RingMind: Interpreting Multimodal Behavior from Ring-Based Speech and Gesture using Large Language Models

Weihao Chen
Department of Computer Science and Technology
Tsinghua University
Beijing, China
chenwh20@mails.tsinghua.edu.cn

Yukun Wang
Department of Computer Science and Technology
Tsinghua University
Beijing, China
wang-yk21@mails.tsinghua.edu.cn

Meizhu Chen
School of Architecture
Tsinghua University
Beijing, China
cmz23@mails.tsinghua.edu.cn

Zhe He
Department of Computer Science and Technology
Tsinghua University
Beijing, China
hz23@mails.tsinghua.edu.cn

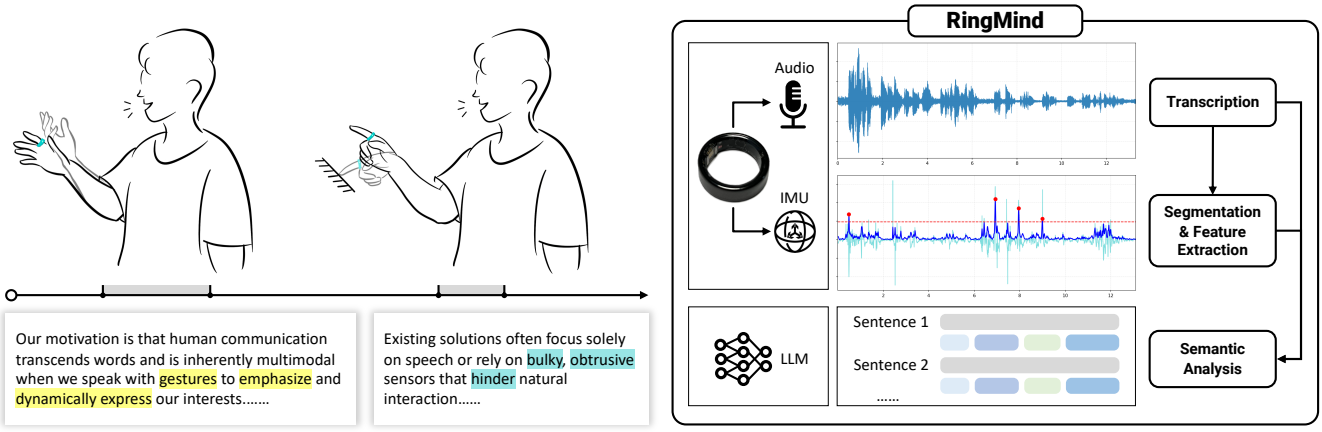

Figure 1: RingMind uses a smart ring with IMU and microphone sensors to capture co-speech gestures and audio, enabling multimodal interaction analysis through LLM-based semantic reasoning.

Abstract

In everyday communication, people combine speech with spontaneous gestures to convey rich, layered meaning. However, capturing and interpreting these multimodal signals in natural contexts remains challenging. We present RingMind, a prototype system that leverages a smart ring with an IMU and a microphone to unobtrusively capture co-speech gestures and audio. By aligning sensor data with transcribed speech and employing large language models (LLMs) for semantic reasoning, RingMind generates structured text reports that highlight expressive moments, summarize conversations, and infer user intent and affect. This work offers a lightweight, context-aware approach for understanding multimodal interaction in real-world settings.

UbiComp Companion '25, Espoo, Finland

ACM ISBN 979-8-4007-1477-1/2025/10
https://doi.org/10.1145/3714394.3750592

CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; **Gestural input**; **Natural language interfaces**.

Keywords

Multimodal Analysis; Co-speech Gestures; Smart Rings; Large Language Models; Context-Aware Computing

ACM Reference Format:
Weihao Chen, Yukun Wang, Meizhu Chen, and Zhe He. 2025. RingMind: Interpreting Multimodal Behavior from Ring-Based Speech and Gesture using Large Language Models. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '25), October 12–16, 2025, Espoo, Finland.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3714394.3750592

1 Introduction

Human communication extends beyond spoken language and is inherently multimodal: we speak, gesture, emphasize, and dynamically express interest and intent [8]. However, seamless methods for capturing and interpreting this multimodal synergy—especially

outside controlled laboratory environments—remain limited. Existing approaches often focus exclusively on speech, thereby missing valuable nonverbal context, or they rely on bulky, obtrusive sensor setups that interfere with natural interaction. This disconnect hinders the broader integration of embodied AI technologies into everyday life, a crucial step toward realizing more intuitive, human-centered ubiquitous computing.

