Enhancing End-User Engagement in Human-Robot Interaction by Performing LLM-driven Expressive Behaviors

SHIPENG LYU, Department of Mechanical Engineering, The Hong Kong Polytechnic University (PolyU), Hong Kong; Ningbo Institute of Digital Twin, Eastern Institute of Technology, China
FANGYUAN WANG, Department of Mechanical Engineering, The Hong Kong Polytechnic University (PolyU), Hong Kong; Ningbo Institute of Digital Twin, Eastern Institute of Technology, China
WEIWEI LIN, Ningbo Institute of Digital Twin, Eastern Institute of Technology, China
GUODONG GUO, Ningbo Institute of Digital Twin, Eastern Institute of Technology, China
DAVID NAVARRO-ALARCON, Department of Mechanical Engineering, The Hong Kong Polytechnic University (PolyU), Hong Kong

We introduce a novel framework with strong generalization capabilities for enabling humanoid robots to perform semantically grounded, context-aware, and physically realizable expressive behaviors, with the goal of enhancing end-user engagement in open-world human–robot interaction (HRI). To achieve this goal, we develop a new LLM-based human cognition module that interprets user dialogues to infer latent intent and generates high-level multimodal behavior descriptions. These semantic representations are mapped to speech and motion outputs through a dual-stage embodied behavior generation pipeline. The pipeline consists of a shape adaptation module that maps human body motions into the robot's kinematic space, followed by a motion retargeting module that generates executable joint trajectories under physical constraints. Additionally, the modular architecture enables seamless integration of state-of-the-art generative models and serves as a practical testbed for evaluating expressive behavior generation in real-world settings. We validate this system on a 58-DoF humanoid platform through both controlled video-based studies and live HRI experiments. The results show significant improvements over rule-based and handcrafted baselines in terms of expressiveness, behavioral appeal, and user engagement. This work helps bridge the gap between high-level language understanding and low-level robot control, thereby enabling scalable and human-aligned expressive behavior generation for embodied agents.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; • **Computing methodologies → Robotic planning**.

Additional Key Words and Phrases: Human-robot interaction, Robotic behavior generation, Large language models, Humanoid robots

Authors' Contact Information: Shipeng Lyu, shipeng.lyu@connect.polyu.hk, Department of Mechanical Engineering, The Hong Kong Polytechnic University (PolyU), Kowloon, Hong Kong; and Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, Zhejiang, China; Fangyuan Wang, fangyuan.wang@connect.polyu.hk, Department of Mechanical Engineering, The Hong Kong Polytechnic University (PolyU), Kowloon, Hong Kong; and Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, Zhejiang, China; Weiwei Lin, wlin@idt.eitech.edu.cn, Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, Zhejiang, China; Guodong Guo, gdguo@eitech.edu.cn, Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, Zhejiang, China; David Navarro-Alarcon, dnavar@polyu.edu.hk, Department of Mechanical Engineering, The Hong Kong Polytechnic University (PolyU), Kowloon, Hong Kong.

ACM 1557-735X/2018/8-ART111
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Shipeng Lyu, Fangyuan Wang, Weiwei Lin, Guodong Guo, and David Navarro-Alarcon. 2018. Enhancing End-User Engagement in Human-Robot Interaction by Performing LLM-driven Expressive Behaviors. *J. ACM* 37, 4, Article 111 (August 2018), 32 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Humanoid robots play a crucial role in human-robot interaction (HRI), particularly in the context of interactions involving people with complex and individualized needs, such as social companionship for elders [29]. For instance, the humanoid robot Sil-Bot has been widely adopted in studies aimed at enhancing cognitive engagement among elderly populations [25]. In these studies, humanoid robots serve as emotional partners to influence users' affective and cognitive states through social companionship and interaction, thereby contributing to improved user's engagement. However, a significant limitation of current humanoid robots in HRI lies in the rigidity and monotony of their movements. In other words, their actions are largely predefined and limited in variety. These predefined constrained behaviors restrict their ability to interact with end-users in the expected manner, especially in uncertain scenarios. Therefore, these constraints reduce their effectiveness in HRI, i.e., low user's engagement during interaction [6]. To solve this problem, this paper focuses on enhancing end-users' engagement during interaction by leveraging adaptive (not predefined) and expressive humanoid robot behaviors, as illustrated in Fig. 1.

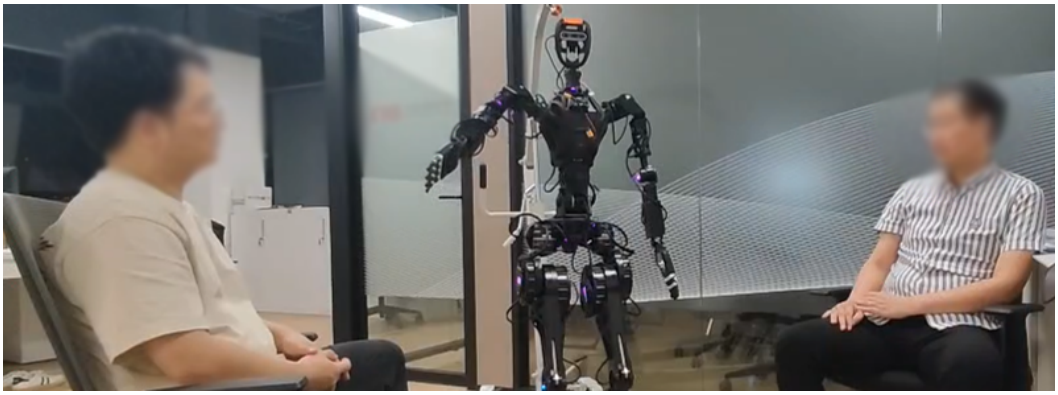

Fig. 1. A HRI scenario for humanoid robot. The life-sized humanoid robot enhances end-user engagement during HRI through expressive behaviors, including voice and action.

To achieve our objective, it is necessary to improve the behavioral expressiveness of humanoid robots [35, 36]. Research has indicated that end-users often engage in functional interactions with robots [1], such as verbal communication and physical action, which present a significant challenge for robots to exhibit expressive behaviors. This challenge arises from the necessity for humanoid robots to exhibit behaviors consistent with human norms in contextually similar scenarios, thereby aligning with the expectations of end-users. However, their behaviors, such as actions, speech, vocal characteristics, and facial expressions, for many recently popular humanoids rely on predefined skills often diverge significantly from natural human behavior, resulting in robotic responses that appear rigid and unnatural. These constrained behaviors hinder the robots' ability to interact with users in a human-expected manner, which in turn reduces their effectiveness in HRI [6].

Given that most commercially available life-sized humanoid platforms, such as GR1, Figure 01, and H1, lack controllable facial features (e.g., controllable eyes or mouth), their ability to convey

expressive behaviors remains significantly constrained. Moreover, these two modalities (speech and action) remain common and powerful channels for demonstrating expressive robot behavior to support our study target [35]. Therefore, this paper focuses on enhancing end-user engagement in HRI by driving humanoids to perform expressive behaviors, which mainly involve body actions and verbal communication.

Although we have simplified expressive feedback behaviors for humanoids, achieving expressive anthropomorphic behaviors remains technically complex. First, due to the morphological differences between humans and robots, a critical challenge lies in transferring expressive behaviors learned from humans onto physical humanoid platforms while minimizing the loss of expressiveness during deployment [4]. Second, ensuring that the performed behaviors align with the expectations of individual users remains a complex and unresolved problem in personalized HRI. The difficulty arises from the fact that end-user expectations are subjective and difficult to measure. For example, unlike traditional robotics tasks, where evaluation metrics are often objective (e.g., precision or accuracy), the assessment of expressive behavior may be entertaining or perhaps frightening, depending on its alignment with their expectations [35]. Therefore, understanding the intentions and expectations of end-users in HRI tasks, along with their subjective preferences for robot feedback, is crucial for robots to obtain a positive evaluation from humans.

To this end, we propose a novel and comprehensive framework (depicted in Fig. 2) to generate expressive behaviors for life-sized humanoid robots. Our approach introduces a human cognition module powered by large language models (LLMs), which captures user intent and preferences to generate semantically rich, context-aware descriptions of expressive speech and action sequences. These descriptive representations are then translated into multi-modal behaviors via two behavior-mapping channels: a speech generation pipeline [28] and a motion synthesis pipeline [26]. In detail, we first generate human motion sequences consistent with the action descriptions using a text-to-motion model in the motion synthesis pipeline. Considering the morphological differences between humans and robots, as well as hardware constraints, directly mapping the generated human motions to the robot often yields poor results. To ensure the generated behaviors can be faithfully reproduced on the physical robot, we propose a shape adaptation method and a motion retargeting mechanism. Based on our proposed framework, we aim to enhance user engagement by enabling humanoid robots to perform expressive, human-like verbal and action behaviors that are contextually appropriate and aligned with user expectations during interaction.

However, our provided method raises several fundamental questions about the role of LLM-driven expressive behaviors in HRI. While augmenting speech with non-verbal motion may increase behavioral richness, it remains unclear whether such automatically generated expressions preserve communicative clarity. Beyond basic understanding, an open question is whether LLM-driven expressive behaviors meaningfully enhance perceived expressiveness and behavioral appeal, and whether their quality can approach that of carefully designed expert behaviors. Moreover, improvements observed in controlled or video-based settings do not necessarily translate to real-world interaction. Therefore, it remains to be determined whether expressive motion increases user engagement during live, real-time interaction, and whether semantically aligned, generative behaviors provide benefits beyond fixed, template-based motion feedback. These questions collectively guide the design of our system and motivate the empirical studies presented in this paper. We summarize our contributions as follows:

(1) We propose a novel HRI framework that integrates LLMs with multimodal behavior generation, enabling humanoid robots to produce semantically grounded and socially appropriate expressive behaviors based on natural language interaction.

(2) We design a dual-stage embodied pipeline comprising shape adaptation and motion retargeting, which ensures faithful and physically feasible execution of generated expressive behaviors on life-sized humanoid robots.
(3) We implement a modular and extensible architecture that bridges high-level semantic reasoning with low-level robot control, providing an executable platform for deploying and evaluating diverse behavior generation models on real-world robots.
(4) We validate our approach through both video-based and real-time interaction studies, demonstrating significant improvements in behavioral expressiveness and user engagement over baseline.

The rest of the article is structured as follows. In Section 2, we review related work on expressive humanoid behaviors and engagement in HRI. Section 3 presents the proposed LLM-driven behavior generation framework, including the reasoning module and the behavior generation pipeline. In Section 4, we report study 1, a controlled video-based evaluation of robot behavioral expressiveness and appeal across feedback modalities. Section 5 introduces study 2, a live interaction study that measures user engagement under different robot feedback strategies. Section 6 evaluates the applicability of our framework to different generative motion models. In Section 7, we discuss the findings and synthesize insights across both studies. Finally, Section 8 concludes the paper and outlines limitations and future directions.

