Modelling Social Intentions in Complex Conversational Setings

Kondyurin, Ivan

DOI
10.1145/3678957.3688614

Publication date
2024

Document Version
Final published version

Published in
ICMI 2024 - Proceedings of the 26th International Conference on Multimodal Interaction

Modelling Social Intentions in Complex Conversational Settings

Ivan Kondyurin
Department of Intelligent Systems, EEMCS, TU Delft
Delft, The Netherlands
i.kondyurin@tudelft.nl

Abstract

Interpreting and managing social interactions is vital for social well-being, yet existing technologies fall short, particularly in group settings. This research aims to develop advanced machine perception systems for Social Signal Processing to accurately model human social behavior. Our multi-modal generative model aims to integrate multi-modal sensory data input data, contextual information and subjective observers' narratives, utilizing them as complex input to an adapted Large Language Model, and producing plausible narratives that reflect various human perspectives. This human-centered approach leverages both low-level cues and high-order events, ensuring adaptability to diverse observers and contexts. The model's potential areas of application include cross-cultural interactions, social group integration, and professional meetings, enhancing social harmony and productivity.

Keywords

Human-centered computing; Social Signal Processing; Large Language Models

ACM Reference Format:
Ivan Kondyurin. 2024. Modelling Social Intentions in Complex Conversational Settings. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 04–08, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3678957.3688614

1 Introduction

The ability to interpret and manage social interactions is paramount to enhancing social well-being and harmony [3]. Despite the substantial time we spend engaging in face-to-face conversations, existing technologies fall short in aiding us to navigate these social encounters effectively [9]. This shortfall is particularly pronounced in understanding and modeling social intentions during group conversations and social events [13, 36, 46], for which socially intelligent systems need to be capable of interpreting human non-verbal social behaviour, even revealed through ambiguous, highly contextual, or subtle cues [29]. This challenge is addressed by Social Signal Processing (SSP) [8] research.

The Human Oriented Machine Intelligence (HOMI) group, to which I belong, is dedicated to advancing the field of SSP by developing sophisticated machine perception systems that can accurately interpret and model human social behavior [25, 35, 36, 41]. HOMI proposes three key components of conversational dynamics in social interactions: modelling participants' intention [25], group involvement [41], and conversational events [45].

ICMI '24, November 04–08, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3688614

2 Research direction

Within the social intention modelling task, our research addresses the challenge of recognizing and modeling plausible human narratives about the intentions of individuals involved in group interactions [13, 25, 35]. The goal is to develop a deep learning model capable of identifying social intentions [27, 40, 48], including those unrealized, based on multimodal cues at a fine-grained temporal scale, without requiring knowledge of the outcomes of these intentions. This model should (1) be grounded in existing theories of intention [6, 33, 37], (2) reflect the specificity of human observations of social behavior [11, 17, 30] in both the ability to infer intentions from subtle cues [29], and taking subjective perspective while disambiguating these cues [14]. Moreover, we aim at (3) producing associated human-readable narratives about these intentions [10] and (4) making the model generalizable to novel social groups and contexts.

2.1 Research question

Following this research direction, we can formulate the overarching research question as follows:

How can multi-modal sensory data and subjective narratives be integrated to create a generative model that accurately reflects human social intentions and adapts to diverse observers and contexts?

The crucial sub-questions that need to be addressed over the course of this research are:

(1) What formalized knowledge structure can be developed to inform the annotation process and to which extent it reflects intuitive background knowledge of humans? Can social intentions be organized hierarchically in a form of taxonomy?
(2) In what ways can human observers' intuitive abilities to infer intentions from nuanced social cues be leveraged to enhance the accuracy and diversity of intention predictions?
   (a) Can existing models account for observer's subjectivity with respect to these intentions in a consistent and explainable way
   (b) Which evaluation approaches can be used to compare human and generated narratives about intentions, maximizing the score for plausible, realistic, perspective-taking, consistent ones
(3) In producing the associated narrative explanations, how can we bridge modality gap between generative language model and multi-modal input that includes sensory data, narrative embeddings and metadata about annotators?

(4) What methodologies can be employed to ensure adaptability to new categories of observers?