To address this, we consider the use of a smart ring platform. Smart rings offer a uniquely promising form factor for next-generation AI wearables due to their unobtrusive, always-on design and their ability to passively capture both core communicative signals and rich contextual cues during natural interactions [14]. This positions them as an ideal modality for inferring a user's social context and communicative intent in everyday, in-the-wild settings.

Despite these advantages, several key challenges persist. Multimodal behavior in open-ended, real-world scenarios is highly variable and cannot be reliably interpreted using a fixed gesture vocabulary or predefined labels. The communicative meaning of gestures often depends on subtle contextual factors, may shift dynamically, or may carry implicit emotional or interpersonal nuance that is difficult to detect using traditional rule-based methods.

To tackle this challenge, we explore the use of large language models (LLMs) for open-ended semantic interpretation. By converting raw signals from different modalities (e.g., motion and speech) into structured textual descriptions, and organizing them in context-aware formats, LLMs can serve as powerful reasoning engines capable of deriving deeper meaning from complex multimodal interactions [3, 4, 10].

We propose RingMind, a prototype system that captures and interprets multimodal behavioral signals—specifically speech and spontaneous hand gestures—using a smart ring equipped with an inertial measurement unit (IMU) and a microphone. By combining time-synchronized signal processing with the semantic reasoning capabilities of large language models (LLMs), RingMind is designed to extract enriched insights into human expression during natural conversations.

RingMind is envisioned as a personal assistant that can be activated by the user during everyday social interactions—such as informal discussions, meetings, presentations, or passive listening—to record multimodal behavioral data in situ. After the session, the system automatically generates a structured, text-based report that includes (but is not limited to):

- Highlighted segments of the speech transcript where gestural expressiveness is particularly high;
- Extractive and abstractive summaries of the spoken content;
- Inferred signals related to the user's interests, emotional states, and communicative traits.

By embedding such semantic insights into natural workflows, RingMind supports a new class of multimodal, lightweight, and meaning-aware wearable applications—enabling more reflective, contextually grounded, and emotionally intelligent computing experiences.

2 Background

2.1 Co-speech Gestures

Co-speech gestures—spontaneous hand movements that accompany speech—play a crucial role in enriching verbal communication by conveying emphasis, emotion, and semantic nuance [8]. Prior research has explored their classification and functional significance, but most systems rely on camera-based setups in constrained environments (e.g., [6]). Recent studies have begun to leverage IMUs for gesture recognition in wearable contexts, yet few have examined their integration with speech to support real-time interpretation of communicative intent. Our work contributes to this space by capturing and analyzing co-speech gestures unobtrusively via a smart ring, enabling more naturalistic multimodal understanding.

2.2 Smart Rings for Multimodal Sensing

Most existing studies on smart rings primarily focus on leveraging a single sensor to perform specific tasks—for instance, using an IMU for planar trajectory reconstruction [5, 11], employing PPG for heart rate monitoring [2], or utilizing active acoustic sensing to infer hand poses [12]. Although some ring-based systems have adopted multimodal sensing, their applications are mostly limited to physiological monitoring. In contrast, we aim to explore how microphones and IMUs on a smart ring can be jointly leveraged to infer a user's state during speech, with the assistance of large language models (LLMs).

2.3 Context-Aware Semantic Analysis using LLMs

LLMs have demonstrated strong capabilities in capturing user intent and interpreting natural language based on situational context. When combined with real-time human activity data from wearable or ubiquitous sensors, they can generate adaptive and personalized responses [3, 4, 10]. LLMs also facilitate emotion inference and emotionally supportive dialogue, aligning with emerging research in context-aware, emotion-driven recommendation and interaction systems [7, 13]. These advances position LLMs as foundational components for next-generation systems that interpret multimodal cues and support rich, adaptive user experiences.

3 RingMind

Figure 2: The smart ring used in our system.

