A Social-interaction World Model Pipeline: Multimodal Data Acquisition for Capturing Bifurcating Social Intents in HRI

1st Yutaka Nakamura
Guardian Robot Project, RI-H
Riken
Kyoto, Japan
yutaka.nakamura@riken.jp

2nd Chenfei Xu
Guardian Robot Project, RI-H
Riken
Kyoto, Japan
chenfei.xu@riken.jp

3rd Yuma Kabe
School of Informatics and Engineering
The University of Electro-Communications
Tokyo, Japan
yuma.kabe@uec.ac.jp

4th Kaoruko Shinkawa
Grad. School of Informatics and Engineering
The University of Electro-Communications
Tokyo, Japan
kshinkawa@uec.ac.jp

5th Yuya Okadome
Faculty of Engineering
Tokyo University of Science
Tokyo, Japan
okadome@rs.tus.ac.jp

6th Yoshihiro Nakata
Grad. School of Informatics and Engineering
The University of Electro-Communications
Tokyo, Japan
ynakata@uec.ac.jp

***Abstract*—Human-shared spaces are fundamentally non-stationary, characterized by "bifurcating social intents" where future actions diverge into multiple potential paths. We propose a multimodal pipeline to capture these dynamics by integrating ego-centric video, acoustic maps, and gaze saliency into a Social-interaction World Model. By integrating ego-centric video, acoustic maps, and gaze saliency, we develop a Social-interaction World Model that maintains interaction narratives. Our framework combines spontaneous human-to-human data with teleoperation to bridge the gap between signal-level dynamics and high-level reasoning. This establishes a foundation for grounding future VLA models in socially-legible, risk-sensitive decision-making.**

***Index Terms*—Human robot interaction, world model, multi-modal dataset**

I. Introduction

In recent years, Vision-Language-Action (VLA) models have emerged as a robust foundation for robotic control [9] [10] [11]. Advances in generalist robot policies have enabled highly precise execution of mobile manipulation tasks in diverse environments [12]. However, for robots to coexist and operate seamlessly within human-shared spaces [13] [18] [19], it is essential to develop foundation models that incorporate "social behaviors," understanding and adhering to the implicit social norms and cues of human interaction.

Unlike traditional manipulation tasks, interacting with humans takes place in a non-stationary environment [4]. Because humans operate based on internal intentions and adapt to evolving contexts [14], their future actions are inherently multimodal and highly uncertain. For instance, in a hallway navigation scenario, humans dynamically choose to pass on the right or left based on the other person's trajectory. Furthermore, the robot's own behavior serves as a critical context that influences subsequent human actions, creating a reciprocal loop of interaction. From an HRI perspective, it is crucial that the robot's motions are not only functional but also interpretable to humans [7], often referred to as legible or predictable movement.

In this paper, we introduce our ongoing efforts to develop a decision-making model grounded in social interaction, aiming to bridge the gap between physical task execution and social intelligence [1] [2].

Specifically, we have developed a data acquisition system where human subjects are equipped with fisheye cameras and microphone arrays (acoustic cameras) to capture a comprehensive view of the surrounding environment and its spatial audio context [2]. Furthermore, we collect multimodal datasets that include eye-tracking data and movement trajectories via IMU (Inertial Measurement Unit) [1]. In addition to human-side data, we have integrated a similar sensor suite into a humanoid robot [15] [16]. This setup allows us to capture rich, ego-centric data across multiple sensor modalities during real-time human-robot interaction, providing a dual perspective on social behavior and action.

The acquired multimodal datasets are processed using feature extraction techniques such as acoustic spatial mapping and semantic segmentation. Leveraging these features, we are currently developing a generative model that simultaneously predicts future video sequences and robot actions [1] [3]. This approach essentially functions as a "Social-interaction world model," simulating how a robot's interventions influence human behavior and vice-versa. Furthermore, by co-generating future saliency maps to mimic human gaze patterns, we investigate how the robot's generated behaviors affect human

This work was supported by the Japan Science and Technology Agency (JST) Moonshot Research and Development Grant through the Development of Semi-Autonomous CA under Grant JPMJMS2011.

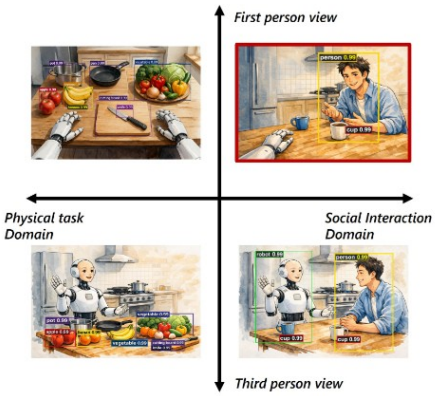

Fig. 1. Conceptual research map for Social-interaction World Model. Note: The images and the overlaid detection boxes/scores are AI-generated for illustrative purposes and do not represent actual experimental results or system outputs.

perception and social impression.