2 Related Work

Despite the myriad concerns expressed towards humanoid robots, quantitative research findings indicate that they are generally acceptable to users [6]. Specifically, presenting expressive behavior by robots is a crucial approach to mitigating these concerns among end-users. The concept and principles of expressive behavior were initially formalized by animators [31], which inspired early research on expressive motion in robots [33]. An important challenge in enabling robots to display expressive behavior lies in identifying motion characteristics associated with expressiveness and understanding how they relate to the intended expressive qualities. Currently, techniques for acquiring expressive motion characteristics can be divided into three categories: manually selected features, expert features, and learned features.

Manually selected features refer to the replication of expressive movements generated by human performers in robots, often combined with imitation [27] and repositioning methods [16]. Despite the robot's attempt to replicate human-generated movements, the expressiveness is constrained by the inherent differences between the robot's physical form and the human body. Furthermore, this imitation-based and repositioning method cannot identify which components of movement contribute to expressiveness and therefore cannot be generalized to new interaction scenarios.

The two most common sources of expert features are the principles of animation [30] and the Laban features [32]. The advantage of expert features overcomes the limitations of manually chosen features. For instance, they can be applied to various types of robots. Nevertheless, the need for skilled professionals to evaluate and design robot actions for specific scenarios restricts the applicability of this method in novel situations.

The method of learned features refers to the approach of attempting to learn relevant features between the motion characteristics and expressiveness from labeled human action data [22]. For example, [26] utilized a text-to-motion retrieval method based on learned textual action features to generate expressive movements tailored to specific scenarios. To guarantee the expressiveness of the generated human motions, the learned feature method depends on extensive human motion data. Currently, this learning-based approach is widely accepted, e.g., in the field of generating expressive motions for digital humans. However, like the other two methods, this learning-based

approach faces similar limitations: it is difficult to produce expressive behavior that meet the expectations of end-users when their feedback is vague during interactions.

Recent advances in LLMs offer a viable way to address this dilemma. Specifically, the remarkable reasoning capabilities of LLMs are introduced to estimate end-user expectations. Recent works have achieved significant success in downstream robot tasks of HRI, such as task planning [7] and navigation [8], by leveraging LLMs for their exceptional contextual analysis and task planning capabilities. [15] has demonstrated that LLMs can effectively leverage contextual information and respond to human feedback. Furthermore, LLMs are capable of achieving social and common sense reasoning [19] and inferring user preferences by summarizing interactions with humans [38].

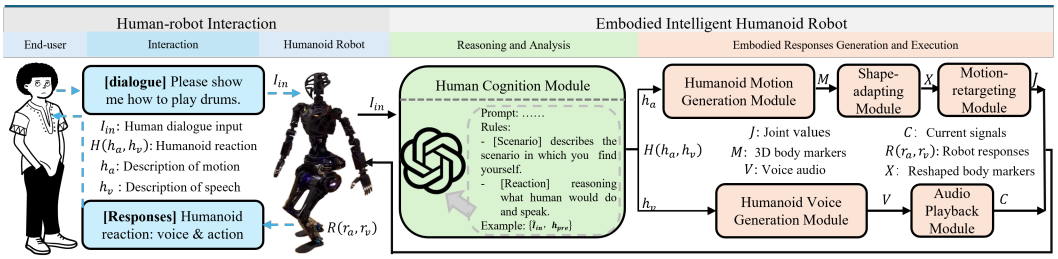

Fig. 2. Overview of the expressive behavior generation framework for a humanoid robot. During interactions, the end-user specifies requirements via dialogue with the humanoid robot. Subsequently, the robot intelligently infers human-like responses within this dialogue scenario and executes responses through embodied information to satisfy the end-user's requirements.

The benefits of LLMs establish the groundwork for robots to interpret unclear user feedback, thereby enabling the generation of expressive behavior aligned with user preferences. A social robot framework [24] utilizes LLMs to analyze end-user social media account data, infer user interests, and engage in expressive dialogue with users. This approach addresses the issue of traditional social robots using pre-set monotonous communication mechanisms, which leads to user disengagement due to boredom during the interaction process. Hence, enabling robots to generate more diverse behaviors using LLMs can enhance user engagement in HRI. Several studies [23, 37] employ LLMs to analyze human behavior patterns and guide nonhumanoid robots to perform expressive and comprehensible actions in response to human vocal or gestural feedback. However, these behaviors typically only involve predefined simple motions, such as nodding up or changing the status of the indicator lights, which are far from capturing the variety of expressive human behaviors. Moreover, the instruction–skill pair mechanism of [23] reduces the reasoning ability of robots as the LLM-generated behavior outside the predefined action library cannot be performed. Consequently, it is better to transform instructions into behaviors directly instead of instruction–skill pair mechanisms, which maximizes the reasoning performance of LLMs and improves the consistency of generated actions with the interaction scenario.

To tackle this issue, [43] introduced a novel approach to produce expressive behavior in humanoid robots by mapping language expressions of complex human behaviors to the robot's body motions. Additionally, this method allows for direct control of the robot's hardware by LLMs, enabling the robot to generate more sophisticated behaviors. However, this approach requires extensive interactions to enable humanoid robots to perform expressive behavior, leading to increased deployment costs. Similarly, [14] uses LLMs to directly generate executable gestures for humanoids with limited demonstrated data captured from vision-pro. However, due to the lack of grounded robot-embodied information, the generation ability of LLMs is reduced. For example, the generated

behavior is hardware-incompatible. To address this issue, we only utilize LLMs for high-level semantic reasoning and a specialized generative model to generate behaviors that are consistent with the semantics of LLM reasoning results, achieving coordinated and multimodal behaviors across tasks and hardware platforms.

3 Methodology

This investigation focuses on the HRI scenario illustrated in Fig. 1, where a humanoid robot fulfills the interaction requirements from end-users by demonstrating expressive behaviors H, encompassing speech h_v and motion h_a. Generally, the expressive behavior refers to multi-modal interactive behavior (motion&speech in our study), designed to naturally convey intentions, emotions, or semantic information through responses aligned with human cognition and social conventions. These behaviors inherently reflect anthropomorphism and contextual adaptability. Specifically, expressive behaviors are affected by a triad of fundamental prerequisites: 1) multimodal coordination, i.e., motion (h_a) and speech (h_v) must be synchronized semantically and temporally. 2) human-likeness, i.e., behaviors must conform to human kinematic characteristics, avoiding mechanical rigidity. 3) contextual appropriateness, i.e., behaviors must precisely match interaction scenarios and user intent, preventing irrelevant responses that are misaligned with user input. To address this challenge, we propose a novel framework, illustrated in Fig. 2. The framework uses LLMs (e.g., GPT-4) to interpret human input dialogue and employs two generation models to produce suitable expressive behaviors for humanoids. However, LLMs possess limited knowledge of robot embodiment. Moreover, humanoid robots differ substantially from human physical structures. As a result, directly translating generated behaviors into executable joint and voice signals remains challenging. Therefore, two design objectives are proposed for this method:

- Enabling the robot brain (LLM) to generate expressive behaviors that align with the human expectations and cognitive understanding. Specifically, this objective requires the behavioral responses to align with users' expectations in dialogue for robot feedback, which are inferred by LLM from their immediate dialogue context (e.g., a specific request for a type of gesture or action). Simultaneously, the behaviors must be intuitively understandable (cognition) to ensure effective communication.
- Maximizing expressiveness when adapting generated human-like behaviors to humanoid robots.

In pursuit of these two objectives, we present a human cognition module (green block) featuring a meticulously crafted prompt to facilitate the generation of human-like behaviors by the LLMs in HRI scenarios. Subsequently, an embodied response generation and execution system (orange block) is developed to retarget humanoid behavior into robot actions. By achieving the two objectives (green and orange blocks in Fig. 2), we establish a connection between the physical structure of the robot and LLMs, thereby allowing the control of humanoid robots to generate expressive behaviors through an advanced decision-making system based on LLMs.

3.1 Problem Statement

Our solution in enhancing end-user engagement in HRI is to develop a framework which can drive humanoids performing expressive behaviors by leveraging LLMs that are augmented with vast human knowledge. To quantify the behavioral expressiveness throughout the generation and retargeting process, we introduce the distance function $dist(R(r_a, r_v), R(r_a, r_v)_{exp})$, where r_v and r_a represent the robot's speech and action responses. It is imperative to underscore that this paper's primary emphasis is on the expressiveness of humanoid robot motions. This function measures the discrepancy between the expert human body trajectory $R(r_a, r_v)_{exp}$ and the robot

behavior trajectory $R(r_a, r_v)$, as referenced in [23]. Expert trajectories are extracted from publicly available human behavior datasets, including the CMU MoCap dataset and the BEAT dataset [20], which are characterized by specific body points. The aim of this research is to maintain the expressiveness of humanoid behavior during generation and retargeting by minimizing the distance $d^* = \min(dist(R(r_a, r_v), R(r_a, r_v)_{exp}))$.

3.2 Reasoning and Analysis

In this section, our goal is to achieve the first objective of guiding robot intelligence (LLMs) to describe natural humanoid behaviors in text (green block in Fig. 2). Furthermore, the described behaviors in text are supposed to meet the end-user's intention and preference. Nevertheless, it is generally infeasible to manually predefine the user's intent and preferences explicitly to the robot in open-world, unconstrained HRI scenarios. Therefore, the robot must actively infer this information in real time. Among the available input signals from end-users, spoken language serves as a critical channel through which users implicitly communicate their goals, preferences, and emotional states. Based on this, our framework begins with user speech as input, from which robots can infer user's latent intent and preference information. This inferred knowledge is then used to drive the generation of expressive behaviors in humanoid robots. To accomplish this objective, we provide a carefully designed prompt within the Human Cognition module to enable the LLMs not only to comprehend and interpret user requirements but also to provide the appropriate behavior a human would do in this context.

Human cognition module: This module utilizes the end-user dialogue $I_{in} \in L$ as its input. This input may encompass the end-user's description of the expressive behavior to be enacted by the humanoid robot (e.g., "waving your hands", involving physical interaction) or everyday conversations (e.g., "Hello, nice to meet you", devoid of physical interaction). The LLM output is represented as a string denoted by $H(h_v, h_a)$. For example, when presented with the input I_{in} = *"See you!"*, the human cognition module would produce h_v = *"See you too!"*, and h_a = *"Wave or nod as a sign of acknowledgment; maintain eye contact to demonstrate engagement with the conversation until its end."* This module leverages LLM's contextual reasoning to implicitly infer preference and intention from user dialogues. Importantly, LLMs have been shown to handle such inference tasks with high accuracy and have already been widely adopted in the robotics domain, such as dialog understanding [15]. It is crucial to highlight that the LLMs use chain-of-thought reasoning to produce H.

Specifically, we adopt GPT-4 (gpt-4-0613) as the backbone language model for the reasoning module. This decision was made based on GPT-4's demonstrated stability, semantic reasoning capacity, and mature API access during our system development phase. Moreover, GPT-4 had been widely validated in generating grounded and coherent responses to complex user instructions. It should be noted that the goal of this module is not to benchmark LLMs, but to utilize them as a semantic reasoning engine for generating contextually appropriate behavior descriptions. Our task involves medium-complexity user instructions (e.g., storytelling, demonstration, expressive gestures), which are well within the reasoning capabilities of current state-of-the-art models. Therefore, GPT-4 can be replaced by other capable LLMs (e.g., DeepSeek) without altering the prompt design.