2.2 Motivation

The motivation for this research operates on several levels, each addressing a critical need in the realm of social interaction and technology. This work aims to model the subjectivity of social perception — a complex and nuanced aspect of human behavior that remains a significant challenge in computational approaches [11, 17, 32].

By developing generative models that integrate multimodal sensory data [26] and subjective narratives [4], this research seeks to overcome the limitations of other approaches that often conflate intentions with outcomes [21, 40]. While discriminative models excel at classifying and predicting based on given input-output pairs, they often fall short in scenarios where understanding, creativity, and context-sensitive interpretation are crucial — qualities that are particularly important in modeling social settings [5]. Social interactions are often ambiguous, with multiple potential interpretations for a given action or statement [29]. By considering past events, environmental factors, and the flow of conversation, generative models can offer more subtle and nuanced interpretations, provided in format that is comprehensible and interpretable for wider audience.

Potential downstream applications include aiding cross-cultural interactions [39], supporting the integration of marginalized social groups [28], and moderating and assisting in professional meetings [19]. To address cross-cultural interactions, the model could be utilized for analyzing students' behavior in schools with higher ethnic and cultural diversity [20], making use of perspective-taking to reveal discrepancies in interpretations of behaviors across this variety. To facilitate societal integration, the model could be deployed to ease the communicative interactions of neurodivergent people in real time, in a form of a digital personal assistant [2].

3 Background and related work

3.1 Theories of intention

Intention, as a concept, has been explored across various disciplines, each providing a unique perspective on its formation and role in human behavior [6, 15, 44]. Here, we delve into key theories that have shaped our understanding of intention.

According to the belief-desire theory [6, 37], intentions are formed through a combination of desires and beliefs, culminating in a choice that includes a commitment to action. The BDI model has been foundational in fields such as artificial intelligence and cognitive science, where it serves as a framework for developing autonomous agents that simulate human decision-making processes [33]. Developing on the original theory, Bratman argues that intentions are formed as part of plans that an agent commits to for future conduct [6]. These plans serve a crucial role in coordinating actions over time [7]. Intention, in this view, involves a belief in the feasibility of the plan and a commitment to executing it.

3.2 Large Language Models in Social Psychology Research

Recent advancements in Large Language Models (LLMs)[49] have significantly impacted social psychology research, providing new methodologies for analyzing and understanding human behavior and social interactions [50]. LLMs have been utilized to explore a variety of social psychology topics, including sentiment analysis [18], discourse analysis [12], and the modeling of intentions [27]. For instance, studies have used LLMs to analyze social media data [31], uncovering trends in public opinion and emotional responses to events. These models can process vast amounts of textual data, enabling researchers to extract patterns and insights about social behaviors at a scale previously unattainable.

3.3 Large Language Models with Multi-Modal Inputs

Recent advancements in vision-language pre-training (VLP) have provided several solutions to effectively combine LLMs with features extracted from computer vision models. One foundational model in this area is **CLIP**, developed by OpenAI [34]. CLIP bridges the gap between vision and language by learning a joint embedding space for images and text, enabling zero-shot inference.

ALIGN (A Large-scale Image-Language model) [22] scales up contrastive learning to hundreds of millions of image-text pairs, achieving state-of-the-art performance on various vision-language tasks. By using a simple dual-encoder architecture, ALIGN learns robust and generalized representations across different domains, significantly improving the zero-shot performance on downstream tasks.

One of the primary challenges in using a frozen LLM is aligning visual features with the text space. To address this, Tsimpoukelli et al. [43] proposed **Frozen**, which fine-tunes an image encoder whose outputs are directly used as soft prompts for the LLM. This method allows the visual features to be interpreted within the text-generative framework of the LLM. Alternatively, **Flamingo** by Alayrac et al. [1] introduces new cross-attention layers into the LLM to inject visual features. These new layers are pre-trained on billions of image-text pairs, ensuring a robust integration of visual data into the language model.