II. Modeling Social Dynamics through Generative World Models

In this study, we propose a Social-interaction World Model that positions the robot not as a mere observer of the environment but as an active participant in social interaction. Conventional VLAs typically focus on physical tasks, seeking optimal solutions within a deterministic action space based on absolute coordinates; in such contexts, any sensor configuration is theoretically applicable, provided it meets resolution requirements. In contrast, social interaction lacks a single 'correct' coordinate or solution, as it is defined by a continuous loop of mutual influence. To capture these social dynamics, we model the system from an ego-centric perspective, mirroring that of a human participant. The following sections describe the specific characteristics of the social interaction space addressed in this research.

A. Bifurcating World-lines in HRI

Interacting with humans occurs in a fundamentally non-stationary environment [17]. As humans act based on internal intent, the future state is not a single deterministic path but a series of bifurcations. For instance, the choice to pass on the right or left creates divergent "world-lines." The robot must not only predict these branches but also understand how its own presence triggers these splits in human behavior [1].

B. Hand-crafted to Pixel-based

Modeling these interactions solely through hand-crafted features, such as coordinate trajectories, is insufficient due to the sheer diversity of real-world scenarios. The number of agents and the relevance of environmental cues (e.g., a person carrying heavy luggage) vary dynamically, making it impossible to pre-define a complete state space. Therefore, we adopt a pixel-based generative approach that captures the full richness of the scene. To mitigate the data-hungry nature of raw video learning, we incorporate mid-level abstractions, such as semantic segmentation, to balance representational capacity with learning efficiency [6] [1].

C. Maintaining Narrative through Social-interaction World Models

Traditional one-step-forward predictive models often fail to capture the "narrative" or long-term context of an interaction [1]. By utilizing video generation as a Social-interaction World Model, we generate extended temporal sequences that maintain consistency across an entire encounter. This allows the model to reason about the multimodal evolution of a social situation rather than just instantaneous reactions.

D. Challenges in Teaching Social Behaviors

In social contexts, conventional kinesthetic teaching is impractical. Because the interaction partner is a dynamic agent, the 'correct' social motion depends entirely on real-time coordination. To capture this fluid "social dance," we employ a dual-pronged data acquisition strategy that combines the observation of spontaneous human-to-human interactions via wearable sensor suites with teleoperated data collection [3], effectively mapping human social intent onto the robot's embodiment. While these methods involve operational and perceptual challenges, they remain the most viable path to extracting the tacit, embodied social intelligence required for authentic human-robot coordination.

E. Reinforcement Learning for Embodied Adaptation

Finally, we must account for the difference in embodiment between humans and robots. A motion that is natural for a human may appear uncanny or be physically infeasible for a robot [16]. We have to employ reinforcement learning to refine the behaviors derived from our world model [5] [6]. By considering the entropy of predicted human responses, the system performs risk-sensitive decision-making, ensuring that the robot's actions are both physically stable and socially legible.

F. Related work

While several large-scale egocentric datasets exist, they differ significantly from our objectives. For instance, Ego4D [21] covers a vast range of daily activities but does not focus on social communication, lacking modalities such as acoustic maps which are crucial for interpreting social intent. Other datasets like Ego-Exo4D [24], EPIC-Kitchens [22] and EgoExoLearn [23] emphasize task execution and object manipulation rather than social behavior in shared spaces. Although EgoCom [25] specifically addresses communication, it is limited to static, pre-defined settings and does not account for the situational variety encountered by a mobile robot. In contrast, our dataset is uniquely designed to model dynamic human-robot interactions, including dialogue, while navigating everyday environments. Although our current sample size is smaller than benchmarks like Ego4D, our pipeline provides a critical multimodal foundation for capturing the nuances of Social-interaction world models.