The RingMind system consists of a smart ring and a Bluetooth-connected PC. The ring is worn on the user's index finger (Figure 2) and is equipped with a 6-axis IMU (ICM-42688P) and a microphone (MP23DB01HP). A Python-based backend runs on the PC to receive and process raw sensor data streamed from the ring. A web-based frontend displays the analyzed results.

3.1 Sense

The IMU samples data at approximately 200 Hz, and the microphone records audio at around 40 kHz. Users initiate and terminate recording of both channels via clicks on the web interface. Upon completion of a session, the analysis pipeline is automatically triggered to process the collected data.

3.2 Analysis Pipeline

We process the raw sensor data from both channels (audio and IMU) to generate a comprehensive report comprising both sentence-level and overall analyses of the recorded session. This process requires internet access, as it invokes external APIs for semantic interpretation.

3.2.1 Automatic Speech Recognition (ASR). The audio stream is first transcribed into word-level timestamped text using an ASR service. In our current implementation, we adopt Alibaba's Paraformer V2 API for this task [1].

3.2.2 IMU Data Segmentation and Behavioral Feature Extraction. The IMU data is segmented into distinct windows aligned with the sentence-level time spans obtained from ASR. For each segment, we extract a set of behavioral features, including binary motion state (stationary vs. active), the presence and timing of physical events (e.g., taps or impacts), and orientation-related information such as detected flips or rotations (roll, pitch, yaw). We also render visualizations of the IMU signals on the frontend interface, presenting time-aligned charts of motion patterns, posture shifts, and impact events.

3.2.3 Semantic Analysis with LLMs. The outputs from the preceding steps are transformed into structured textual prompts, which are fed into an LLM for high-level semantic interpretation. We use OpenAI's GPT-4.1 model [9] (with a temperature setting of 0.2) to perform this analysis.

We instruct the LLM to perform two layers of analysis:

- Sentence-level reasoning: For each utterance, the model provides an interpretation of the user's likely intention and inferred physical behavior corresponding to that sentence.
- Session-level summary: The model generates an overall analysis of the user's behavioral patterns, communicative intent, attitude, and emotional tone across the entire session.

To support this, we construct a temporally ordered prompt that contains all ASR sentences, and we embed within each sentence a corresponding gesture descriptor. This descriptor includes the motion state, posture details, detected taps or impacts, and the words temporally associated with those events.

This structured prompting approach creates a flexible analysis framework that can be extended to incorporate additional behavioral signals or higher-level semantic tasks.

Finally, the LLM-generated analysis is parsed and rendered on the frontend interface. Users can click on individual sentences to explore inferred motivations and intent, or browse the aggregated summary for a holistic view of the interaction.

4 Case Study: Poster Presentation

We illustrate the application of RingMind in a university poster presentation setting, where a student introduces their research to peers and faculty members. This scenario is particularly relevant for testing real-time, naturalistic multimodal interaction: the presenter speaks dynamically, emphasizes key ideas with hand gestures, occasionally points to visual elements, and reacts to listener engagement. Importantly, these behaviors unfold spontaneously and are often subtle, making them challenging to capture through traditional methods such as cameras or manual annotation.

In this case, the student wears a smart ring embedded with a microphone and an IMU sensor. The ring allows for continuous, unobtrusive sensing of both speech and hand movements throughout the presentation, without disrupting the speaker's natural flow.

4.1 Analysis Results

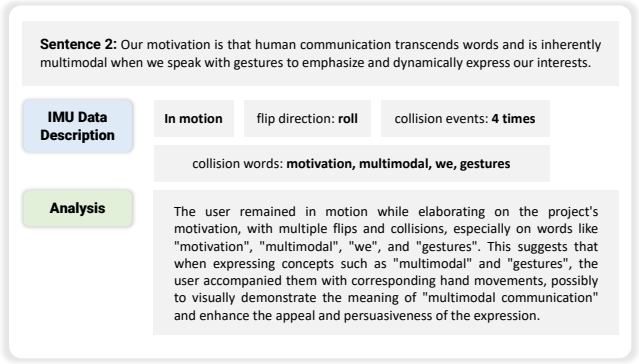

Figure 3: Example result at the sentence level.

To interpret the presenter's behavior, we structure the analysis results into two complementary levels: sentence-level micro analysis and overall behavioral summarization.

