SBM

Social Behavior Model for Human-Like Action Generation

Chew, Jouh Yeong; Lin, Zhi Yi; Zhang, Xucong

DOI
10.1145/3747327.3763038

Publication date
2025

Document Version
Final published version

Published in
ICMI 2025 - Companion Publication of the 27th International Conference on Multimodal Interaction

SBM: Social Behavior Model for Human-Like Action Generation

Jouh Yeong Chew
Honda Research Institute Japan
Saitama, Japan
jouhyeong.chew@jp.honda-ri.com

Zhi-Yi Lin
Computer Vision Lab
TU Delft
Delft, Netherlands
z.y.lin@tudelft.nl

Xucong Zhang
Computer Vision Lab
TU Delft
Delft, Netherlands
xucong.zhang@tudelft.nl

Abstract

Humans use verbal and nonverbal cues for effective communication, particularly during group interactions. Enabling intelligent systems — such as robots and virtual agents — to understand and generate such cues is crucial to facilitate natural and trustworthy human-robot interactions. We propose Social Behavior Model (SBM), a novel framework to generate socially appropriate actions in multi-party scenarios. Specifically, SBM takes into account the contextual information from surrounding individuals and the history of interaction data to generate socially coherent actions for an intelligent agent, including dialogue content and nonverbal cues like pose. To adapt pre-trained LLMs to the domain of social behavior, we fine-tune them using the Low-Rank Adaptation (LoRA) technique on a newly curated, labeled dataset containing multiparty social cues such as text and pose data. This method preserves the base model's capabilities while enabling domain-specific adaptation with minimal computational cost. Given the lack of prior work on multi-party social behavior generation, we benchmark our model against state-of-the-art methods in dyadic pose generation. Our results demonstrate superior performance, establishing SBM as the first foundation model that integrates multiparty verbal and nonverbal social cues generation grounded in context understanding.

CCS Concepts

• **Computing methodologies** → **Learning from demonstrations**; • **Human-centered computing**;

Keywords

Pose generation, nonverbal cues, multi-party interaction, multimodal interaction, social context, large language model

ACM Reference Format:
Jouh Yeong Chew, Zhi-Yi Lin, and Xucong Zhang. 2025. SBM: Social Behavior Model for Human-Like Action Generation. In *Companion Proceedings of the 27th International Conference on Multimodal Interaction (ICMI Companion '25), October 13–17, 2025, Canberra, ACT, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3747327.3763038

1 Introduction

Understanding the dynamics of multiparty social interaction is a complex challenge, yet it is essential for intelligent systems — such as robots and virtual agents.This challenge aligns with the broader goals of human-centered computing, which aims to steer the development of AI systems toward improving human well-being. Within this framework, AI agents must be capable of engaging with users in intuitive and human-like ways. A promising direction in this pursuit is the development of human-like AI agents to mimic human-human interactions [8, 21, 24]. Such agents can uniquely capture and hold a user's attention, especially in the early phases of interaction, often more effectively than text or speech alone [1]. This capability has particular relevance for applications in healthcare—e.g., to support individuals with depression, autism spectrum disorders, or social anxiety [9, 11]—as well as to enhance user familiarity and trust, which are critical to fostering long-term engagement [3, 14]. To facilitate acceptance by humans, AI systems must be able to perceive and interpret human behavior, and respond with human-like sensitivity and social fluency.

Existing AI agents can perform a wide range of tasks, including listening [10], planning and executing robot motions [13, 16, 23], performing collaborative lifting [17], greeting people in social environments [22], and even engaging in physical gestures such as hugging [2]. However, there remains a significant gap: no existing method serves as a foundation model capable of understanding human behavior and generating rich, multi-modal responses that integrate multiple behavioral features in complex, real-world social settings. This is challenging as human social behavior includes subtle and coordinated verbal and nonverbal cues—including speech, body language, and fine-grained movements that reveal underlying intent. Prior work has addressed this challenge using text prompt engineering and heuristic rules [20], temporal graph modeling [10], and denoising diffusion techniques [15]. However, many of these solutions are not end-to-end and rely on external modules, which limits their robustness and scalability. While recent diffusion-based methods offer more integrated solutions, they focus on dyadic interactions and do not generalize to group settings.