To ensure that the robot brain (LLMs, Human Cognition Module) emulates human-like thinking and aligns with end-user expectations, we designed prompts to guide robot responses. The prompt prefix encapsulates the definition of the LLMs' role and the output format. Initially, the prompt enables the robot to construct an identity role within its cognitive framework, that is, a sociable individual. Subsequently, we establish rules to ensure that the robot's focus remains on the dialogue content of the end-user and generates feedback dialogue and actions. These rules and examples are

provided on the project website. Specifically, the robot must recognize the conversational context I_{in} and reason about appropriate verbal and physical responses as a sociable individual in this scenario. To guide the model outputs, we provide a small number of predefined example pairs $\{\hat{I}_{in}, h_{pre}\}$ for the system (LLMs) prompting and fine-tuning process. Therefore, LLMs can infer the end-user's intentions and preferences based on the prompt requirement, and then generate feedback $H(h_v, h_a)$ based on the estimated information. The mechanism draws theoretical support from LLM's ability to summarize interaction history as shown in [38], which we explicitly cited as the foundation for preference estimation in HRI contexts. Furthermore, this implicit modeling aligns with our design principle of contextual appropriateness, ensuring behaviors match user intent without explicit profiling.

3.3 Generation and Execution of Action Behavior

In this section, our goal is to achieve the second objective, which is to convert the descriptive text generated by LLMs into control signals for humanoid robots. As illustrated by the orange block in Fig. 2, we utilize two separate pipelines to handle the responses of speech and action. We introduce three modules designed to process robot action responses, thereby allowing the humanoid robot to generate expressive humanoid behavior and retarget it to its body while preserving the behavioral expressiveness. In detail, within the motion generation module, the text-to-motion/gesture model is employed to produce motion M from text input h_a. Then the modified motions M would be processed by the shape-adapting module to get body marks X. Finally, the processed human behavior X is converted into the robot control signal J by the motion-retargeting module.

Motion generation module: This module aims to generate human-like actions M based on the input text h_a. In this field, a multitude of outstanding studies have been performed, including those focusing on text-to-motion [45] and speech-to-gesture [18]. Consequently, we have chosen to leverage these proposed state-of-the-art models as our motion generation model instead of developing a new one. This approach not only reduces development costs, but also offers a practical framework for transitioning research in the motion generation field from virtual humans to real robots. To adapt these state-of-the-art motion generation techniques for humanoid robots, two primary challenges must be addressed:

- Action stability: The action data utilized for model training are collected from human-demonstrated behaviors, which pose significant challenges to stability control of robots. For example, executing actions such as "performing Tai Chi" may not be reliably achieved by the robot controller, potentially resulting in demonstration failures.
- Action retargeting: Due to the disparity in physical dimensions between virtual humans and humanoid robots, accurately mapping anthropomorphic actions from virtual models to physical robots presents a significant challenge (discussed in Sections 3.3 and 3.3).

To address the issue of action stability for generated motions, we employ a modification operation to ensure the generated motion could be performed by humanoid robots. The primary approach of this modification operation (see Eq. 1) is to align the lower body (leg) movements of demonstrated actions with three fundamental motion modes imposed by humanoid robots, i.e., standing, walking, and turning gait. Specifically, the standing gait is used when no leg movement is required, the walking gait is adopted when forward locomotion is presented, and the turning gait is used when the robot walks in a non-straight direction.

$$g_t = \begin{cases} g_{\text{standing}}, & \text{if the robot remains stationary or gesturing in place} \\ g_{\text{forward}}, & \text{If the robot is walking in a straight direction} \\ g_{\text{turn}}, & \text{If the robot is walking in a non-straight direction} \end{cases} \tag{1}$$

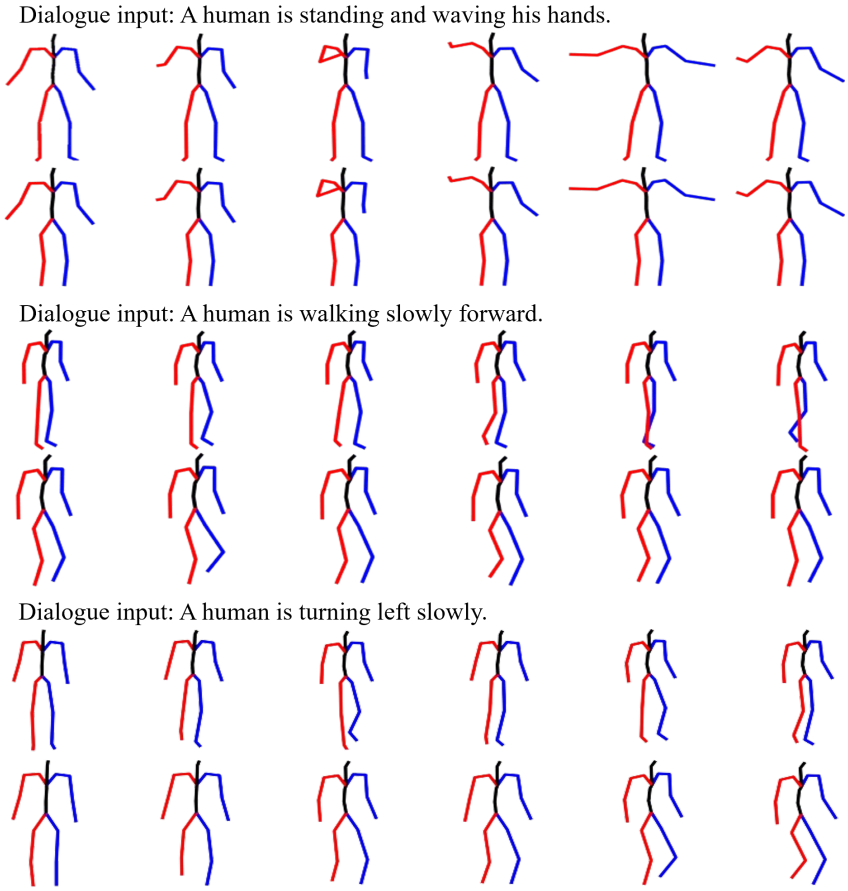

Fig. 3. Demonstrations of the original and modified humanoid motions corresponding to end-user dialogue inputs. These humanoid motions have been generated using the Motiondiffus model with pre-trained parameters. The first block represents the original and modified humanoid motion for standing gait. The second block corresponds to straight walking gait,. The third block corresponds to turning gait.

As illustrated in Fig. 3, we provide an example demonstrating the modification of three generated human actions by using humanoid robot motion gaits. In Fig. 3, the first line of each block illustrates the original generated motions, while the second line depicts the modified motions. By applying this modification operation to the generated action, we can rigorously leverage pre-trained generative models from existing works without the need for retraining. Before mapping these actions to the robot's body, we ensure the stability of the modified actions through simulation validation. This validation process is conducted within a physics-based simulation environment (i.e., Pybullet), where the generated actions are systematically evaluated. Specifically, we check for violations of joint kinematic constraints, self-collisions, and dynamic instability. These issues may compromise safe execution on humanoid platforms. Actions that fail to meet these safety criteria are discarded, and the generation module is triggered to produce alternative candidate actions. Only actions that successfully pass all validation checks are considered safe and subsequently deployed in humanoid robots. Although the leg movements are reduced into three basic gaits, we intentionally focus on

Fig. 4. Different human body shapes for motion generation models [39].Different human morphology models have different body parameters, such as height and arm length.

upper-body expressive behaviors due to both practical and interactional considerations. From a technical perspective, generating and executing full-body behaviors, especially complex lower-body actions (e.g., stepping or crouching), poses significant challenges to current commercial humanoid robots. These include instability, hardware limitations, and a lack of reliable controllers that can safely execute such motions. From a social interaction perspective, [35] shows that upper-body movements and speech are the primary channels for emotional expression and communicative intent during human-robot interactions. Therefore, we focus on the upper-body motion of humanoid robots in this study to maximize behavioral expressiveness and platform safety in real-world HRI scenarios.

Body shape-adapting module: Once the humanoid motion generation module is deployed, whether it utilizes state-of-the-art pre-trained models or self-trained ones, it is essential to implement body shape adaptation prior to translating the motion to robots. This is necessary because these generative models can accommodate various human body shapes (as depicted in Fig. 4), relying on different parametric human body models, such as SMPL [21] and GHUM [39]. Given that the joint structure (link size) of these digital avatars with varying body shapes differs from that of the humanoid robot, direct retargeting of these motions to the robot's body can lead to several issues, such as inconsistencies in the relative positions of the hands between the generated digital human model and the robot. Consequently, the adaptability of the motion retargeting model is reduced, and the expressiveness is diminished too. For example, [4] retargets the human motion to the humanoid robot by directly mapping joint rotations without structural adaptation. This often leads to distorted or infeasible motions due to discrepancies in limb proportions and joint limits. To address this issue, we introduce a shape-adapting module to map human body shapes to robot morphology before retargeting motion to humanoids.

Specifically, we firstly selected ten key markers from the generated digital human model and the humanoid robot (see Fig. 5) respectively. Subsequently, we normalize the body shape of digital models of various sizes to fit a humanoid robot platform. As illustrated in Fig. 5, the body position p_i^r of the humanoid is determined using the markers of the digital human model (refer to Eq. 2), by leveraging the kinematic chains within the coordinate system of the robot's torso [44].

$$p_i^r = p_{i-parent}^r + \frac{p_i^m - p_{i-parent}^m}{L_i^m} * L_i^r \tag{2}$$

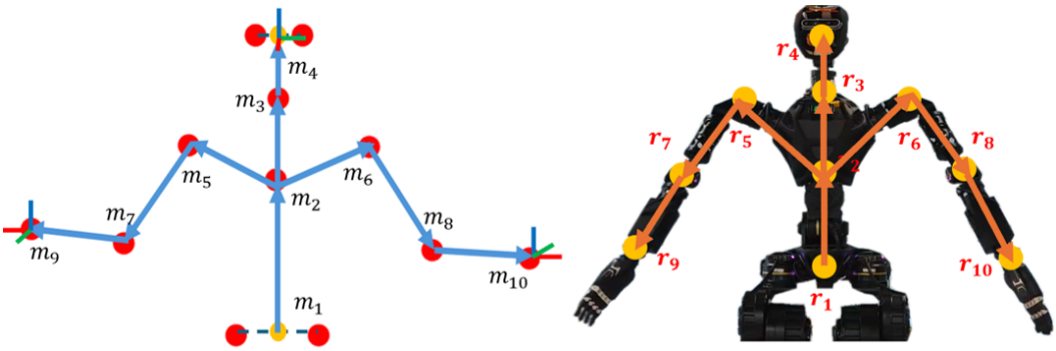

Fig. 5. The 10 key markers selected for shape-adapting operation. The left shows the human model used to generate motions M. The right is the humanoid robot used to perform generated motions.