At the sentence level (see Figure 3), RingMind combines IMU features—such as motion state, hand roll orientation, and ring-surface collision events—with synchronized speech segments. Using LLMs, it generates contextualized interpretations of each utterance. For example, it identifies emphatic gestures aligned with keywords like "motivation" or "gestures," indicating the speaker's intentional stress and cognitive focus.

Beyond individual sentences, RingMind synthesizes the interaction into high-level behavioral patterns (see Figure 4). It recognizes that the speaker tends to use larger gestures and more fluent movements when expressing confidence or enthusiasm (e.g., describing system contributions), while periods of stillness and repetitive tapping coincide with critique or problem statements. Emotional tone, communicative intent, and presentation rhythm are all inferred from these cross-modal patterns.

4.2 Insights

These multimodal insights unlock a range of applications. For example, presentation feedback systems can highlight moments of strong emphasis or detect hesitation based on mismatches between

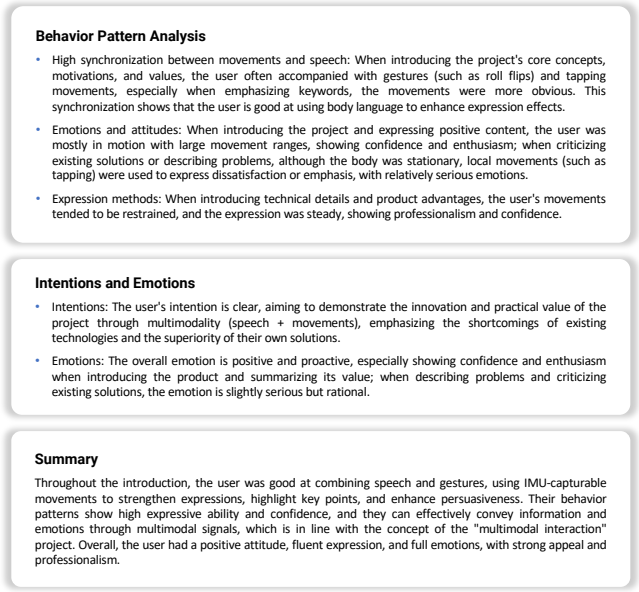

Figure 4: Example overall report.

speech and gesture fluency. Behavior modeling can identify habitual communication patterns—such as a user's tendency to tap when expressing critique—enabling personalized coaching or reflection. Emotion tracking becomes more robust when subtle hand motions are integrated with prosodic speech features, improving affective computing in real-world settings. Moreover, accessibility aids can leverage such insights to infer intent in users with speech or motor impairments, supporting expressive communication through partial signals.

Beyond this case study, the same framework generalizes to naturalistic, socially embedded scenarios involving speech-gesture coordination—not only in individual expression (e.g., remote teaching, scientific pitching, or guided tours), but also in multi-party contexts such as discussions, meetings, or professional consultations. RingMind enables fine-grained, context-aware interpretation of subtle communicative behaviors across these settings.

5 Discussion

Ethical Considerations. RingMind collects and analyzes speech and motion data, which may contain sensitive information. To protect privacy, recordings are user-initiated only, with no always-on listening. A visible LED indicates recording status to bystanders. Behavioral insights are presented as optional interpretations rather than definitive labels to ensure user agency. While early prototypes use cloud-based LLMs, we aim to support local processing on smartphones or laptops to minimize data exposure.

Future Work. We plan to assess the long-term wearability of the ring in daily use, including its social acceptability in various settings. Additionally, we aim to validate the semantic accuracy of RingMind's interpretations across diverse, real-world conversational contexts. Future iterations will also explore improved multimodal fusion techniques and real-time feedback.

Acknowledgments

We thank Weinan Shi for his guidance and Chengchi Zhou for his support in ring development. This work is supported by Beijing Key Lab of Networked Multimedia and Beijing National Research Center for Information Science and Technology (BNRist).

References

[1] Keyu An, Zerui Li, Zhifu Gao, and Shiliang Zhang. 2024. Paraformer-v2: An improved non-autoregressive transformer for noise-robust speech recognition. doi:10.48550/arXiv.2409.17746 arXiv:2409.17746 [eess].