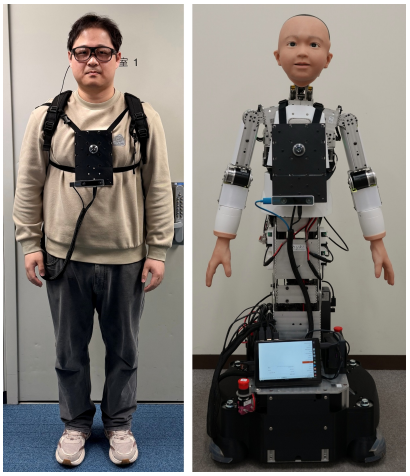

Fig. 2. EgoSAS on a human and an andrdoid Yui

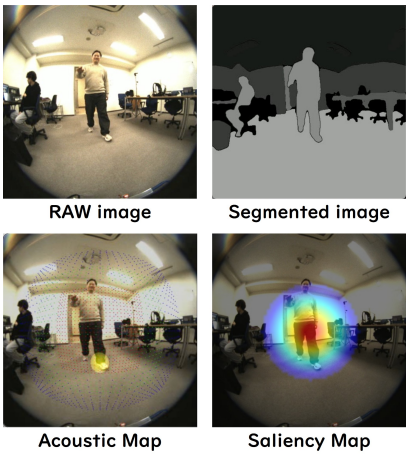

Fig. 3. Observed information

III. Data acquisition system and model

Figure 2 shows the sensor system implemented as a wearable system. The system consists of a fisheye camera, a 16-ch microphone array, and an IMU sensor (Intel Realsense T265) [1]. The microphone array generates acoustic maps [2], while the IMU sensor records the movement of the robot. Furthermore, in teleoperation [3] and wearable systems, saliency maps are obtained by using eye-tracking devices such as Tobii Pro Glasses. All of this information, except for the movement, is then mapped onto the images captured by the fisheye camera as shown in Figure 6

Our model adopts a generative architecture for video, where multi-modal features are structured as a temporal sequence of high-dimensional 'frames.' Unlike conventional RGB videos, our approach constructs a composite frame $x(t)$ by concatenating the FPV image $o_{img}(t)$, acoustic map $o_{sound}(t)$, and saliency map $a_{gaze}(t)$ along the channel dimension (Figure 4) The movement vector $a_{move}(t)$ (e.g., the translation parameter and the rotating speed) is treated as a separate input stream associated with each frame: $x(t) = [o_{img}(t), o_{sound}(t), a_{gaze}(t), a_{move}(t)]$. By integrating these observations and actions into a unified multi-channel format, the model can process the entire interaction history as a single spatio-temporal sequence: $X(t) = [x(t), x(t-1), \ldots, x(t-T)]$, where T is the time window for each clip (Figure 4).

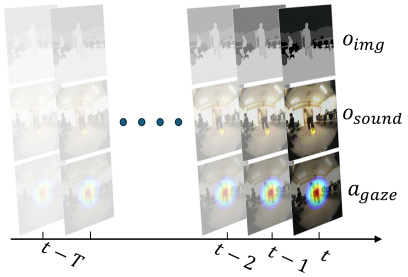

Fig. 4. Data format for the spatio-temporal generative model

We implemented a generative model for this data format using a Diffusion Model [20]. Since Diffusion Models can reconstruct missing information from available data, our model can generate both future observations and actions based on a history of arbitrary length. Utilizing this generative capability, the model performs both forecasting (inference of future observations) and action planning (inference of future actions, i.e., imitation learning).

This approach is particularly valuable in real-world scenarios involving human interaction. In such environments, "multi-modal" prediction is essential to account for the diverse potential futures influenced by factors such as human intent. Furthermore, unlike simple one-step-forward predictions, our method can generate interpretable action candidates that represent coherent, goal-oriented behaviors over a sustained duration.

IV. A preliminary tele-operation experiment using an Android

As a Proof of Concept (PoC) for the proposed system, we developed a model to control the behaviors of a robot in a reception-style interaction [5] [6] [7] [8]. Figure 5 illustrates the data collection environment. The robot is positioned in a room facing a corridor and engages in dialogue with a person who enters from the corridor to approach it. The objective is to construct a model that appropriately controls the robot's gaze, gestures, and relative positioning in such scenarios. For simplicity, the robot's locomotive functions were not utilized in this experiment.

The model was trained on an integrated dataset combining the data collected in this scenario with egocentric data from other sources and data collected with different robot platforms [1] [3]. Figure 6 displays the results of the model generating appropriate gaze directions as saliency maps for the data recorded in this environment. While the robot's gaze is not always perfectly directed at the relevant person, partly due to data scarcity, the model is expected to realize appropriate gaze behavior that considers the overall surroundings.