Where, robot key point $p^r_{i-parent}$ is the parent point of p^r_i and the definition of the human key point $p^m_{i-parent}$ is the same. L^r_i (L^m_i) refers to the link length between the key point of p^r_i (p^m_i) and $p^r_{i-parent}$ ($p^m_{i-parent}$). For example, the coordinates of the left elbow p_8 of the robot can be determined using Eq. 3.

$$p_8 = p_6 + \frac{m_8 - m_6}{\|m_8 - m_6\|} * \|r_8 - r_6\| \tag{3}$$

Upon obtaining the coordinates of the 10 robotic body markers, we utilize quaternions q_i to represent the rotation pose of the markers m_4, m_9, m_{10}. Lastly, we reshape the data of the resized body markers into a vector denoted X, which is the input of the motion-retargeting module.

Motion-retargeting module: To transfer the shape-adapted expressive behavior onto a humanoid robot, we need to perform IK resolution to find joint values J based on the pose condition X. Due to the variations between humans and humanoid robots, including the restrictions imposed by robotic hardware, it is challenging to utilize only the IK algorithm to retarget human actions on humanoid robots while preserving the quality of expressive motions [4]. To address this problem, this study develops a local search method using a genetic algorithm (GA) to determine joint values that align with the intended humanoid actions, thereby preserving the expressiveness. In detail, we introduce the local search algorithm 1 to determine the joint values J of the humanoid robot in pose X. First, we calculate the initial joint J using an IK solver with certain constraints. As the results discussed in [4], significant morphological differences exist between the pose of the humanoid robot calculated by the IK solver and the input of the human pose X. To address these differences, we have employed a three-step genetic algorithm (GA) due to our observation that conducting separate searches for body joints, elbow joints, and wrist joints is more efficient in terms of time cost compared to using a single GA to search for all joints simultaneously. When all fitness values f_i calculated by Eq. 4 are below the threshold F_i, we obtain the optimal joint configuration J. F_i is introduced as a posture modulation factor to control the adjustment step between the current robot pose and the desired target pose. This factor directly affects convergence results, i.e., a smaller value leads to smaller deviations between the computed posture and the target. In our implementation, F_i is empirically set to 0.001 for all scenarios. This conservative setting balances motion expressiveness with kinematic stability.

$$f_i = 1 - \frac{1}{N}\sum_{j=1}^{N} exp^{(-\|(p_j - \hat{p}_j)\|)} \tag{4}$$

Algorithm 1: Local Searching Method for Joint Angle J

input : Cartesian space pose X
output : Joint space pose J

$J = IK_{constraint}(X)$;
repeat ;
  $[J_1, f_1] = GA\_body(J)$;
  $[J_2, f_2] = GA\_elbow(J_1)$;
  $[J, f_3] = GA\_wrist(J_2)$;
until $f_i \leq F_i$;
Note : GA_body encodes six joints of the body and head; GA_elbow encodes six shoulder joints of both arms; GA_elbow encodes eight joints of both arms (3 wrist joints and elbow joint for each arm);

Although the algorithm effectively minimizes the loss of expressiveness in the motion retargeting process, its deployment in real HRI scenarios is time-consuming. Nevertheless, the real-time HRI demands minimal response latency. Therefore, we create an MLP model with three advantages to process the IK-solving problem. First, the MLP can speed up the IK-solving process when compared with the traditional IK algorithm. This benefit fundamentally resolves interaction lag and ensures responsive robot behavior. Second, the MLP demonstrates clear advantages in handling this complex nonlinear transformation, particularly in managing one-to-many mappings inherent in systems with redundant degrees of freedom. It offers strong fitting capacity while avoiding trajectory distortion often introduced by constraint simplifications in analytical methods. Finally, the MLP balances precision and efficiency when compared with other learning-based methods, such as transformers. Importantly, the use of MLPs in inverse kinematics tasks has already gained practical consensus in the robotics community, as evidenced by prior work such as CycleIK [13].

To train this MLP, we create a customized dataset. For conventional IK tasks, as shown in [12], the pose condition in the dataset mainly includes Cartesian poses of the robot's end effectors with some joint constraints. In contrast, our dataset includes the positions of ten upper-body markers and the rotation poses of three end effectors (both hands and head) to keep the expressiveness of the human motions. Therefore, the solution based on our dataset can precisely map the generated expressive human-like motions onto humanoid robots. This distinction fundamentally differentiates our dataset from those used in conventional IK-solving tasks. Unlike other methods that build datasets from uniformly random samples in the robot's joint space, we create a tailored dataset by gathering data pairs generated by our algorithm 1. We gather data from the Cartesian space X and the joint space J as samples $\{X, J\}$ to train the MLP depicted in Fig. 6. Training, validation, and test sets are created separately. The training set includes $400,000$ samples, while the validation and test sets consist of $40,000$ and $80,000$ samples, respectively.

The MLP contains six to eight layers, and the exact number of layers is part of optimized hyperparameters, as is the number of neurons in each layer [13]. The detailed information of final MLP parameters is provided on our website. The network input is a batch of 1-dimensional vectors X with 42 parameters (see Fig. 6 Cartesian space). The vector consists of 10 up-body 3D marker positions $p_n = (x_n^p, y_n^p, z_n^p)$ and 3 rotation poses $q_n = (x_n^r, y_n^r, z_n^r, w_n^r)$ for the head and both hands. The network output is the robot joint space J which consists of 20 joint values (3 for body, 3 for head, and 7 for each arm), normalized into the interval $[-1, 1]$ activated by the function $tanh$.

Referring to the work of the neural IK solver CycleIK [12], we train the MLP based on the framework in Fig. 6. The MLP learns to solve the IK problem by using the inverse property between the functions $FK(J)$ and $IK(X)$ in Eqs. 5 and 6.

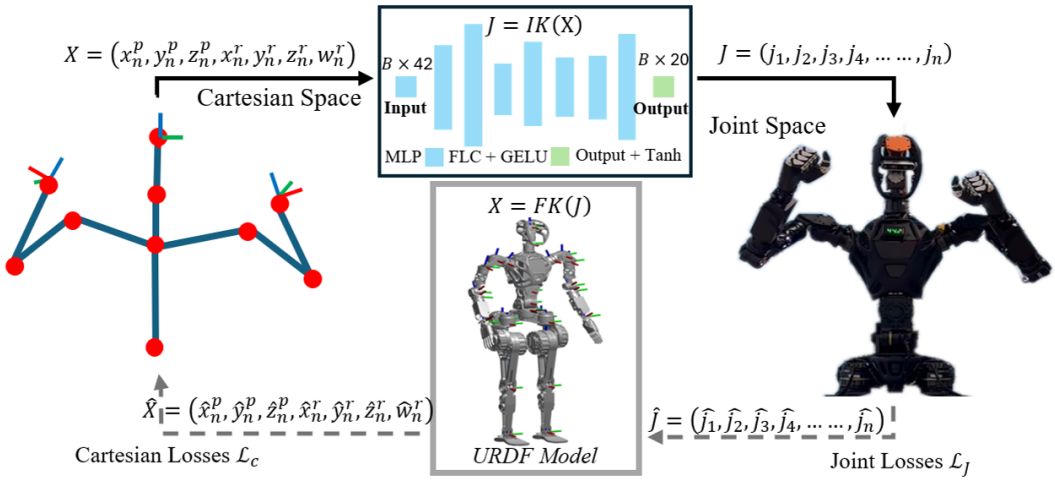

Fig. 6. Overview of the MLP training process based on CycleIK algorithm for robot motion-retargeting module. The algorithm aims to predict a set of valid robot configurations J under constraints $\mathcal{L}_C$ and $\mathcal{L}_J$ by inferring a batch of Cartesian poses X.

$$\hat{X} = FK(IK(X)) \tag{5}$$

$$e_{IK} = \|X - \hat{X}\| \tag{6}$$

When training the MLP network, various metrics can be utilized to update the training losses. Specifically, we can select the Mean Squared Error (MSE) as the metric for joint losses $\mathcal{L}_J$, or select the smooth loss L_1 [9] as the primary metric for Cartesian losses $\mathcal{L}_C$. Drawing from the approach in [12], we treat the joint space as a semi-hidden domain and calculate the positional and rotational losses solely in the Cartesian space. This is accomplished by mapping back to Cartesian space in a full cycle, thereby minimizing the significant errors linked to redundant manipulators [17], which exhibit a one-to-many relationship within the redundant nullspace manifold J. Consequently, we select Cartesian losses $\mathcal{L}_C$ (Eq. 7) as a loss function for training the MLP.

$$\mathcal{L}_C = \omega_{pos} * \mathcal{L}_{pos} + \omega_{rot} * \mathcal{L}_{rot} \tag{7}$$

In Eq. 7, $\mathcal{L}_{pos}$ denotes the positional loss, computed as the smoothed L_1 distance $l(a, b)$ between the body position p_n and the predicted $\hat{p}_n$. In addition, ω_{pos} and ω_{rot} are two weight parameters. The rotational loss, denoted as $\mathcal{L}_{rot}$, is calculated using the SMQL method introduced in [12]. Furthermore, $q_{i,j}$ refers to the j-th dimension of the i-th orientation in Eq. 8.

$$\begin{cases} \mathcal{L}_{pos} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\frac{1}{3}\sum_{j=1}^{3} l(p_{i,j}, \hat{p}_{i,j}) \\ \mathcal{L}_{rot} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} min\{\sum_{j=1}^{4} l(q_{i,j}, \hat{q}_{i,j}, \sum_{j=1}^{4} l(-q_{i,j}, \hat{q}_{i,j})\} \end{cases} \tag{8}$$

Generation and execution of speech behavior: To facilitate interaction between the end user and the humanoid robot, a six-channel microphone array (M260C, produced by Iflytek Co., Ltd), a USB sound collection card, and a speaker are installed on the head of the humanoid robot. During

HRI, the robot utilizes the Whisper model [28] to translate the end-user's dialogue collected by the microphone array into text. Subsequently, a Text-to-Speech (TTS) model[1] is employed to convert the text h_v generated by LLMs into voice audio V within the voice generation module. Finally, voice audio V is processed into current control signals C to drive the speaker in the audio playback module. Following this procedure, the humanoid robot can generate voice responses to facilitate interaction.

4 Study 1: Perceptual Study on Robot Behavior

4.1 Study Objective and Hypotheses

This study aimed to evaluate whether our method can enable humanoid robots to perform expressive behaviors that align with end users' personalized preferences (i.e., expectations) and cognitive understanding. To test this objective, we formulate four hypotheses (H0–H3). H0 and H3 concern practical equivalence and are evaluated using bootstrap-based equivalence testing with predefined equivalence bounds, whereas H1 and H2 are directional hypotheses evaluated using classic null-hypothesis significance testing.