The emergence of generative AI—particularly Large Language Models (LLMs) and Vision-Language Models (VLMs)—has accelerated research in human-AI interaction. These models have demonstrated impressive capabilities for generating natural verbal interactions [12, 18], however, they fall short in scenarios involving embodied agents like robots. This shortcoming arises because LLMs and VLMs are not inherently designed to drive physical behavior or embodiment, which is critical for natural and engaging human-robot interactions [19]. LLM generally interprets the extensive knowledge that can empower the behavior model to be able to work with unseen objects and also can serve as the backbone of the model [6]. With recent rapid developments, variants of LLMs can accept not only text but also image, video, speech, etc. [4, 5, 25]. However, it is still not explored yet on how to integrate temporal human behavior into LLMs due to the strong causality, which is there are spatial correlations between behavior features and temporal correlations across time. For example, the hand gesture of a person should obtain: i) semantically align with body, facial, and speech,

ICMI Companion '25, Canberra, ACT, Australia

ACM ISBN 979-8-4007-2076-5/25/10
https://doi.org/10.1145/3747327.3763038

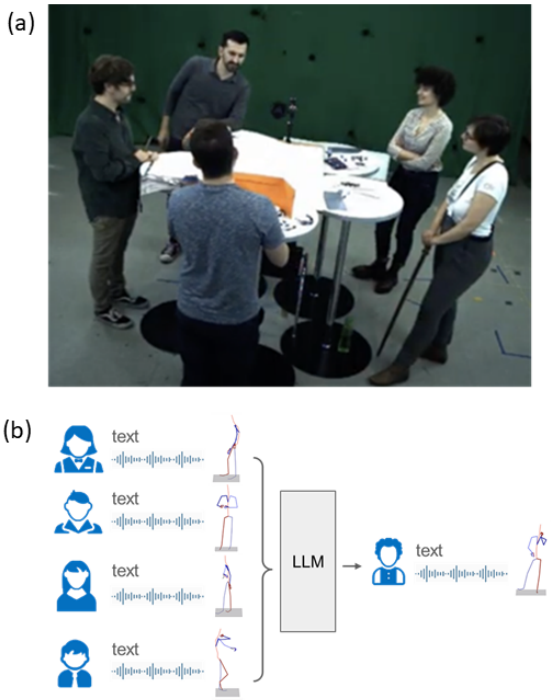

Figure 1: We show (a) a frame grabbed from the DnD dataset [15] to demonstrate the problem formulation of action generation during multiparty social interaction, and (b) the overview of the SBM . Given the multi-modal observations from multiple people and historical information, the SBM outputs speech, text and body motion for a selected person, who can be replaced by an intelligent agent or robot.

ii) movements are continuous, and iii) be similar but not the same as a deja vu situation according to personality.

We propose SBM , a novel foundation model for generating socially appropriate actions in group interactions involving multiple people. Our model leverages the learned semantics of a pre-trained LLM to generate realistic body poses in response to social context. We fine-tune the LLM in an auto-regressive manner, incorporating interaction history to generate temporally coherent actions. Unlike previous methods, SBM employs lightweight adapters trained on a compact, labeled dataset of human-human interactions—including aligned text, pose, and audio—to model the correlation and synchronization across multiple social modalities. This design enables SBM to explicitly capture multi-modal dependencies among participants in a scene, ensuring that the generated verbal and nonverbal behaviors are contextually appropriate, temporally aligned, and socially coherent. Our key contributions are as follows:

- SBM to generate nonverbal cues considering the social context from multiple people, and
- the representation of social context, which we validate using experiments with an open-source dataset.

2 Social Behavior Model (SBM)

The proposed SBM framework to generate socially appropriate actions for a selected target person during multiparty interactions is illustrated in fig. 1. The core task is to generate verbal and nonverbal actions—including body pose, hand pose, and speech transcript—for a single target person, based on the context of the surrounding people and the recent history of the interaction. We assume that the visual and auditory features—such as body pose, hand pose, and speech transcripts—are already extracted from input videos using state-of-the-art methods. These features serve as the input to SBM , which focuses solely on action generation. The model is designed with the goal to enable a social robot to replace one of the humans in a group interaction by generating socially coherent actions.

2.1 Social Context Encoding

We explore three variants of SBM in fig. 2 to evaluate how different representations of social context influence action generation quality. We aim to systematically investigate how the granularity and format of contextual information affect the model's performance in generating socially aligned multimodal behaviors.

