Real-Time Multimodal Processing for Interpreting Embodied Actions

Hannah VanderHoeven, Videep Venkatesha, Abhijnan Nath, and Nikhil Krishnaswamy
Colorado State University
Fort Collins, CO USA
hannah.vanderhoeven@colostate.edu

Abstract

In this paper, we demonstrate how real-time integration of language with embodied gesture and action in a collaborative task enables the generation of AI agent interventions that result in "positive friction", or reflection, deliberation, and more mindful collaboration. Further, we demonstrate how the same framework can be adapted toward agent action generation for real-time task guidance.

1. Introduction

As artificial intelligence has become increasingly integrated into various workflows, there has been consistent interest in creating flexible agents with the ability to collaborate with humans across a wide range of diverse domains. To achieve this goal, there is a need for deeper, more nuanced understanding of human expressivity on the part of the agent. Human communication extends far beyond words or visual cues—it is inherently embodied, involving a rich combination of language, gesture, movement, and other embodied signals, all of which humans intuitively interpret based on lived experience and embodied cognition [3]. The ubiquity of language data in multimodal pretraining [8], especially compared to other embodied communicative channels, can limit an AI system's ability to understand and respond to human behavior in context, and can thus lead it to be somewhat biased, particularly toward linguistic input, especially when in real-time interactions with humans.

In this paper, we explore how a real-time interpretation of embodied communicative signals contributes to high quality agent interventions. We specifically examine human-AI collaborative task settings with LLM-driven agents whose actions consist of dialogue "interventions" throughout the task. Our results show how using multimodal embodied signals in interpreting tasks dialogues allows LLM-driven agents to generate higher-quality interventions that are judged to have higher impact on the collaborative task. We conclude with future directions and use-cases for multimodal interpretation of embodied actions for adaptive, context-aware human-AI collaboration.

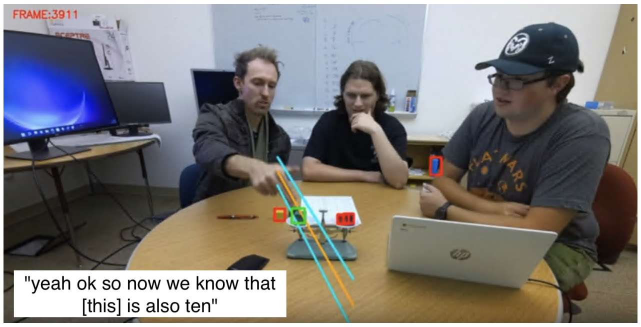

Figure 1. Participants performing the Weights Task [5, 6] with detected deixis, objects, and the associated raw utterance overlaid.

2. Methodology

We use the TRACE platform [13] for multimodal *common ground tracking* [7] as the foundation for our approach. TRACE enables the creation of a contextually complete dialogue history that merges natural language with embodied signals such as gesture, body language, and actions over task-relevant items. E.g., in situated dialogue, objects are often referenced with demonstratives ("this," "that one," etc.). To fully interpret statements using these values an agent would require knowledge of what is being referenced at any given time; in conjunction with demonstrative pronouns, this information is typically provided through gesture or action (e.g., pointing to or manipulating an object). TRACE integrates these channels as gesture and object "features" wherein the outputs of a gesture detector and an object detector are unified to determine which objects in the task context are the likely denotata of deixis or foci of actions [12]. This list of objects is then overlapped with the output of the speech channel, and the demonstrative tokens are replaced with the specific names of the objects, creating a *multimodal dense paraphrase* (MMDP) [11]. Fig. 1 shows a visual example in a collaborative reasoning task known as the Weights Task (WTD; [5, 6]), in which the raw utterance is augmented with gesture and object locations. Here, a participant is seen pointing at the blue block, and so TRACE augments the utterance to "yeah ok so now we know that [blue block] is also ten."

Raw Friction	Embodied Friction
We're assuming this block is either 10 or 20, but what if it's something entirely different? Have we considered other possibilities?	We can't assume the green block's weight is the sum of the blue and red blocks' weights just yet. What if there's another combination of blocks that adds up to 20?

Table 1. Example "raw" and "embodied" friction interventions.

3. Experiments

How does live interpretation of embodied communication signals help LLMs act as better collaborators? Within a generation-as-action-taking framework [9], embodied signals provide better "perception" for the perception-action loop, even when the AI "action" is limited to text generation, as our experiments show. We focused on the problem of inserting "positive friction" into collaborative task dialogues (e.g., [5, 6]), which consists of interventions intended to prompt user reflection on goals [4]. We gave subsets of raw dialogue histories and their equivalents augmented with embodied signals to an instance of LLaMA 3-8B Instruct [1] that was aligned with Direct Preference Optimization [10], and prompted it to generate positive friction as described above. We compared the respective outputs—"raw" friction vs. "embodied" friction (Table 1)—using a reward-modeling framework [2] in which outputs are scored using an OPT 1.3B-based model that was fine tuned to judge how strongly a friction intervention prompts participants to reassess their assumptions. Higher values indicate greater predicted impact. Across all 10 groups of the WTD, we compared the respective average rewards for friction outputs generated by the model when given "raw" and "embodied" dialog inputs. Table 1 shows differences in an example output (associated with a dialogue history containing the utterance shown in Fig. 1) under each condition. The output associated with embodied signals is more specific in its guidance, citing specific blocks and combinations to experiment with. Fig. 2 shows the distribution of average groupwise rewards for generated interventions across all 10 WTD groups. A paired t-test shows that the interventions returned given embodied inputs have statistically significant higher average reward (p=0.023).