- H0: Participants' ratings of understanding confidence and behavioral communication clarity of robot feedback under the generated motion and voice feedback (GMVF) condition are statistically equivalent to their ratings under the other three conditions: only voice feedback (VF), only motion feedback (MF), and designed motion and generated voice feedback (DMVF).
- H1:Robot behaviors in the GMVF condition are perceived as more appealing and more expressive than behaviors in the VF condition.
- H2: Robot behaviors in the GMVF condition are perceived as more appealing and more expressive than behaviors in the MF condition.
- H3: Participants' perceived expressiveness and appeal of robot behaviors in the GMVF condition are statistically equivalent to those in the DMVF condition.

4.2 Study Methodology

This study employed an online within-subjects, video-based experimental design to examine the effects of robot feedback modalities on users' perception for behavioral expressiveness and appeal under controlled conditions. This approach provides a controlled, repeatable, and scalable way to compare different feedback modalities under consistent conditions, while minimizing variability introduced by real-time interaction. Thereby, this study allowing us to focus on perceived quality. Importantly, this study used a within-subjects design with one four-level independent variable, namely the robot feedback modality during interaction, consisting of: VF, MF, DMVF, and GMVF.

All four response types were generated for the same interaction situation (i.e., identical dialogue input). Accordingly, observed differences across conditions can be attributed to the feedback modality rather than to variability in LLM outputs. The dependent variables are four questionnaire scores collected in our online study, i.e., understanding confidence for robot behaviors, communication clarity, behavioral expressiveness, and behavioral appeal of robot behaviors. The following subsections describe the detailed content of our study method, including the experimental design, interaction scenarios, study procedure, participants, and data analysis strategies.

Experimental design: A within-subjects experimental design was adopted, in which all participants watched four robot feedback modalities in the same interaction situation. The four feedback modalities were defined as follows:

- VF: Only **V**oice h_v **F**eedback.
- MF: Only **M**otion h_a **F**eedback.

[1]https://docs.coqui.ai/en/latest/

- DMVF: **D**esigned **M**otion h_d and Generated **V**oice h_v **F**eedback.
- GMVF: **G**enerated **M**otion h_a and **V**oice h_v **F**eedback (our method).

In particular, the motion h_d has been carefully designed through a collaborative effort involving a professional animator and two robotics experts. This design process takes into account the contextual dynamics inherent in human conversational input. The motion h_a and the voice h_v are generated by our framework. For each scenario, we called the LLM once to generate a single behavior description, which was then instantiated into four modality conditions (VF, MF, DMVF, GMVF) to eliminate stochastic variation. We chose this setup specifically to ensure experimental control, i.e., calling the LLMs once per scenario allowed us to remove any variability that might arise from stochastic generation. For the VF and DMVF condition, the robot executes only the speech signal h_v generated by the framework, while the corresponding motion output h_a is omitted. Conversely, in the MF condition, only the generated motion signal h_a is executed, and the speech output h_v is omitted.

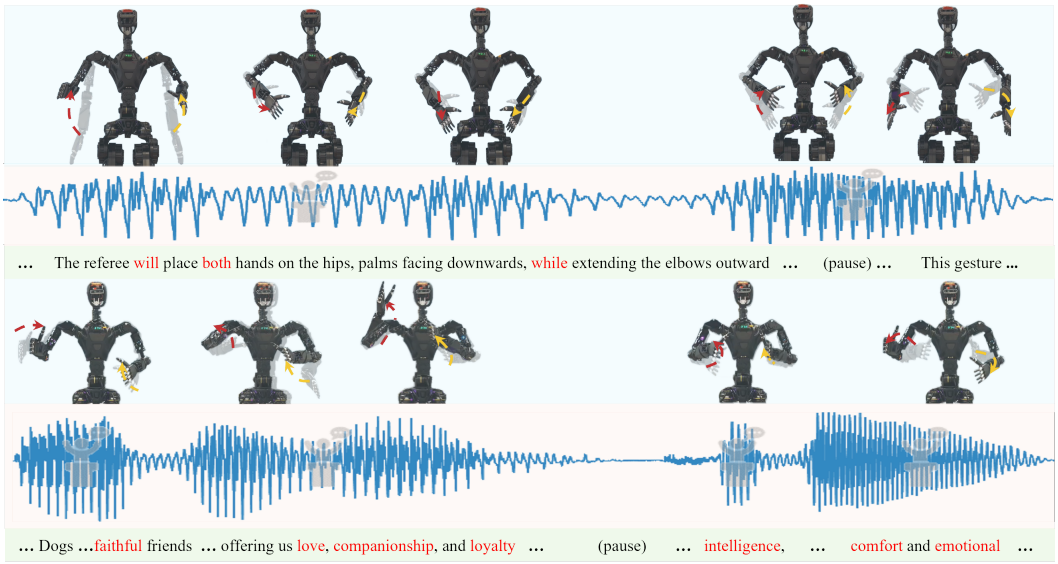

Fig. 7. Demonstration of robot expressive behaviors in an HRI scenario. For the results in the first part, the end-user's dialogue input is *"Could you show me some basketball gestures?"*. For the behaviors in the second part, the end-user's dialogue input is *"Could you tell me a story about dragons?"*. For each part, the first line is the GR1 action response while the second line is the speech response. The third line is the text of the robot speech response.

Interaction situation: We constructed seven situations representing two types of robot behaviors. The first type, instrumental-demonstrative actions, comprises object-directed or action-directed behaviors that convey demonstrative or instrumental information. They largely overlap with referential co-speech gestures, e.g., iconic/metaphoric depictions and deictic pointing. The second type, called discourse-oriented co-speech gestures, refers to gestures that accompany speech and primarily support discourse management rather than object-directed meaning. They are largely overlap with non-referential gestures (e.g., beat gestures and other discourse-management gestures), which structure and emphasize speech rather than depict object content.

For instrumental-demonstrative actions, we selected five representative scenarios of increasing complexity: social etiquette (e.g., greeting), common-sense gestures (e.g., basketball gesture),

animal mimicry (e.g., avian flight), musical performance (e.g., playing drums), and dramatic expression (e.g., enacting "I am a little teapot"). These actions in selected scenarios serve instrumental-communicative operations (e.g., teaching basketball gesture), which can serve as role models for the observer. Moreover, these categories of the five particular situations align with the established classification of the widely recognized CMU Mocap dataset. By covering the representative behavioral spectrum from basic social etiquette (greeting) to complex performance (drama), we aimed to rigorously test the framework's ability to generate expressive behaviors. In this process, we integrated the MotionDiffuse model [45] into our framework, which serves as a generative model.

For discourse-oriented co-speech gestures, we designed two speech-centric situations ,e.g., a TED-talk style presentation and casual storytelling, that primarily elicit non-referential gestures used for discourse functions such as emphasis, rhythmic structuring, and turn management. Importantly, "non-referential" does not mean that the gestures are meaningless; rather, it indicates that they are not primarily object/action-directed. Specifically, these two settings effectively represent the behavioral spectrum of discourse-oriented co-speech gestures. The TED Talk scenario imposes a structured, professional speaking context demanding measured and supportive gestures. Conversely, the storytelling scenario represents a more spontaneous, informal context allowing for potentially more varied gestures. This dichotomy ensures robust testing of the framework's ability to generate appropriate co-speech gestures tied to conversational style rather than specific semantic content. Within these scenarios, we integrated the QPGesture model [42] into our framework, which functions as the generative model for gestures semantically aligned with speech.

To support online video-based survey, we record the interaction scenarios as survey materials. As our study mainly focuses on the expressiveness of robot behavioral feedback, the interaction scenarios were presented as repeated conversational exchanges within predefined situations. In this setup, a standardized greeting utterance (e.g., "Hello, GR1, happy to meet you") was used in the greeting scenario. Then, the robot provided behavioral feedback in four feedback modalities, respectively. We recorded this conversation as stimulus videos for online study (see the project website), and the participants evaluated the robot feedback behaviors based on recorded video clips, instead of interacting with the robot directly. Generally, robot speech responses were generated using the GPT-4 API (gpt-4-0613), and motion responses were produced either by our framework (GMVF, MF) or expert animators (DMVF). All responses were executed on the GR1 humanoid robot and recorded as video clips. To ensure comparability across conditions, all recordings were made in the same physical environment with controlled lighting and viewpoints.

Study procedure: Before the study, all participants provided written informed consent. Participants were first briefed on one interaction situation and then viewed four modality-specific videos (i.e., VF, MF, DMVF, GMVF) corresponding to that situation. Specifically, the four videos present the robot's responses under all four feedback modalities in the same interaction situation. This procedure was repeated for all seven interaction situations, resulting in 28 (7 situations $\times$ 4 modalities) videos watched by each participant. The presentation order of the seven interaction situations was randomized for each participant. To minimize potential order effects in the video presentation, a balanced Latin square design was used to counterbalance the order of the four feedback modalities (i.e., the four videos) within each situation across participants. Following the viewing of each video, participants were required to respond to four items rated on a 7-point scale, designed to measure participants' confidence in their understanding of the robot's behaviors and their perceptions of the behaviors' expressiveness.

- Q1. How confident are you that you correctly understood the meaning or intention behind the robot's behavior?

- Q2. To what degree did the robot's behavior convey a clear and unambiguous message to you?
- Q3. How well did the robot's behavior demonstrate expressiveness that was appropriate to the scenario's requirements?
- Q4. How engaging or motivating did you find the robot's interactive behavior in this context?

In particular, questions 1 and 2, derived from [34], aim to gauge semantic clarity and comprehension of intent and thus assess participants' understanding of the robot's feedback behavior. These two dimensions establish the "cognitive foundation", i.e., if users do not understand the robot's intent, subsequent evaluations of behavioral expressiveness and appeal are unreliable. Question 3 focuses directly on expressiveness and its fit with the given context. It evaluates both the richness of the motion and alignment between motion and the dialogue scenario. This item reveals whether the robot's behaviors genuinely reinforce the conversational intent rather than merely appearing "decorative". Question 4 evaluates how effectively robot behaviors encourage users to continue the interaction, i.e., the behavioral appeal. This dimension highlights that clarity and contextual appropriateness alone may not suffice. Moreover, robot behaviors must also be engaging to sustain effective human–robot interaction, even when actions are clear and contextually appropriate.

Participants: We invited 27 users (15 men, 12 women) to participate in this survey, and they are aged 7 to 72 ($M = 34$ and $SD = 15.45$; $7-16:2; 18-27:7; 28-37:11; 38-47:3; 47-72:4$). For participants under the age of 18, written informed consent was obtained from a parent or legal guardian in addition to the child's assent. 24 participants self-identified as Asian, 3 self-identified as White. The wide age range was intentional, as we aimed to include users from different stages of life and cognitive backgrounds to better assess the generalizability of the result. Participants were recruited through voluntary sign-ups via public advertisements posted on the university campus and local online community platforms. Before the study, all participants provided informed consent and were briefly introduced to the purpose of the experiment. The study was approved by the institutional ethics review board.