Pose-Only Context (Variant A).. The simplest version of our model uses only the past sequences of body and hand poses from all participants as input. This variant predicts the target person's future pose based solely on these low-level physical features. While this representation is lightweight, it lacks semantic information, which often leads to outputs that are less socially appropriate or temporally aligned with the underlying group dynamics.

Pose + Transcript Context (Variant B).. The second variant extends the input by including the speech transcripts of all participants in the interaction. This provides semantic context in addition to physical pose data. The model is trained to generate not only the target person's body and hand poses, but also their expected speech transcript. While the addition of semantic content helps align physical and verbal behaviors, this approach introduces a large volume of raw text that can overwhelm the model and reduce interpretability due to noise and redundancy in human conversations.

Pose + Summarized Social Context (Variant C).. The most advanced variant uses high-level summaries of the group interaction as input. These summaries are generated using a ChatGPT model applied to the raw transcripts. The summary captures both coarse and fine-grained context, such as overall group tone (e.g., friendly discussion), speaker-target relationships (e.g., who is addressing whom), and nonverbal cues (e.g., a participant folding their arms). This holistic representation offers a balance between semantic richness and input tractability, enabling the model to generate coherent and socially compliant responses. The output remains the same: body pose, hand pose, and speech transcript of the target person.

2.2 Model Architecture

The backbone of SBM is an LLM, chosen for its powerful semantic understanding and flexible token-based input/output architecture. While LLMs are inherently designed to process textual data, we extend their capabilities to handle structured physical information such as body and hand poses by converting these signals into tokenized forms. To achieve this, we design a pose encoder that transforms raw joint angle vectors into discrete tokens that are compatible with the LLM input format. Each token encodes a specific configuration of body or hand joints, allowing the LLM to reason about physical dynamics alongside textual semantics. The model is trained to auto-regressively predict future actions, including motion and speech, for a given target individual based on recent history

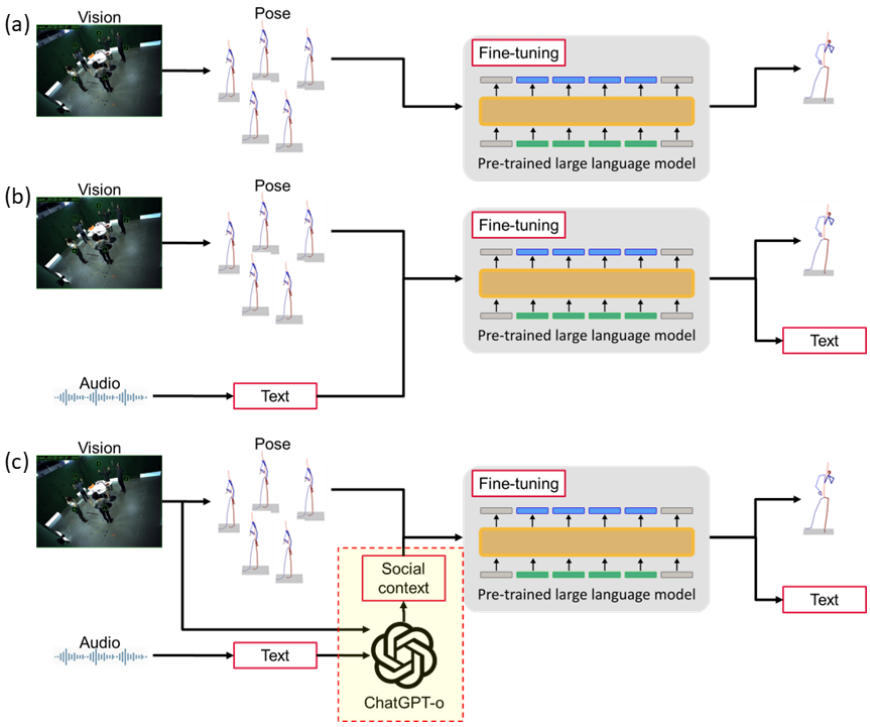

Figure 2: The experiments to evaluate the SBM framework using different representation of social context using (a) Variant A: only the pose of multiple people, (b) Variant B: the pose and text of multiple people, and (c) Variant C: the pose and social context of multiple people generated by ChatGPT-o. The backbone of our model is a pre-trained LLM model.

and contextual input. During decoding, the output tokens corresponding to motion are fed into a pose decoder that reconstructs continuous joint angles for both body and hands. Simultaneously, text tokens are directly interpreted as the speech.