Q-Former [24] is another innovative solution, acting as a lightweight transformer with a set of learnable query vectors designed to extract the most useful visual features from a frozen image encoder. It functions as an "information bottleneck" [24], feeding the essential visual data to the LLM while filtering out irrelevant information. This approach reduces the complexity of vision-language alignment learning by feeding only the most useful information to the LLM. It shows zero-shot learning capacity, potentially facilitating the framing of tasks as meta-learning.

4 Methodology

Following our research direction, we design a formalized taxonomy [10, 47] of conversational intentions to guide the adaptation of our labeling framework [45] to compile a dataset of social intentions related to behaviors in complex conversational scenes.

The model will integrate high-level features extracted from various sensory inputs [26, 42] such as video, movement data, and

low-frequency audio, forming a unified representation of the objective data. To incorporate subjective perspectives, annotator narratives will be converted into embeddings using pre-trained language models, and combined with metadata about the annotators.

We will employ a meta-learning paradigm [35] to train a generative model that learns from both objective sensory data and subjective narratives. The model will be fine-tuned to generate plausible narratives for different settings and perspectives, ensuring adaptability to new environments and diverse points of view.

4.1 Operational Framework

4.1.1 Plausible narratives. In this research, we define a plausible narrative as a descriptive account that verbally articulates how the components of a social intention manifest in observable actions, influenced by the individual's or group's context.

4.1.2 Intention taxonomy. Our taxonomy of intentions is stratified by different conversational stages, and inherently multi-dimensional, considering factors such as function, modality, means of conveying, spatial zones, and scope. On a high level, it consists of:

- Pre-initiating Intentions involve joining the conversation either as an active participant or as an affiliate of the conversation [23], setting the stage for engagement.
- Initiating Intentions include greeting and social rituals to initiate interaction according to social norms.
- Maintaining Intentions, such as showing agreement or support to continue interaction.
- Modifying Intentions that alter the course of the conversation: suggesting a topic shift or ending the current topic.
- Ending Intentions include farewells and social rituals to leave, disengagement signals indicating a desire to end the conversation.
- Post-ending Intentions involve actions taken after the conversation has formally ended.

4.2 Modelling approach

Our modelling approach leverages both bottom-up and top-down paradigms to effectively capture and represent social intentions. By integrating low-level cues with high-order events or actions, we condition data-driven modeling on high-order descriptions produced by annotators, referred to as "plausible narratives." This approach introduces the subjectivity of observers into the model, ensuring that the resulting narratives are reflective of human perspectives.

4.2.1 Annotation Collection. We will collect annotations through two distinct experimental modes to capture both intuitive and structured perspectives on social intentions.

In the first experiment, annotators will be asked to describe the intentions of participants without any detailed instructions about the types of intentions or their typical cues. This approach relies on the annotators' social intuition and natural interpretive skills to provide spontaneous and unstructured descriptions of observed behaviors.

In the second experiment, annotators will be prompted with our taxonomy of intentions and instructed to identify and categorize the types of intentions described. This structured approach will leverage the detailed framework provided by the taxonomy to guide annotators in recognizing and labeling specific intentions, ensuring consistency and alignment with the defined categories.

Comparative analysis of the narratives generated in both modes will provide insights into the validity of our structured taxonomy in guiding human observers into more insightful understanding of social intentions

4.3 Model Design

4.3.1 Limitations of Zero-Shot LLM Inference. While zero-shot inference with LLMs performs well in uninformed tasks [49], it is suboptimal for our goals [50]. These models lack access to temporal dynamics and sensory inputs, which are crucial for understanding social interactions. Additionally, LLMs have a limited ability to incorporate the background and subjective perspective of annotators. Although prompt engineering can help [38], it doesn't address these core issues.

4.3.2 Proposed Multi-Modal Generative Model. Our proposed model leverages diverse data modalities, including video, movement data (accelerometer and proximity), and low-frequency audio (intonation and loudness) while preserving privacy by rendering words indistinguishable. Additionally, metadata about the events during which this data was collected is used. High-level NLP features such as frequency of topic changes, relative vectors of topic change, and sentiment analysis, which do not disclose specific topics and words, further enrich the "objective" data.