Data analysis strategy: Although participants rated robot behaviors in multiple interaction scenarios, our confirmatory analyses target differences between feedback modalities that are independent of specific scenarios (H0 to H3). Therefore, for the primary confirmatory analyses, participants' ratings were aggregated across the seven scenarios within each feedback modality. Scenario-level analyses are reported as supplementary analyses to examine the robustness and consistency of the observed effects across different interaction contexts. Four 7-point questionnaire items were collected for each feedback condition, assessing understanding confidence (Q1), communication clarity (Q2), perceived expressiveness (Q3), and behavioral appeal (Q4). Internal consistency of the questionnaire was evaluated using Cronbach's α, and sampling adequacy was assessed using the Kaiser–Meyer–Olkin (KMO) measure and Bartlett's test of sphericity. These indices were computed both across all scenarios and separately for each interaction scenario to ensure that the questionnaire functioned consistently across contexts.

We used two complementary inferential approaches. Equivalence hypotheses (H0 and H3) were evaluated using equivalence testing, whereas directional hypotheses (H1 and H2) were evaluated using conventional null-hypothesis significance testing. To test H0, we conducted bootstrap-based equivalence tests with equivalence bounds of $\Delta = \pm 0.5$ on the 7-point scale. For each pairwise comparison (e.g., GMVF vs. DMVF), we computed 90% percentile bootstrap confidence intervals for the mean paired differences. Practical equivalence was concluded if the entire confidence interval lay within the predefined bounds. To test H1 and H2 (directional improvements in perceived expressiveness and behavioral appeal), one-tailed Wilcoxon signed-rank tests were conducted due to violations of normality in paired difference scores (Shapiro–Wilk tests, $p < 0.05$). For each

comparison, we report the Wilcoxon test statistic (W), standardized test statistic (Z), effect size (r), and bias-corrected and accelerated (BCa) 95% bootstrap confidence intervals for the median differences. To test H3, we applied the same bootstrap-based equivalence testing procedure as H0. All statistical analyses were conducted using SPSS.

4.3 Results

This section reports the results of the video-based study on robot behavioral expressiveness, addressing H0 through H3. Participants' perceptions were evaluated using four questionnaire measures: understanding confidence, communication clarity, perceived behavioral expressiveness, and behavioral appeal. Descriptive statistics and inferential results for all comparisons are presented in the following subsections.

Questionnaire reliability and sampling adequacy: As a supportive check of the questionnaire's internal coherence and sampling adequacy, reliability and sampling adequacy indices were computed separately for each interaction scenario and are summarized in Table 1. Across all scenarios, KMO values exceeded 0.85 and Bartlett's tests were significant ($p < 0.05$), indicating adequate sampling adequacy and stable response patterns within each interaction context. Cronbach's α values were above 0.75 for all scenarios, suggesting satisfactory internal consistency of the questionnaire items. These indices serve as supportive checks to ensure that the questionnaire functioned consistently across interaction scenarios of varying complexity and do not constitute evidence for a latent factor structure.

Table 1. Supportive reliability and sampling adequacy indices computed separately for each interaction scenario.

Parameters	Greeting	Basketball	Animal mimicry	Drums	Drama	TED	Story
KMO	0.963	0.923	0.932	0.876	0.872	0.912	0.921
p	0.011	0.008	0.003	0.019	0.024	0.020	0.018
α	0.891	0.853	0.881	0.832	0.794	0.833	0.843

Results for H0: The H0 examined whether participants' understanding confidence (Q1) and communication clarity (Q2) ratings for the GMVF condition were statistically equivalent to those for the VF, MF, and DMVF conditions. Fig. 8 presents the mean and standard deviation of Q1 and Q2 ratings across feedback modalities.

Overall, participants reported high levels of understanding confidence and communication clarity across all four feedback conditions (mean ratings > 5 on a 7-point scale). Bootstrap-based equivalence tests indicated that the GMVF condition was statistically equivalent to the DMVF condition for communication clarity, with a mean difference of 0.005 and a 90% bootstrap confidence interval of $[-0.016, 0.026]$, which lay entirely within the predefined equivalence bounds of ± 0.50. Equivalence was likewise observed between GMVF and the VF and MF conditions for both Q1 and Q2. These results support H0 and establish that differences observed in subsequent analyses of expressiveness and appeal are not attributable to disparities in behavioral comprehensibility or clarity.

Results for H1: The H1 examined whether the GMVF condition led to higher perceived expressiveness and behavioral appeal than the voice-only feedback (VF) condition.

For perceived expressiveness (Q3), ratings in the GMVF condition (median = 5.71) were significantly higher than those in the VF condition (median = 5.57), $W = 378$, $Z = 4.54$, $p < 0.01$, with a large effect size ($r = 0.87$). The 95% BCa bootstrap confidence interval for the median difference was $[0.224, 0.492]$. Scenario-level analyses (Figure 9) indicated that this effect was observed in

six of the seven interaction scenarios, with the greeting scenario constituting the sole exception. For behavioral appeal (Q4), ratings were likewise higher in the GMVF condition (median = 5.57) than in the VF condition (median = 4.68), $W = 378$, $Z = 4.54$, $p < 0.01$, $r = 0.87$, with a 95% BCa bootstrap confidence interval of $[0.571, 0.857]$. This pattern was consistently observed across all seven interaction scenarios. Together, these results support H1.

Results for H2: The H2 examined whether the GMVF condition led to higher perceived expressiveness and behavioral appeal than the motion-only feedback (MF) condition. Analyses followed the same procedure as for H1. Comparisons between the GMVF and MF conditions revealed significantly higher ratings for GMVF on both perceived expressiveness (Q3) and behavioral appeal (Q4), with large effect sizes. Scenario-level analyses showed that these differences were consistently observed across interaction scenarios of varying complexity. These results support H2.

Results for H3: The H3 examined whether the perceived expressiveness and behavioral appeal of GMVF were statistically equivalent to those of the DMVF condition. Overall, participants reported high and similar rating for robot's behavioral expressiveness and appeal in both GMVF and DMVF conditions. Bootstrap-based equivalence tests indicated that the GMVF condition was statistically equivalent to the DMVF condition for behavioral expressiveness, with a mean difference of 0.212 and a 90% bootstrap confidence interval of $[0.159, 0.270]$, which lay entirely within the predefined equivalence bounds of ± 0.50. Equivalence was likewise observed between GMVF and the DMVF conditions for results of robot's behavioral appeal. These results support H3 and suggest that the behavior generated by our method (GMVF) can achieve the same performance as expert-designed one (DMVF).

Scenario-level patterns: As a descriptive trend, ratings of perceived expressiveness tended to decrease as interaction scenarios increased in behavioral complexity, ranging from simple social gestures to more elaborate dramatic behaviors. Across scenarios, multimodal feedback conditions (GMVF and DMVF) consistently received higher expressiveness and appeal ratings than unimodal conditions (VF and MF). In addition, voice-only feedback was generally rated as more expressive and appealing than motion-only feedback in more complex interaction contexts.

5 Study 2: Interaction Study on User Engagement

In contrast to the video-based evaluation in Study 1, this study investigates user engagement in a live, real-world HRI setting. From a behavioral and psychological perspective, user engagement reflects sustained attention, motivation, and willingness to continue interaction, all of which are strongly influenced by non-verbal social cues. Building on these insights, this study examines whether increasing the expressiveness of robot feedback behaviors enhances user engagement during live interaction.

5.1 Study Objective and Hypotheses

The objective of this interaction study is to examine whether expressive robot behaviors can enhance user engagement, measured through both observable interaction behavior and perceived engagement experience. Participants completed identical interaction tasks under multiple robot feedback conditions, allowing differences in engagement to be attributed specifically to the expressiveness of the robot's non-verbal feedback rather than task content or verbal responses. To test this objective, we propose two directional hypotheses corresponding to a motion-gain effect and an expressiveness-enhancement effect. These hypotheses are evaluated using paired significance tests, i.e., paired t-tests when normality holds; otherwise Wilcoxon signed-rank tests.

- **H4 (motion-gain effect)**: The fixed-motion and voice feedback (FMVF) condition elicits higher user engagement than the voice-only feedback (VF) condition.