Backbone and Adaptation Strategy. We use LLaMA-3 [6] as the LLM backbone, which offers state-of-the-art performance in reasoning and text generation. However, adapting such a large pre-trained model to a new domain (i.e., social interaction modeling) is computationally expensive. To address this, we employ the Low-Rank Adaptation (LoRA) technique [7], which enables efficient fine-tuning of large models by injecting small, trainable low-rank matrices into the frozen weights of the pre-trained network.

Low-Rank Adaptation (LoRA).. LoRA fine-tunes large pre-trained models by injecting trainable low-rank matrices into the frozen weight matrices. Instead of updating the full weight matrix $W \in \mathbb{R}^{d\times k}$, LoRA decomposes the update into two smaller matrices:

$$\Delta W = AB, \quad \text{where } A \in \mathbb{R}^{d\times r},\ B \in \mathbb{R}^{r\times k},\ r \ll \min(d,k) \tag{1}$$

The adapted weight becomes:

$$W' = W + \Delta W = W + AB \tag{2}$$

This significantly reduces the number of trainable parameters during fine-tuning, while retaining the general capabilities of the pre-trained model. In our case, LoRA is applied to selected projection layers of the LLaMA 3 model to specialize it for social interaction modeling. We train these adapters using synchronized annotations of body pose, hand pose, and speech for each participant. The adapters allow SBM to specialize in modeling the intricate dynamics of social behavior while maintaining the general-purpose reasoning strength of the original LLaMA-3 model.

3 Experiments

We use the DnD Group Gesture Dataset [15], which provides a rich dataset of multi-person interactions. We follow the protocol defined in the paper , using an 80%/10%/10% split for training/validation/testing to ensure comparability using:

- **L1 Divergence (L1Div)**: Computes the ability of models to span the space of gesture motions with enough coverage. This represents the diversity of the pose. L1Div is computed for both the predicted joints of a human skeleton and the Ground Truth (GT) joints in a given dataset, and the difference is smaller the better.
- **Mean Per Joint Position Error (MPJPE)**: Measures the average distance between the predicted joints of a human skeleton and the ground truth. MPJPE is smaller the better.

3.1 Pre-processing

We segment the continuous motion and transcript data into fixed-length clips. Each clip consists of 64 consecutive frames, corresponding to a few seconds of interaction. This chunking allows the model to learn from temporal windows while capturing meaningful interaction context. Only segments with utterances are used to train model to generate synchronized verbal and nonverbal response.

3.2 Comparison with the Benchmark

We first evaluate the performance of SBM against the state-of-the-art baseline ConvoFusion [15]. Table 1 shows the results of the comparison across three SBM variants. L1Div quantifies the distributional difference between ground-truth (GT) and predicted (Pred) joint motion. A smaller absolute difference (GT - Pred) indicates a better match between the diversity of generated and human motion.

The preliminary results are promising as they demonstrate improvement of SBM over the ConvoFusion in terms of motion diversity. While ConvoFusion shows a large divergence gap of 3.93 between ground-truth and predicted motion distributions, all variants of SBM maintain a smaller gap below 0.05, highlighting the ability of our model to more accurately replicate the natural variability in human motion. Among the SBM variants, we observe a consistent trend as more lower level context such as pose and raw speech transcripts are introduced into the input, the divergence gap decreases. The pose-only variant already performs well, indicating that our approach can model reasonable motion distributions even from low-level signals. However, adding raw speech transcripts improves the alignment of motion with the underlying interaction dynamics, reducing the gap further to 0.0171.

The performance of the SBM deteriorates when the higher level semantic context which incorporates the description of the social scene generated by ChatGPT-o is added to the pose. Although the results of L1Div gap (0.0474) is still better than the ConvoFusion benchmark, it is noteworthy the semantically rich context could make it difficult for the model to capture the nuanced relationships between the different modalities and people in social scenes. ConvoFusion exhibits a much higher divergence due to its reliance on heuristic and modular processing pipelines, which can limit its capacity to model the full range of human-like gestures in multiparty interactions. Our end-to-end framework—grounded in LLM-based semantic modeling and LoRA-based fine-tuning—offers a more coherent integration of motion and social context. These findings suggest the effectiveness of SBM as a foundation model for social behavior generation and highlight the importance of incorporating structured, context-aware inputs in multimodal interaction.