4. Use Cases

Sec. 3 showed how embodied inputs improved the quality of interventions returned from an LLM-driven agent. Using TRACE's flexible custom feature processing, this general approach serves as a basis for multiple different active efforts in both embodied group work analysis and explicit task guidance. In task guidance, multimodal signals can be interpreted to verify that each individual stage of a physical task has been successfully completed, allowing agent interventions as necessary to support proper execution. Fig. 3 shows an individual washing their hands, as in preparation for a medical examination. Relevant signals for this task may include hand motion relative to objects such as the soap dispenser and faucet. Integration of embodied features would enable an agent to infer if a task has been completed successfully (such as hand-washing to WHO specifications[1]), and intervene when required with higher-quality feedback.

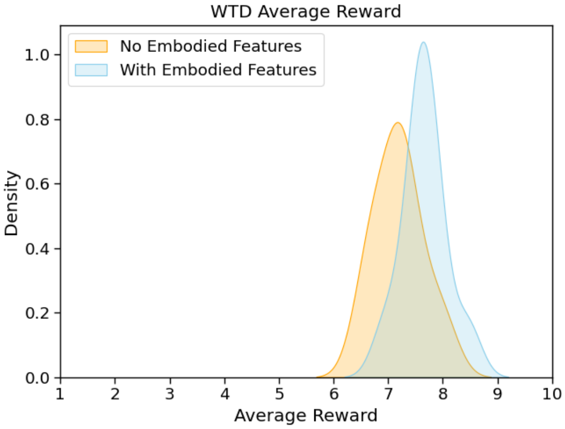

Figure 2. Kernel Density Estimation plot showing average per-group reward values for interventions returned by the LLM agent.

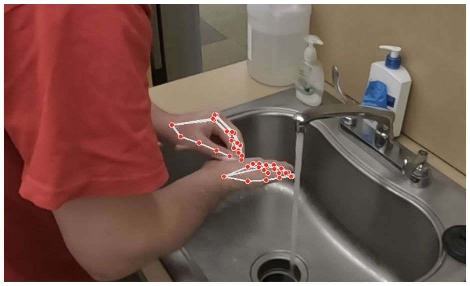

Figure 3. Example of potential task guidance, in which an individual is washing their hands, as in preparation for a medical examination. Hand overlay is detected using MediaPipe [14].

5. Conclusion

In this paper we present a method for integrating language, gestures and actions in real time during collaborative tasks to create detailed embodied signals. We show that these signals allow AI agents to generate higher quality outputs, specifically interventions that support encouraging reflection, thoughtful decision-making, and more mindful collaboration. Additionally, we illustrate how this framework can be adapted to support a variety of different uses, from task analysis to real-time task guidance.

[1] https://www.who.int/publications/m/item/how-to-handwash

Acknowledgments

This material is based in part upon work supported by Other Transaction award HR00112490377 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program, and by Other Transaction award 1AY2AX000062 from the U.S. Advanced Research Projects Agency for Health (ARPA-H) Platform Accelerating Rural Access to Distributed Integrated Medical Care (PARADIGM) program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

[1] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2

[2] Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, 2024. 2

[3] Autumn B Hostetter and Martha W Alibali. Visible embodiment: Gestures as simulated action. *Psychonomic bulletin & review*, 15:495–514, 2008. 1

[4] Mert İnan, Anthony Sicilia, Suvodip Dey, Vardhan Dongre, Tejas Srinivasan, Jesse Thomason, Gökhan Tür, Dilek Hakkani-Tür, and Malihe Alikhani. Better slow than sorry: Introducing positive friction for reliable dialogue systems. *arXiv preprint arXiv:2501.17348*, 2025. 2

[5] Ibrahim Khebour, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Mariah Bradford, Brittany Cates, Paige Hansen, Changsoo Jung, Brett Wisniewski, Corbyn Terpstra, Leanne Hirshfield, Sadhana Puntambekar, Nathaniel Blanchard, James Pustejovsky, and Nikhil Krishnaswamy. The weights task dataset: A multimodal dataset of collaboration in a situated task. 2023. 1, 2

[6] Ibrahim Khebour, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Mariah Bradford, Brittany Cates, Paige Hansen, Changsoo Jung, et al. When text and speech are not enough: A multimodal dataset of collaboration in a situated task. *Journal of open humanities data*, 10, 2024. 1, 2

[7] Ibrahim Khalil Khebour, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard A Brutti, Christopher Tam, Jingxuan Tu, Benjamin A Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, et al. Common ground tracking in multimodal dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3587–3602, 2024. 1

[8] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023. 1

[9] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 2

[10] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023. 2

[11] Jingxuan Tu, Kyeongmin Rim, Bingyang Ye, Kenneth Lai, and James Pustejovsky. Dense Paraphrasing for Multimodal Dialogue Interpretation. *Frontiers in Artificial Intelligence*, 7, 2024. 1

[12] Hannah VanderHoeven, Nathaniel Blanchard, and Nikhil Krishnaswamy. Point target detection for multimodal communication. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management.* Springer, 2024. 1

[13] Hannah VanderHoeven, Brady Bhalla, Ibrahim Khebour, Austin Youngren, Videep Venkatesha, Mariah Bradford, Jack Fitzgerald, Carlos Mabrey, Jingxuan Tu, Yifan Zhu, et al. Trace: Real-time multimodal common ground tracking in situated collaborative dialogues. *arXiv preprint arXiv:2503.09511*, 2025. 1

[14] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020. 2