[2] Assim Boukhayma, Anthony Barison, Serj Haddad, and Antonino Caizzone. 2021. Ring-Embedded Micro-Power mm-Sized Optical Sensor for Accurate Heart Beat Monitoring. *IEEE Access* 9 (2021), 127217–127225. doi:10.1109/ACCESS.2021.3111956

[3] Weihao Chen, Yuanchun Shi, Yukun Wang, Weinan Shi, Meizhu Chen, Cheng Gao, Yu Mei, Yeshuang Zhu, Jinchao Zhang, and Chun Yu. 2025. Investigating Context-Aware Collaborative Text Entry on Smartphones using Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3706598.3713944

[4] Weihao Chen, Chun Yu, Huadong Wang, Zheng Wang, Lichen Yang, Yukun Wang, Weinan Shi, and Yuanchun Shi. 2023. From Gap to Synergy: Enhancing Contextual Understanding through Human-Machine Collaboration in Personalized Systems. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 110, 15 pages. doi:10.1145/3586183.3606741

[5] Zhe He, Zixuan Wang, Chun Yu, Chengwen Zhang, Xiyuan Shen, and Yuanchun Shi. 2025. WritingRing: Enabling Natural Handwriting Input with a Single IMU Ring. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 731, 15 pages. doi:10.1145/3706598.3714066

[6] Xiyun Hu, Dizhi Ma, Fengming He, Zhengzhe Zhu, Shao-Kang Hsia, Chenfei Zhu, Ziyi Liu, and Karthik Ramani. 2025. GesPrompt: Leveraging Co-Speech Gestures to Augment LLM-Based Interaction in Virtual Reality. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. Association for Computing Machinery, New York, NY, USA, 59–80. doi:10.1145/3715336.3735769

[7] Nils Klüwer, Irina Nalis, and Julia Neidhardt. 2025. Context over Categories: Implementing the Theory of Constructed Emotion with LLM-Guided User Analysis. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 151, 7 pages. doi:10.1145/3706599.3721205

[8] David McNeill. 1995. Hand and Mind. In *Advances in Visual Semiotics: The Semiotic Web 1992-93*, Thomas A. Sebeok and Jean Umiker-Sebeok (Eds.). De Gruyter Mouton, Berlin, Boston, 351–374. doi:doi:10.1515/9783110874259.351

[9] OpenAI. 2025. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/

[10] Aurora Polo-Rodríguez, Laura Fiorini, Erika Rovini, Filippo Cavallo, and Javier Medina-Quero. 2025. Enhancing Smart Environments with Context-Aware Chatbots using Large Language Models. arXiv:2502.14469 [cs.CL] https://arxiv.org/abs/2502.14469

[11] Xiyuan Shen, Chun Yu, Xutong Wang, Chen Liang, Haozhan Chen, and Yuanchun Shi. 2024. MouseRing: Always-available Touchpad Interaction with IMU Rings. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 412, 19 pages. doi:10.1145/3613904.3642225

[12] Anandghan Waghmare, Ishan Chatterjee, and Shwetak Patel. 2023. Z-Pose: Continuous 3D Hand Pose Tracking Using Single-Point Bio-Impedance Sensing on a Ring. In *Proceedings of the 2nd Workshop on Smart Wearable Systems and Applications* (Madrid, Spain) *(SmartWear '23)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3615592.3616851

[13] Zhiyuan Wang, Katharine E. Daniel, Laura E. Barnes, and Philip I. Chow. 2025. CALLM: Understanding Cancer Survivors' Emotions and Intervention Opportunities via Mobile Diaries and Context-Aware Language Models. arXiv:2503.10707 [cs.CL] https://arxiv.org/abs/2503.10707

[14] Zeyu Wang, Ruotong Yu, Xiangyang Wang, Jiexin Ding, Jiankai Tang, Jun Fang, Zhe He, Zhuojun Li, Tobias Röddiger, Weiye Xu, Xiyuxing Zhang, huan-ang Gao, Nan Gao, Chun Yu, Yuanchun Shi, and Yuntao Wang. 2025. Computing with Smart Rings: A Systematic Literature Review. doi:10.48550/arXiv.2502.02459 arXiv:2502.02459 [cs] version: 1.