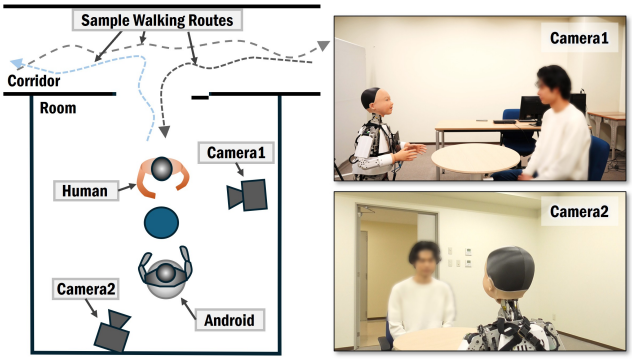

Fig. 5. Data recording environment

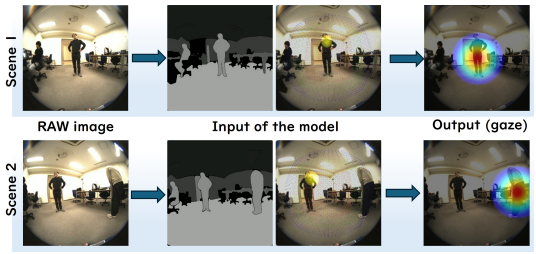

Fig. 6. Observed information

V. Conclusion and Discussion

In this paper, we introduced a novel framework for modeling social intelligence within unstructured, human-centric environments. By leveraging first-person (ego-centric) multimodal data, including video, acoustic spatial maps, and gaze saliency, we have developed a Social-interaction World Model that moves beyond simple physical manipulation. Our approach uniquely combines natural human-to-human interaction data with teleoperated robot data, providing a rich foundation for the robot to understand the implicit social norms and dynamic cues of shared spaces.

Regarding the safety and stability of generated actions, our goal is not to achieve deterministic perfection but to explicitly handle the inherent uncertainty of social interactions. It is crucial to acknowledge that humans and robots possess different embodiments; therefore, the robot will not necessarily replicate human behavior, nor will humans respond in a uniform manner. Rather than pursuing a single 'correct' path, our focus is on minimizing the 'surprise' or cognitive load experienced by humans when a robot's actions deviate from their expectations [1].

Building on our previous findings in reinforcement learning for social interactions [8] [6], we hypothesize that the integration of our Social-interaction World Model into an RL loop will enable the robot to navigate more complex, bifurcating social scenarios that require a deeper understanding of human intent. By considering the entropy of predicted human behaviors and the potential bifurcations of intent [1], for example, the system might make risk-sensitive decisions. Through this model, we aim to realize a robot that can operate in non-stationary environments by actively reasoning about its social predictions and maintaining a consistent narrative, rather than simply reacting to sensory signals.

As a next step, we are currently deploying this model onto a physical humanoid platform to validate its real-time performance. A key focus of our upcoming experiments is the integration of high-level linguistic contexts with low-level sensorimotor control. By leveraging the VLM-based automated annotation pipeline we have established, we aim to bridge the gap between abstract semantic instructions and signal-level social dynamics. This will allow the robot to not only execute physically stable movements but also to reason about the long-term narrative of an interaction, such as understanding the social implications of a specific gesture or a verbal cue.

Furthermore, we will evaluate how these entropy-aware, language-conditioned policies influence human perception in the wild. Specifically, we will investigate whether grounding a Vision-Language-Action (VLA) model in our Social-interaction World Model leads to more legible and socially acceptable behaviors. Ultimately, we seek to foster a truly co-existent relationship between humans and autonomous agents through a unified understanding of both physical and social environments.

Acknowledgment

We sincerely thank Huthaifa Ahmad, Liliana Villamar Gomez, Zhichao Chen, Haru Nakamura, Li Xuhai, Franco Antonio, and Rio Taguchi for their invaluable technical support and contributions to the development of the hardware, teleoperation systems, and experimental platforms used in this study. Additionally, Gemini was utilized as an AI-powered tool for language translation and stylistic refinement to improve the clarity of the paper.

Fig. 1 was generated with the assistance of the image generation feature in ChatGPT (OpenAI) and was subsequently refined and edited by the authors to ensure its conceptual accuracy. The final version has been reviewed and approved by all authors.

References

[1] C. Xu, H. Ahmad, Y. Okadome, H. Ishiguro, Y. Nakamura, "Action-inclusive multi-future prediction using a generative model in human-related scenes for mobile robots," IEEE Access, 13, 2025.

[2] Z. Chen, C. Xu, H. Ahmad, Y. Okadome, H. Ishiguro, Y. Nakamura, "A feasibility study with in-the-wild data in human interaction settings: Acoustic-visual fusion for predictive sound source positioning," IEEE Access, 13, 2025.