The "subjective" input comprises metadata about the observers, capturing the specification of their point of view (POV) [16]. This subjective input also includes raw descriptions of the conversational scene, which are agnostic to sensory context and annotator details but may be prompted with example narratives about intentions. These descriptions are converted into embeddings using LLMs and concatenated with the observer metadata.

An important aspect of our model is the consideration of involvement features, which are part of the subjective component. These features will be defined more concretely at a later stage, since involvement bears a subjective perception of its degree.

4.3.3 Model Training Approach. The model will be trained in a meta-learning paradigm to produce subjective narratives for various settings (changes in "objective" data) and different points of view (changes in "subjective" data). The training process is outlined as follows:

Feature extraction model: The first step involves training an "objective model" for feature extraction from all multi-modal data. This model requires intention labels established by informed and qualified annotators. The model is trained with a classification head to extract salient features relevant to social intentions. Once the model is trained, the classification head is removed at the transfer learning stage, retaining the pre-trained feature weights from the hidden layers.

Integrating subjectivity and narratives: The next step involves enriching the model with subjective data and narratives. This integration combines three vectors: pre-trained feature weights from the objective model, metadata about the annotators, and embeddings of the raw descriptions prompted with example narratives.

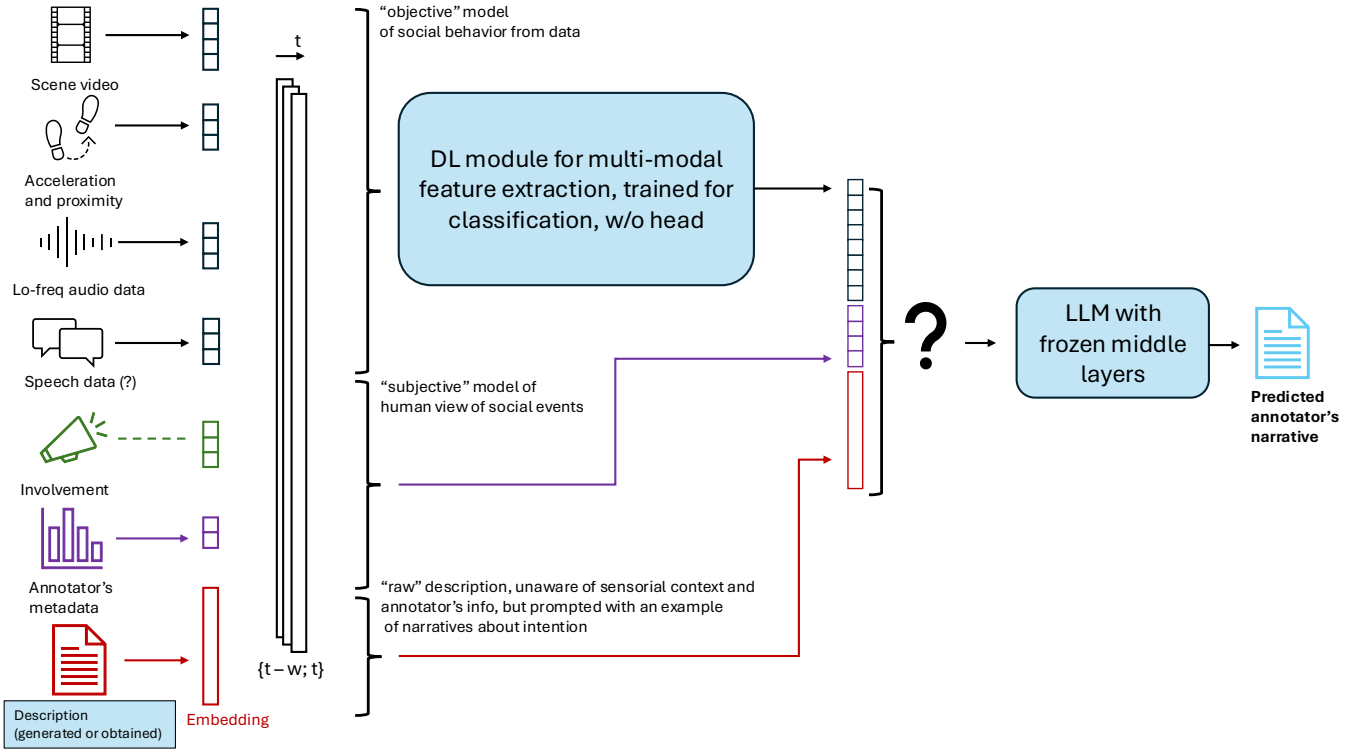

Figure 1: Proposed model architecture with "?" indicating the modality gap that we aim to overcome with Q-Former architecture

These combined vectors are jointly input into the subsequent component of the model, effectively merging objective and subjective data.