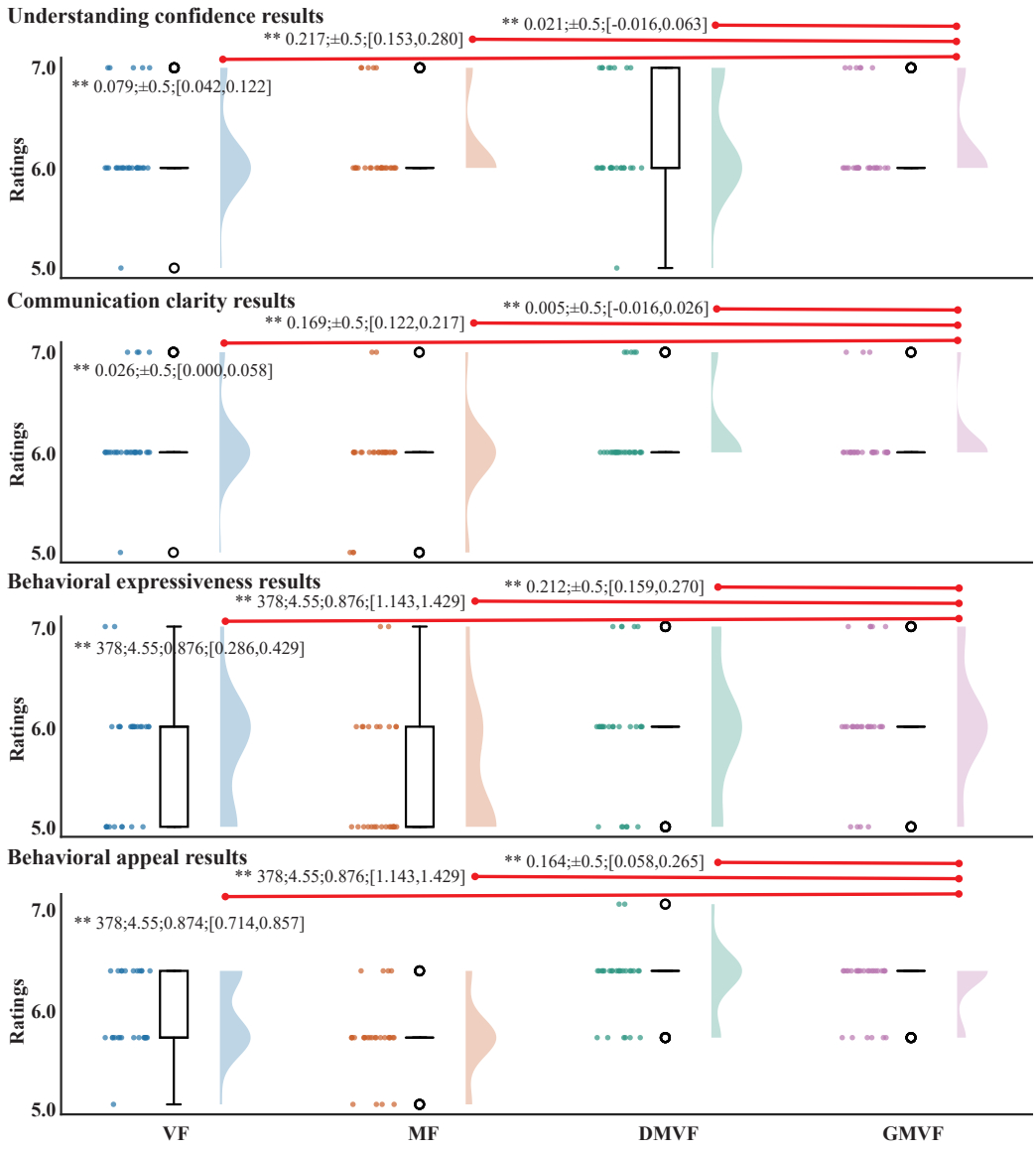

Fig. 8. The average questionnaire scores, i.e., understanding confidence, communication clarity, behavioral expressiveness, and appeal, for four distinct response patterns of humanoid robots across various conversational scenarios. The two terminals of the line "—" indicate the paired-situation of significance testing ($*$ indicates $p \leq 0.05$ and $**$ indicates $p \leq 0.01$). In this setup, a bootstrap-based equivalence test is used for testing of understanding confidence and communication. The data in the left terminal of line "—" is grouped with form "mean difference;equivalence margin;90% Bootstrap Confidence Interval". Generally, p-values correspond to bootstrap-based equivalence tests indicating statistical equivalence, not the significance of differences. A one-tailed Wilcoxon signed-rank test is used for testing of behavioral expressiveness and appeal. The data in the left terminal of line "—" is grouped with form "Wilcoxon signed-rank statistic; standardized test statistic; effect size ;95% BCa CI". The equivalence bounds used in the equivalence analyses were ±0.5 on the 7-point scale. Specifically, the testing between GMVF and DMVF in behavioral expressiveness and appeal is also processed by bootstrap-based equivalence test.

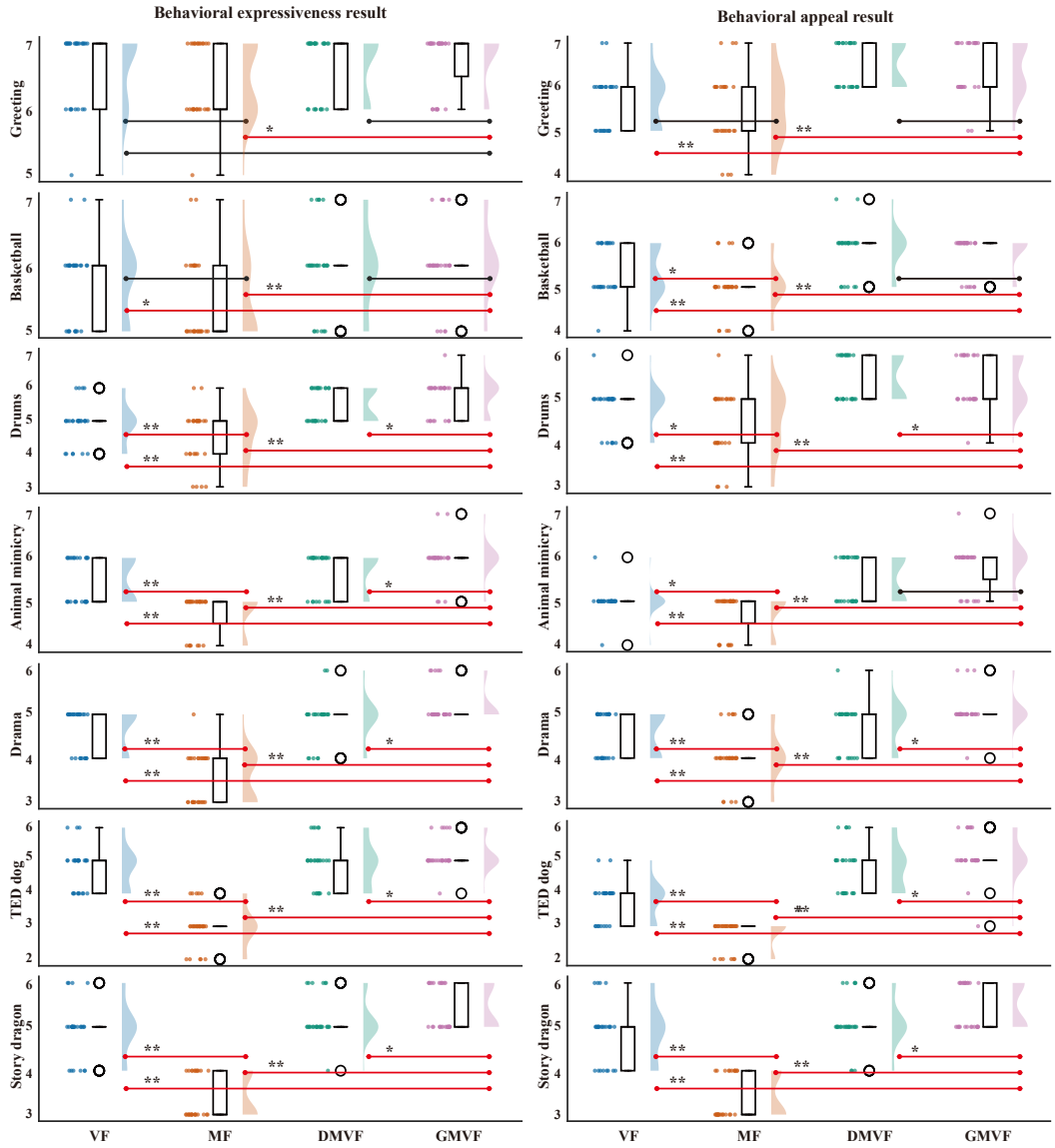

Fig. 9. Questionnaire scores, i.e., expressiveness and behavioral appeal, for four distinct response patterns of humanoid robots in seven conversational scenarios. The error bars represent standard deviation of participant ratings. The two terminals of the line "—" indicate the significance testing ($*$ indicates $p \leq 0.05$ and $**$ indicates $p \leq 0.01$). In this setup, the black line "—" means the difference of paired-situation is not significant, while the red line "—" indicates that the difference of paired-situation is significant.

- **H5 (expressiveness-enhancement effect)**: The generated-motion and voice feedback (GMVF) condition elicits higher user engagement than both comparison conditions (FMVF and VF).

5.2 Study Methodology

To rigorously evaluate the effect of expressive robot feedback on user engagement, we conducted a live, real-world HRI study. In contrast to the video-based evaluation in Study 1, this study focuses on user engagement by allowing participants to interact freely with a physical humanoid robot. Specifically, participants followed a semi-structured topic prompt but were free to ask follow-up questions, extend the conversation, and end the session at any time. Such a live interaction setting enables the capture of both behavioral engagement indicators (e.g., interaction duration and turn-taking behavior) and perceived engagement experiences, which cannot be reliably assessed through passive observation alone. Importantly, this study employed a within-subjects design with one three-level independent variable, namely the robot feedback modality during interaction. The three feedback conditions were voice-only feedback (VF), fixed-motion and voice feedback (FMVF), and generated-motion and voice feedback (GMVF).

Across conditions, the interaction task, language generation mechanism, and overall interaction structure were held constant. As a result, any observed differences in user engagement can be attributed to variations in the expressiveness of the robot's non-verbal feedback rather than differences in task content or verbal responses. The dependent variables consisted of both objective behavioral engagement measures and subjective self-reported engagement ratings. The following subsections describe the experimental design, interaction task, procedure, participants, measures, and data analysis strategy in detail.

Experimental design: A within-subjects experimental design was adopted, in which all participants experienced three robot feedback conditions while performing a semi-structured question–answering task. The interaction context (including task content and question sources), the language generation mechanism of robot feedback, and the overall task format were kept consistent across conditions. The only systematic manipulation was the expressiveness of the robot's feedback modalities. The three feedback modalities were defined as follows:

- **Voice-only Feedback (VF)**: The robot responded using speech generated by a large language model, without any accompanying gestures or non-verbal behaviors.
- **Fixed-Motion and Voice Feedback (FMVF)**: In this condition, the robot produced spoken responses accompanied by gestures selected from a predefined template-based gesture library. This library contains a set of semantically meaningful gesture primitives commonly used in humanoid robot animation systems, such as waving, nodding, head shaking, pointing, shrugging, thinking poses, and presentation, style beat gestures—analogous to the classic gesture templates in Aldebaran's Animated Speech. Specifically, the contextually appropriate gestures are selected by LLM from this predefined library based on the dialog content.
- **Generated-Motion and Voice Feedback (GMVF)**: The robot produced both speech and semantically aligned gestures, which are generated by our framework.

Across all conditions, voice feedback was generated using GPT-4 with a fixed temperature of 0.8. DMVF was not included as an additional comparison condition. DMVF requires predefined mappings between anticipated user inputs and corresponding tailored motion sequences, which is impractical under real-time constraints with unpredictable user inputs. Given these constraints, FMVF serves as a more feasible baseline for real-world deployment and is consistent with established multimodal HRI practices.

Interaction situation: The interaction followed a dialogue-based scenario consisting of multiple rounds of question-and-answer exchanges between the user and the robot. To encourage participants to initiate interaction and produce natural language input, a semi-structured question–answering

task was designed. Participants were provided with a list[2] of 20 suggested everyday topics (e.g., "Could you tell me about the basic rules of basketball?" and "What are good vacation destinations for this season?"). Each participant selected at least three topics of personal interest. For each selected topic, participants engaged in an independent interaction session with the robot. They initiated the interaction with the selected topic question and were free to ask follow-up questions within the same topic. Participants were encouraged to extend the conversation based on the robot's responses. The number of turns and interaction duration were not constrained, and each session ended when the participant voluntarily chose to stop.

Study procedure: Before the study, all participants provided informed consent. Each participant completed multiple interaction sessions and experienced all three feedback conditions (VF, FMVF, and GMVF) at least once. To mitigate potential order effects, feedback conditions were assigned to interaction topics in a randomized manner for each participant, such that the order in which conditions were encountered varied across participants. Each topic could appear at most once within a participant (i.e., topics were non-repeating within-participant), although the same topics could be reused across different participants. The order of interaction sessions (topics) is self-selected by each participant, i.e., participants chose which topic to interact with next. All interactions were conducted in Chinese and recorded for subsequent analysis.

After finishing each interaction, we assessed participants' engagement using both objective and subjective measures. Objective engagement was assessed using average engagement time (AET) and average interaction turns (AIT). Perceived engagement was measured using a short questionnaire consisting of two 7-point items, which formed the average perceived engagement (APE) score. For each participant and condition, AET, AIT, and APE were computed by averaging across that participant's sessions assigned to the condition.

- Objective metrics
  - **Average Engagement Time (AET)**: The mean duration (in minutes) from the user's first question to the moment the user voluntarily ended the interaction session.
  - **Average Interaction Turns (AIT)**: The number of user-initiated question turns during an interaction session, reflecting interaction tempo and user initiative.
- Subjective metrics
  - **APE#1**: "Did you feel highly engaged throughout the preceding interaction?"
  - **APE#2**: "If possible, would you like to interact with this robot again?"