Method	L1Div		
	GT	Pred	(GT-Pred) ↓
ConvoFusion	5.1200	1.1900	3.9300
SBM (pose)	0.5369	0.5175	0.0194
SBM (pose+transcript)	**0.5369**	**0.5199**	**0.0171**
SBM (pose+context)	0.5369	0.4895	0.0474

Table 1: Comparison of L1Div performance metrics for body motion prediction between ConvoFusion [15] and SBM . The last column shows the difference between the ground-truth (GT) and the prediction (Pred), the number is lower the better.

3.3 Pose Tokenizer Ablation

table 2 shows the ablation study comparing model performance with and without the pose tokenizer, which refers to an encoder that converts continuous joint angle values into discrete tokens, enabling more effective integration with the LLM input format. The results suggest removing the pose tokenizer leads to better performance. Models without the tokenizer achieve lower MPJPE and L1Div, indicating more accurate and diverse motion generation. These findings suggest directly feeding normalized joint angles into the model could be more effective than discretizing them through a tokenizer.The added quantization introduced by the tokenizer may result in information loss, limiting the model's ability to generate fine-grained, human-like motion. While pose tokenization offers a convenient mechanism for text-based models, our results show that retaining continuous representations yields superior performance in both accuracy and motion diversity in our context.

4 Discussions

The preliminary results of SBM outperforms the ConvoFusion baseline across all metrics which demonstrates SBM is promising to produce more human-like and varied gestures that closely align with the natural distribution found in the dataset, thus suggesting the effectiveness of our end-to-end generative approach. L1Div reflects how closely the distribution of generated motions matches that of real human behavior, making it suitable for assessing diversity and realism—critical attributes in social behavior modeling. MPJPE quantifies per-joint accuracy and serves as a precise measure of motion quality at the frame level. Together, these metrics evaluate diversity and spatial accuracy.

Models trained without pose tokenizer perform significantly better than those with it. This suggests that direct modeling of continuous joint angles could be more effective than converting them into discrete tokens for LLM input. We hypothesize that the tokenization process introduces quantization errors, which hinder the model ability to generate nuanced and fine-grained motion patterns. While tokenization is beneficial in cases with strong symbolic structure (e.g., language), in the case of continuous human motion, preserving precision appears more important.

Interestingly, the low level representation of social scene using speech transcript and pose yields the best performance over the semantically rich context of descriptions generated by ChatGPT-o. This suggests the semantic information provided by high-level context—such as group mood or who is addressing whom — offers limited benefit beyond what is already captured by raw motion and transcript inputs. There could be over-abstraction of the social scene which is too general for the SBM to learn the nuanced relationships within the social interaction. There are several potential reasons for this. First, the summaries are generated via a LLM and may lack detailed temporal alignment with the actual interaction flow. Second, the model may already infer sufficient semantic structure directly from the sequence of past poses and speech, diminishing the marginal value of abstracted context. Lastly, current LLMs may not be fully optimized to leverage structured group-level summaries, especially if such inputs are loosely formatted.

Method	Tokenizer	(GT-Pred) ↓	(MPJPE) ↓
SBM (pose)	0	0.0194	0.1682
	1	0.0792	0.4702
SBM (pose+transcript)	**0**	**0.0171**	**0.1161**
	1	0.0792	0.4681
SBM (pose+context)	0	0.0474	0.1899
	1	0.0796	0.4685

Table 2: Ablation tests of SBM with and without pose tokenizer in the second column.

5 Conclusion

Incorporating low level social context such as pose key points and speech transcription (i.e., pose+transcript), leads to improved social behavior generation. High level abstraction of social scene is challenging for the SBM to learn. However, this limitation likely depends on the quality and granularity of the context summary, which warrants further investigation. Our model is limited to scenarios where the target person is actively speaking. Extending the model to predict speaking turns and transitions between speakers remains an important direction for future work.

References

[1] Gary Bente, Sabine Rüggenberg, Nicole C Krämer, and Felix Eschenburg. 2008. Avatar-mediated networking: Increasing social presence and interpersonal trust in net-based collaborations. *Human communication research* 34, 2 (2008), 287–318.