Input to LLM: The combined data is then fed into an LLM configured with frozen middle layers. The top layers of the LLM remain unfrozen to allow fine-tuning, enabling the model to adapt to new input and generate task-specific outputs. This final stage predicts the subjective narrative based on the received annotator metadata.

5 Remaining Work

Along with the implementation of the feature extraction model, the key remaining research gap is bridging the modality gap and enhance the integration of multi-modal data in our generative model, we propose adapting the Q-former architecture, enhancing it with multi-modal integration.

6 Research Contributions

This research contributes to the field of SSP and human-machine interaction by advancing our understanding of social intentions through multi-modal data integration and generative modeling. The theoretical framework being developed in this study will offer a novel and comprehensive taxonomy of conversational intentions, providing a unified description that benefits social science research. Our multi-modal generative model represents a sophisticated approach to automatically estimate social intentions.

Acknowledgments

This paper was funded by ERC Consolidator. Grant NEON:1010891

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. https://doi.org/10.48550/arXiv.2204.14198 arXiv:2204.14198 [cs].

[2] Nuria Aresti-Bartolome and Begonya Garcia-Zapirain. 2014. Technologies as Support Tools for Persons with Autistic Spectrum Disorder: A Systematic Review. *International Journal of Environmental Research and Public Health* 11, 8 (Aug. 2014), 7767–7802. https://doi.org/10.3390/ijerph110807767 Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[3] Elliot Aronson, Timothy D. Wilson, and Sam Sommers. 2019. *Social psychology* (tenth edition ed.). Pearson, New York, NY. OCLC: 1007494785.

[4] Lisa Feldman Barrett. 2017. *How Emotions Are Made: The Secret Life of the Brain.* Houghton Mifflin Harcourt, Boston.

[5] Ben M. Bensaou, Charles Galunic, and Claudia Jonczyk-Sédès. 2014. Players and Purists: Networking Strategies and Agency of Service Professionals. *Organization Science* 25, 1 (2014), 29–56. https://www.jstor.org/stable/43660866 Publisher: INFORMS.

[6] Michael E. Bratman. 1987. *Intention, Plans, and Practical Reason.* Cambridge, MA: Harvard University Press, Cambridge.

[7] Michael E. Bratman. 1993. Shared Intention. *Ethics* 104, 1 (1993), 97–113. https://www.jstor.org/stable/2381695 Publisher: The University of Chicago Press.

[8] Judee K. Burgoon, Nadia Magnenat-Thalmann, Maja Pantic, and Alessandro Vinciarelli (Eds.). 2017. *Social Signal Processing.* Cambridge University Press, Cambridge. https://doi.org/10.1017/9781316676202

[9] Federico Cabitza, Andrea Campagner, and Martina Mattioli. 2022. The unbearable (technical) unreliability of automated facial emotion recognition. *Big Data & Society* 9, 2 (July 2022), 20539517221129549. https://doi.org/10.1177/

