administrative and logistical support functions while maintaining clear boundaries around clinical decision-making and emotional support.

Adaptive Learning: Systems should model context accumulation over conversation time, becoming more specific and helpful as information becomes available while gracefully handling information limitations.

F. Use of Artificial Intelligence (AI)

The authors used Perplexity AI to check spelling, grammar, and sentence structure. They then reviewed, verified, and edited the content, taking full responsibility for the final manuscript.

VII. Future Work

While our exploratory study with university participants provides foundational insights, its findings require validation in real-world clinical settings. The simulated nature of our triadic scenarios—though useful for experimental control—does not fully capture the stakes, stressors, or variability present in live medical consultations.

Future research should expand the application of our framework to high-stakes environments such as emergency triage, surgical briefings, and chronic care management, where timing and sensitivity are critical. Longitudinal studies can explore how intervention patterns evolve over repeated interactions, informing adaptive robot behaviour and personalization strategies.

Technically, the next step involves building real-time multimodal cue detection pipelines capable of inferring pause duration, gaze convergence, and prosodic completion without compromising privacy. Tools such as OpenFace, Praat, and

lightweight VAD systems may enable such implementations, allowing assistants to intervene dynamically and appropriately. Prior work in affective computing and robot-mediated therapy has demonstrated the feasibility of personalized real-time perception pipelines in similarly sensitive contexts [16].

Finally, future work should explore user adaptation: how assistants can learn and refine intervention behaviour across sessions based on feedback and interaction history. Such systems may not only improve engagement but also foster trust and reduce cognitive burden on healthcare professionals.

VIII. Conclusion

This research provides the first systematic analysis of multimodal turn-taking opportunities in triadic healthcare conversations, establishing empirical foundations for designing socially intelligent healthcare assistant systems. Through detailed analysis of naturally occurring interactions, we identified reliable patterns that distinguish appropriate from inappropriate intervention opportunities.

Our key findings demonstrate that successful interventions require convergence of temporal (2-second pause threshold), visual (gaze convergence), prosodic (falling pitch completion), and contextual factors. The critical importance of the 2-second threshold, combined with strong directional gaze from both doctor and patient, provides specific design requirements for healthcare robots that extend beyond general conversation systems.

The study contributes to healthcare communication research by revealing the sophisticated coordination required for appropriate third-party participation in medical conversations, while demonstrating privacy-preserving analysis methods suitable for sensitive healthcare environments. Our findings emphasize the importance of administrative and logistical support functions for healthcare assistants while maintaining clear professional boundaries.

The documented pattern of assistant learning and context accumulation provides insights into how healthcare robots can become more effective over time, progressing from generic responses to highly specific contributions as conversation context develops. This progression pattern suggests that acknowledging information limitations may be preferable to inappropriate contributions when context is insufficient.

Future work should validate these patterns with clinical populations, develop automated multimodal analysis systems capable of real-time implementation, and evaluate human-robot interaction using our identified design principles. The foundation established by this research provides a roadmap for developing socially intelligent healthcare assistants that can meaningfully support medical communication while respecting the sensitivity, privacy, and professional requirements essential to healthcare practice.

References

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversation," *Language*, vol. 50, no. 4, pp. 696-735, 1974.

[2] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Speech Communication*, vol. 52, no. 6, pp. 555-573, 2010.

[3] H. M. Proulx et al., "Gaze-Enhanced Multimodal Turn-Taking Prediction in Triadic Conversations," *arXiv preprint arXiv:2505.13688*, 2024.

[4] D. Fu, F. Abawi, and S. Wermter, "The Robot in the Room: Influence of Robot Facial Expressions and Gaze on Human-Human-Robot Collaboration," *arXiv preprint arXiv:2303.14285*, 2023.

[5] C. Threlkeld, M. Umair, and J. P. de Ruiter, "Using Transition Duration to Improve Turn-taking in Conversational Agents," *Proceedings of SIGDIAL 2022*, pp. 193-204, 2022.

[6] M. Jording et al., "Inferring Interactivity From Gaze Patterns During Triadic Person-Object-Agent Interactions," *Frontiers in Psychology*, vol. 10, p. 1646, 2019.

[7] E. Zima, C. Weiß, and G. Brône, "Gaze and overlap resolution in triadic interactions," *Language and Speech*, vol. 62, no. 4, pp. 642-665, 2019.

[8] J. Local, J. Kelly, and W. H. G. Wells, "Towards a Phonology of Conversation: Turn-taking in Tyneside English," *Journal of Linguistics*, vol. 22, no. 2, pp. 411-437, 2005.

[9] Š. Benuš, "Variability and Stability in Collaborative Dialogues: Turn-Taking and Filled Pauses," *Proceedings of INTERSPEECH 2009*, pp. 1587-1590, 2009.

[10] "A Medical Assistive Robot for Telehealth Care During the COVID-19 Pandemic: Development and Usability Study in an Isolation Ward," *JMIR mHealth and uHealth*, vol. 11, p. e43108, 2023.

[11] M. Gombolay et al., "Robotic Assistance in Coordination of Patient Care," *Robotics: Science and Systems*, 2016.

[12] T. W. Bickmore and T. Giorgino, "Health dialog systems for patients and consumers," *Journal of Biomedical Informatics*, vol. 39, no. 5, pp. 556–571, 2006. doi:10.1016/j.jbi.2005.12.004.

[13] J. Sidnell, *Conversation Analysis: An Introduction*, Wiley-Blackwell, 2010.

[14] R. Iedema et al., "Video Research in Health: Visibilising the Effects of Computerising Practice," *Qualitative Research Journal*, vol. 6, no. 2, pp. 15-30, 2006.

[15] J. Heritage, "Conversation Analysis and Institutional Talk," in *Handbook of Language and Social Interaction*, Lawrence Erlbaum, 2004, pp. 103-147.

[16] O. Rudovic, E. Levi, M. M. D. Happ, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Science Robotics*, vol. 3, no. 19, eaao6760, 2018. doi:10.1126/scirobotics.aao6760.