[3] C. Xu, Z. Chen, Y. Okadome, H. Ishiguro, Y. Nakamura, "A Egocentric Situational Awareness Sensor (EgoSAS): Collecting Human-Human Interaction Data for Mobile Robots' Co-existence," IEICE Tech. Rep., CQ2025-18, 2025.

[4] Y.Okadome, Y. Alkatshah, Y. Nakamura, "Generating interaction gestures in dyadic conversations using a diffusion model," PLoS One, 20 (12), e0339579, 2025.

[5] Z. Chen, Y. Nakamura, H. Ishiguro, "Android as a Receptionist in a Shopping Mall Using Inverse Reinforcement Learning," IEEE Robotics and Automation Letters, 7 (3), 7091 – 7098, 2023.

[6] Z. Chen, Y. Nakamura, H. Ishiguro, "Outperformance of Mall-Receptionist Android as Inverse Reinforcement Learning is Transitioned to Reinforcement Learning," IEEE Robotics and Automation Letters, 8 (6), 3350 – 3357, 2023.

[7] A. H. Qureshi, Y. Nakamura, Y. Yoshikawa, H. Ishiguro, "Show, attend and interact: Perceivable human-robot social interaction through neural attention Q-network," in 2017 IEEE International Conference on Robotics and Automation (ICRA), 1639 -– 1645, 2017.

[8] A. H. Qureshi, Y. Nakamura, Y. Yoshikawa, H. Ishiguro, "Intrinsically motivated reinforcement learning for human–robot interaction in the real-world," Neural Networks, 107, 23–33, 2018.

[9] A. Brohan, et. al., "RT-1: Robotics Transformer for Real-World Control at Scale," in Proceedings of Robotics: Science and Systems (RSS), 2023.

[10] A. Brohan, et. al., "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control," arXiv preprint arXiv:2307.15818, 2023.

[11] C. Chi, et. al., "Universal Manipulation Interface: In-The-Wild Robot Teaching Without In-The-Wild Robots," in Proceedings of Robotics: Science and Systems (RSS), 2024.

[12] Generalist AI Team, "GEN-0: Embodied Foundation Models That Scale with Physical Interaction," Generalist AI Blog, 2025

[13] T. Minato, K. Sakai, T. Uchida, H. Ishiguro, "A study of interactive robot architecture through the practical implementation of conversational android," Frontiers in robotics and AI, 2022.

[14] T. Uchida, T. Minato, Y. Nakamura, Y. Yoshikawa, H. Ishiguro, "Female-Type Android's Drive to Quickly Understand a User's Concept of Preferences Stimulates Dialogue Satisfaction: Dialogue Strategies for Modeling User's Concept of Preferences," International Journal of Social Robotics, 13, 1499 – 1516, 2021.

[15] M. Nakajima, K. Shinkawa, Y. Nakata, "Development of the Lifelike Head Unit for a Humanoid Cybernetic Avatar 'Yui' and its Operation Interface," IEEE Access, 12, 23930 – 23942, 2024.

[16] R. Taguchi, K. Shinkawa, Y. Nakata, "Personal Space Toward Human-like and Non-human-like Robots: Effects of Robot Appearance and Likability," in 2026 IEEE/SICE International Symposium on System Integration (SII), 483 – 490, 2026.

[17] Y. Nishimura, Y. Nakamura, H. Ishiguro, "Human interaction behavior modeling using Generative Adversarial Networks," Neural networks, 132, 521 – 531, 2020.

[18] Y. Nakata, et. al., "Development of 'ibuki' an electrically actuated childlike android with mobility and its potential in the future society," Robotica, 40 (4), 933 – 950, 2021.

[19] H. Ishiguro, et. al., "Cybernetic Avatar," Singapore: Springer, 2025

[20] H. Lu, et. al., "VDT: General-purpose Video Diffusion Transformers via Mask Modeling," International Conference on Learning Representations, 2023.

[21] K. Grauman, et. al., "Ego4D: Around the World in 3, 000 Hours of Egocentric Video," https://arxiv.org/abs/2110.07058, 2021.

[22] D. Damen, et. al., "The EPIC-KITCHENS Dataset: Collection, Challenges and Baselines," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020

[23] Y. Huang, et. al., "EgoExoLearn: A Dataset for Bridging Asynchronous Ego- and Exo-centric View of Procedural Activities in Real World," https://arxiv.org/abs/2403.16182, 2025

[24] K. Grauman, et. al., "Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives," https://arxiv.org/abs/2311.18259, 2024

[25] C. G. Northcutt, et. al., "EgoCom: A Multi-Person Multi-Modal Egocentric Communications Dataset," IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(6), pp. 6783–6793, 2023