20539517221129549 Publisher: SAGE Publications Ltd.
[10] Jean Decety and Philip Jackson. 2004. The Functional Architecture of Human Empathy. *Behavioral and cognitive neuroscience reviews* 3 (June 2004), 71–100. https://doi.org/10.1177/1534582304267187
[11] Yayue Deng, Jinlong Xue, Fengping Wang, Yingming Gao, and Ya Li. 2023. CMCU-CSS: Enhancing Naturalness via Commonsense-based Multi-modal Context Understanding in Conversational Speech Synthesis. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. Association for Computing Machinery, New York, NY, USA, 6081–6089. https://doi.org/10.1145/3581783.3612565
[12] Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023. Understanding Social Reasoning in Language Models with Language Models. https://doi.org/10.48550/arXiv.2306.15448 arXiv:2306.15448 [cs].
[13] Ekin Gedik and Hayley Hung. 2018. Detecting Conversing Groups Using Social Dynamics from Wearable Acceleration: Group Size Awareness. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4 (Dec. 2018), 163:1–163:24. https://doi.org/10.1145/3287041
[14] Maria Gendron, Debi Roberson, Jacoba Marietta Van Der Vyver, and Lisa Feldman Barrett. 2014. Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion* 14, 2 (2014), 251–262. https://doi.org/10.1037/a0036052
[15] Margaret P. Gilbert. 2007. Searle on Collective Intentions. https://papers.ssrn.com/abstract=3523055
[16] Antonio Liz Gutiérrez and Margarita Vázquez Campos. 2015. The Notion of Point of View. In *Temporal Points of View: Subjective and Objective Aspects*, Margarita Vázquez Campos (Ed.). Springer.
[17] Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller. 2017. From Hard to Soft: Towards more Human-like Emotion Recognition by Modelling the Perception Uncertainty. In *Proceedings of the 25th ACM international conference on Multimedia (MM '17)*. Association for Computing Machinery, New York, NY, USA, 890–897. https://doi.org/10.1145/3123266.3123383
[18] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, Seattle WA USA, 1122–1131. https://doi.org/10.1145/3394171.3413678
[19] Jess Hohenstein, Rene F. Kizilcec, Dominic DiFranzo, Zhila Aghajari, Hannah Mieczkowski, Karen Levy, Mor Naaman, Jeffrey Hancock, and Malte F. Jung. 2023. Artificial intelligence in communication impacts language and social relationships. *Scientific Reports* 13, 1 (April 2023), 5487. https://doi.org/10.1038/s41598-023-30938-9 Publisher: Nature Publishing Group.
[20] Maria João Hortas. 2008. Territories of integration: the children of immigrants in the schools of the Metropolitan Area of Lisbon. *Intercultural Education* 19, 5 (Oct. 2008), 421–433. https://doi.org/10.1080/14675980802531630 Publisher: Routledge _eprint: https://doi.org/10.1080/14675980802531630.
[21] Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. 2015. Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology* 6 (July 2015), 1049. https://doi.org/10.3389/fpsyg.2015.01049
[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. https://doi.org/10.48550/arXiv.2102.05918 arXiv:2102.05918 [cs].
[23] Adam Kendon. 1990. *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge University Press.
[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. https://doi.org/10.48550/arXiv.2301.12597 arXiv:2301.12597 [cs].
[25] Litian Li, Jord Molhoek, and Jing Zhou. 2024. Inferring Intentions to Speak Using Accelerometer Data In-the-Wild. http://arxiv.org/abs/2401.05849 arXiv:2401.05849 [cs].
[26] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2023. Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. https://doi.org/10.48550/arXiv.2209.03430 arXiv:2209.03430 [cs].
[27] Adyasha Maharana, Quan Tran, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, and Mohit Bansal. 2022. Multimodal Intent Discovery from Livestream Videos. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 476–489. https://doi.org/10.18653/v1/2022.findings-naacl.36
[28] Adam S. Miner, Nigam Shah, Kim D. Bullock, Bruce A. Arnow, Jeremy Bailenson, and Jeff Hancock. 2019. Key Considerations for Incorporating Conversational AI in Psychotherapy. *Frontiers in Psychiatry* 10 (Oct. 2019). https://doi.org/10.3389/fpsyt.2019.00746 Publisher: Frontiers.