[2] Alexis E Block, Hasti Seifi, Otmar Hilliges, Roger Gassert, and Katherine J Kuchenbecker. 2023. In the arms of a robot: Designing autonomous hugging robots with intra-hug gestures. *ACM Transactions on Human-Robot Interaction* 12, 2 (2023), 1–49.

[3] Linnda R Caporael. 1986. Anthropomorphism and mechanomorphism: Two faces of the human machine. *Computers in human behavior* 2, 3 (1986), 215–234.

[4] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 958–979.

[5] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15180–15190.

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

[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

[8] Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. 2020. Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[9] Michelle R Kandalaft, Nyaz Didehbani, Daniel C Krawczyk, Tandra T Allen, and Sandra B Chapman. 2013. Virtual reality social cognition training for young adults with high-functioning autism. *Journal of autism and developmental disorders* 43 (2013), 34–44.

[10] J. Taery Kim, Archit Naik, Isuru Jayarathne, Sehoon Ha, and Jouh Yeong Chew. 2024. Modeling social interaction dynamics using temporal graph networks. In *33rd IEEE International Conference on Robot & Human Interactive Communication (RO-MAN 2024)*. arXiv:2404.06611 [cs.HC] https://arxiv.org/abs/2404.06611

[11] Julian Leff, Geoffrey Williams, Mark A Huckvale, Maurice Arbuthnot, and Alex P Leff. 2013. Computer-assisted therapy for medication-resistant auditory hallucinations: proof-of-concept study. *The British Journal of Psychiatry* 202, 6 (2013), 428–433.

[12] Jan Leusmann, Chao Wang, and Sven Mayer. 2024. Comparing Rule-based and LLM based Methods to enable Active Robot Assistant Conversations. In *Workshop@CHI 2024: Building Trust in CUIs – From Design to Deployment*. Workshop position paper of CHI 2024.

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

[14] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & automation magazine* 19, 2 (2012), 98–100.

[15] Muhammad Hamza Mughal, Rishabh Dabral, Ikhsanul Habibie, Lucia Donatelli, Marc Habermann, and Christian Theobalt. 2024. ConvoFusion: Multi-Modal Conversational Diffusion for Co-Speech Gesture Synthesis. arXiv:2403.17936 [cs.CV] https://arxiv.org/abs/2403.17936

[16] Eley Ng, Ziang Liu, and Monroe Kennedy. 2023. It takes two: Learning to plan for human-robot cooperative carrying. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 7526–7532.

[17] Lorenzo Rapetti, Carlotta Sartore, Mohamed Elobaid, Yeshasvi Tirupachuri, Francesco Draicchio, Tomohiro Kawakami, Takahide Yoshiike, and Daniele Pucci. 2023. A control approach for human-robot ergonomic payload lifting. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 7504–7510.

[18] Petri Tikka, Andrea Alesani, Vladimir Goriachev, Jaakko Karjalainen, and Kaj Helin. 2024. LLM based virtual assistant for human-robot interaction. https://fcai.fi/ai-day-2024 FCAI AI Day + Nordic AI Meet 2024 ; Conference date: 21-10-2024 Through 22-10-2024.

[19] Joshua Wainer, David J. Feil-seifer, Dylan A. Shell, and Maja J. Mataric. 2006. The role of physical embodiment in human-robot interaction. In *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*. 117–122. doi:10.1109/ROMAN.2006.314404

[20] Chao Wang, Stephan Hasler, Daniel Tanneberg, Felix Ocker, Frank Joublin, Antonello Ceravola, Joerg Deigmoeller, and Michael Gienger. 2024. LaMI: Large Language Models for Multi-Modal Human-Robot Interaction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, 1–10. doi:10.1145/3613905.3651029

[21] Jieyeon Woo, Catherine Pelachaud, and Catherine Achard. 2023. Asap: Endowing adaptation capability to agent in human-agent interaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 464–475.

[22] Yang Xue, Fan Wang, Hao Tian, Min Zhao, Jiangyong Li, Haiqing Pan, and Yueqiang Dong. 2021. Proactive interaction framework for intelligent social receptionist robots. In *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 3403–3409.

[23] Yang Yang, Xibai Lou, and Changhyun Choi. 2022. Interactive robotic grasping with attribute-guided disambiguation. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 8914–8920.

[24] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. Few-shot adversarial learning of realistic neural talking head models. 9459–9468.

[25] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*. PMLR, 2165–2183.