[29] Monica M. Moore. 1985. Nonverbal courtship patterns in women: Context and consequences. *Ethology and Sociobiology* 6, 4 (Jan. 1985), 237–247. https://doi.org/10.1016/0162-3095(85)90016-0
[30] Nora A. Murphy and Judith A. Hall. 2021. Capturing Behavior in Small Doses: A Review of Comparative Research in Evaluating Thin Slices for Behavioral Measurement. *Frontiers in Psychology* 12 (April 2021), 667326. https://doi.org/10.3389/fpsyg.2021.667326
[31] Heinrich Peters and Sandra Matz. 2024. Large Language Models Can Infer Psychological Dispositions of Social Media Users. https://doi.org/10.48550/arXiv.2309.08631 arXiv:2309.08631 [cs].
[32] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (Sept. 2017), 98–125. https://doi.org/10.1016/j.inffus.2017.02.003
[33] Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. Machine Theory of Mind. https://doi.org/10.48550/arXiv.1802.07740 arXiv:1802.07740 [cs].
[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. https://doi.org/10.48550/arXiv.2103.00020 arXiv:2103.00020 [cs].
[35] Chirag Raman, Hayley Hung, and Marco Loog. 2022. Social Processes: Self-Supervised Meta-Learning over Conversational Groups for Forecasting Nonverbal Social Cues. https://doi.org/10.48550/arXiv.2107.13576 arXiv:2107.13576 [cs].
[36] Chirag Raman, Jose Vargas Quiros, Stephanie Tan, Ashraful Islam, Ekin Gedik, and Hayley Hung. 2022. ConfLab: A Data Collection Concept, Dataset, and Benchmark for Machine Analysis of Free-Standing Social Interactions in the Wild. https://openreview.net/forum?id=CNJQKM5cV2o
[37] Anand Rao and Michael Georgeff. 2001. Modeling Rational Agents within a BDI-Architecture. (Jan. 2001).
[38] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. https://doi.org/10.48550/arXiv.2402.07927 arXiv:2402.07927 [cs].
[39] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2023. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence* 5, 1 (Jan. 2023), 46–57. https://doi.org/10.1038/s42256-022-00593-2 Publisher: Nature Publishing Group.
[40] Gopendra Vikram Singh, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. EmoInt-Trans: A Multimodal Transformer for Identifying Emotions and Intents in Social Conversations. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 31 (Nov. 2022), 290–300. https://doi.org/10.1109/TASLP.2022.3224287
[41] Stephanie Tan, David M.J. Tax, and Hayley Hung. 2022. Conversation Group Detection With Spatio-Temporal Context. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. ACM, Bengaluru India, 170–180. https://doi.org/10.1145/3536221.3556611
[42] Sara Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2020. Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE transactions on affective computing* 11, 2 (2020), 200–213. https://doi.org/10.1109/TAFFC.2017.2784832
[43] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal Few-Shot Learning with Frozen Language Models. https://doi.org/10.48550/arXiv.2106.13884 arXiv:2106.13884 [cs].
[44] Raimo Tuomela and Kaarlo Miller. 1988. We-Intentions. *Philosophical Studies* 53, 3 (1988), 367–389. https://doi.org/10.1007/bf00353512 Publisher: Springer.
[45] Jose Vargas-Quiros, Laura Cabrera-Quiros, Catharine Oertel, and Hayley Hung. 2022. Impact of annotation modality on label quality and model performance in the automatic assessment of laughter in-the-wild. https://doi.org/10.48550/arXiv.2211.00794 arXiv:2211.00794 [cs, eess].
[46] Jose Vargas-Quiros, Chirag Raman, Stephanie Tan, Ekin Gedik, Laura Cabrera-Quiros, and Hayley Hung. [n. d.]. REWIND Dataset: Privacy-preserving Speaking Status Segmentation from Multimodal Body Movement Signals in the Wild. *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING* ([n. d.]).
[47] Anuradha Welivita and Pearl Pu. 2020. A Taxonomy of Empathetic Response Intents in Human Social Conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 4886–4899. https://doi.org/10.18653/v1/2020.coling-main.429
[48] Anuradha Welivita, Yubo Xie, and Pearl Pu. 2020. Fine-grained Emotion and Intent Learning in Movie Dialogues. https://doi.org/10.48550/arXiv.2012.13624 arXiv:2012.13624 [cs].
[49] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. https://doi.org/10.48550/arXiv.2303.18223 arXiv:2303.18223 [cs].
[50] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? https://doi.org/10.48550/arXiv.2305.03514 arXiv:2305.03514 [cs].