Participants: The same 27 participants as in Study 1 took part in this interaction study. All participants provided informed consent prior to participation.

Data analysis strategy: Directional hypotheses were tested using one-tailed paired statistical procedures for all engagement indicators (AET, AIT, and APE). Specifically, average engagement time (AET) was analyzed using one-tailed paired t-tests, as the normality assumption was satisfied. Average interaction turns (AIT) and average perceived engagement (APE) violated normality assumptions and were therefore analyzed using one-tailed Wilcoxon signed-rank tests.

5.3 Results

This section reports the results of the interaction study on user engagement, addressing H4 and H5. User engagement was evaluated using average engagement time (AET), average interaction turns (AIT), and appeal of engagement (APE). Descriptive statistics and effect sizes for all comparisons are reported in Table 2.

Results for H4: The H4 examines the motion-gain effect, i.e., whether the adding non-verbal motion (fixed motion feedback, FMVF) increase engagement relative to voice-only feedback (VF).

[2]The list is provided on our project website.

Table 2. Descriptive statistics and effect sizes for engagement metrics in the interaction study.

Hypothesis	Conditions	AET				AIT			
		Mean	SD	p	Efect size	Mean	SD	p	Efect size
H4	VF	3.982	0.866	<0.01	0.801	3.815	0.879	<0.05	0.554
	FMVF	4.656	1.081			4.185	0.834		
H5	VF	3.982	0.866	<0.01	0.859	3.815	0.879	<0.01	0.839
	GMVF	4.820	1.069			4.778	1.013		
H5	FMVF	4.656	1.081	<0.05	0.402	4.185	0.834	<0.01	0.882
	GMVF	4.820	1.069			4.778	1.013		
Hypothesis	Conditions	APE#1				APE#2			
		Mean	SD	p	Efect size	Mean	SD	p	Efect size
H4	VF	4.370	0.629	<0.05	0.550	4.222	0.641	<0.01	0.675
	FMVF	4.741	0.764			4.704	0.724		
H5	VF	4.370	0.629	<0.01	0.686	4.222	0.641	<0.01	0.925
	GMVF	4.963	0.649			5.037	0.518		
H5	FMVF	4.741	0.764	<0.05	0.564	4.704	0.724	<0.01	0.724
	GMVF	4.963	0.649			5.037	0.518		

Results showed that FMVF significantly improved user engagement across all metrics. Participants spent more time interacting with the robot under FMVF than VF (AET: M = 4.656, SD = 1.081 vs. M = 3.982, SD = 0.866; one-tailed paired t-test, $p < 0.01$). They also completed more interaction turns (AIT: M = 4.185, SD = 0.834 vs. M = 3.815, SD = 0.879; one-tailed Wilcoxon signed-rank test, $p < 0.05$). Similarly, perceived engagement ratings were higher for FMVF than VF on both APE measures (APE#1: $p < 0.05$; APE#2: $p < 0.01$; one-tailed Wilcoxon signed-rank tests). These findings support H4, indicating that the addition of fixed non-verbal motion enhances both objective and perceived engagement.

Results for H5: The H5 examines the expressiveness-enhancement effect), i.e., whether the semantically aligned generated motion (GMVF) further enhance engagement relative to fixed-motion feedback (FMVF) and voice-only feedback (VF). Compared to voice-only feedback, GMVF resulted in significantly higher engagement across all measures. Participants showed longer engagement time (AET: M = 4.820, SD = 1.069 vs. M = 3.982, SD = 0.866; one-tailed paired t-test, $p < 0.01$), more interaction turns (AIT: M = 4.778, SD = 1.013 vs. M = 3.815, SD = 0.879; one-tailed Wilcoxon signed-rank test, $p < 0.01$), and higher APE ratings on both scales (all $p < 0.01$). Importantly, GMVF also outperformed fixed motion feedback. AET was significantly higher under GMVF than FMVF (AET: $p < 0.05$), as were interaction turns (AIT: $p < 0.01$). Perceived engagement ratings likewise favored GMVF over FMVF on both APE measures (APE#1: $p < 0.05$; APE#2: $p < 0.01$; one-tailed Wilcoxon signed-rank tests). Together, these results support H5 and demonstrate that semantically generated expressive motion provides additional engagement benefits beyond both voice-only and fixed-motion feedback.

6 Generative Model Evaluation

Our approach addresses the pressing need within the motion generation domains to validate various motion generation models on a life-sized humanoid robot platform, thus solving the difficult problem of integrating generated actions with the physical environment. To demonstrate the utility of our framework, we implemented several generation models on the humanoid robot GR1 with our framework and evaluated their performance. These models can be categorized into two groups:

instrumental-demonstrative motion generation models and discourse-oriented co-speech gesture generation models.

Motion model selection: To ensure the fairness of our evaluation, we selected behavior generation methods that are open-source and accompanied by an official pre-trained model. For the motion generation model, we have meticulously chosen five models, namely TM2T [11], MotionDiffuse [45], MLD [3], MoMask [10], and MotionLCM [5]. In the domain of gesture generation, we have similarly selected five models, including Gesticulator [18], Qpgesture [42], DiffuseStyleGesture [41], UnifiedGesture [40], and SemanticGesticulator [46].

All models were executed on a computer equipped with an NVIDIA GeForce RTX 4090 GPU. We did not perform any additional training or fine-tuning beyond the officially released pre-trained models. To ensure consistency, all models in the same motion category were tested using the same human input dialogue, guaranteeing identical input text for each generative model. In this section, we successfully implemented these selected works within our system and perform objective and subjective evaluations.

Objective evaluation metrics: In this section, we will evaluate successful implementations with both quantitative and qualitative metrics. We employ a straightforward and effective objective metric, the *percentage of matched beats* (PMB) [2], to assess the rhythmic performance of motion synthesis. In PMB, a motion beat is considered matched if the distance to a nearby audio beat is within the range of δ (Eq. 9). Since PMB primarily evaluates whether the temporal beats of body motion and speech feel natural and rhythmically coherent, higher PMB scores indicate more lifelike and expressive robot behaviors. This alignment of robot feedback behaviors enhances the audience's emotional resonance and is consistent with human cognitive habits. Therefore, introducing PMB into our evaluation framework facilitates a more comprehensive understanding of cross-modal coordination and contributes to evaluating the expressiveness of generative models. As supported by prior work in gesture synthesis and motion generation [2, 18, 35], the greater motion diversity and rhythmic alignment lead to higher perceived naturalness. Consequently, we define the richness of the motion (RM) as a metric (see Eq. 10) to evaluate the diversity and naturalness of the robot behaviors in response to specific dialogue input.

$$\mathrm{PMB}(B^m, B^a) = \frac{1}{N_m} \sum_{i=1}^{N_m} \sum_{j=b_*^a+1}^{N_a} \mathbb{1}[\| b_j^m - b_j^a \|_1 < \delta] \tag{9}$$

$$\mathrm{RM}(B^m, T_i) = B^m / T_i \tag{10}$$

Where B^m and B^a represent the motion beats implemented by the robot and audio beats, respectively, while N_m and N_a denote the number of beats. Motion beats are identified by detecting the local minima of joint deceleration. b_*^a indicates the last matched audio beat. The threshold δ is initially set to 0.2 seconds. This approach provides a more comprehensive understanding of the rhythmic performance.

We introduce two temporal metrics, T_g and T_i, to rigorously assess the timeliness and smoothness of the performance exhibited by the ersity and rhythmic alignmselected models within our framework. T_g represents the generation time required for the selected model to produce robot actions based on input dialogue, which should be minimized to ensure that the robot can respond to human input effectively and quickly. T_i represents the time taken by the humanoid robot to perform the generated action. Generally, our goal is for T_i to be sufficiently small to achieve synchronization of the interaction between the robot and end-users.

Subjective evaluation metrics: To evaluate the visual effects of humanoid robots executing gestures (motions) generated by various models, we conducted a user study on the robot behaviors produced by each comparative method. For each method, we selected 4 short video segments of

robot responses for evaluation, each varying in length from approximately one to four minutes and corresponding to a specific user input dialogue.

The experiment involved 27 participants as discussed in section 4. For the human-likeness evaluation, each assessment page with one video segment asked participants, "How human-like does the robot behavior appear?" In the appropriateness evaluation, each assessment page asked, "How appropriate does the robot behavior appear?" Each page displayed five options, rated on a scale from 5 to 1, with 1-point intervals, and labeled as "excellent," "good," "fair," "poor," and "very poor" (from best to worst). We used the Mean Opinion Score (MOS) to reflect the human-likeness and appropriateness of robot responses under different models in the same category. To mitigate any prior expectation biases associated with different motion types, we evaluate the performance of different models within the same motion category, not across categories. At the start of every video, participants received the same spoken prompt and situational description, ensuring that all motions were judged under identical contextual framing. Within each category, the order in which model videos were presented was randomized for each participant, eliminating order and fatigue effects. Moreover, the rating criteria remained neutral, i.e., participants evaluated each clip solely on human-likeness and contextual appropriateness, with no indication of motion category. Consequently, their judgments focused entirely on model performance. As a result, any observed MOS differences can be attributed to variations in model quality rather than to pre-existing expectations about motion type.

Table 3. Evaluation Metrics for Motion Generation Model

Model	PMB	RM	$T_g/[s]$	$T_i/[s]$	S
TM2T	0.224	0.487	11.231	124	4.1
MotionDiffuse	0.361	0.699	9.194	186	4.3
MLD	0.312	0.518	7.461	137	3.8
MoMask	0.386	0.708	13.183	226	4.4
MotionLCM	0.379	0.682	7.754	132	3.9

Result: The experimental results confirm the successful deployment of advanced humanoid motion generation models on the GR1 robot within our expressive behavior generation framework. Specifically, Fig. 10 illustrates the process of retargeting the results of the selected models to robot actions. In the interaction case, the human dialogue input was *"Can you play percussion for me, such as play drums?"* The GR1's humanoid cognition module initiated reasoning and subsequently generated the robot voice response *"Sure, here we go! Boom, Boom, Tsh, Boom, Boom, Tsh!"* along with the action response *"Position hands as if holding drumsticks. Mimic common drumming movement."* The description of the action response h_a was then sent to the generation module (MoMask for Fig. 10) and processed by the shape-adapted module. The actions X in Fig. 10 (first row) are adapted to the standing gait of GR1. Subsequently, the Cartesian poses X were mapped to the robot joints J, and the mapping results are shown in Fig. 10 (second row). Upon completion of the processing of h_a and h_v, the GR1 robot was able to fulfill the human request through voice and actions, as depicted in Fig. 10 (third row).

The results of the evaluation metrics are shown in Table 3 and Table 4. Furthermore, some of the GR1 behaviors of selected models are also shown in Fig. 11. It can be observed that gesture generation models perform better than motion generation models in PMB metrics. The reason is that the gesture generation model takes into account the temporal alignment between gesture actions and language semantics during its construction. As most of the selected models exhibit an average processing time T_g of more than 5 seconds, this prolonged latency during HRI results

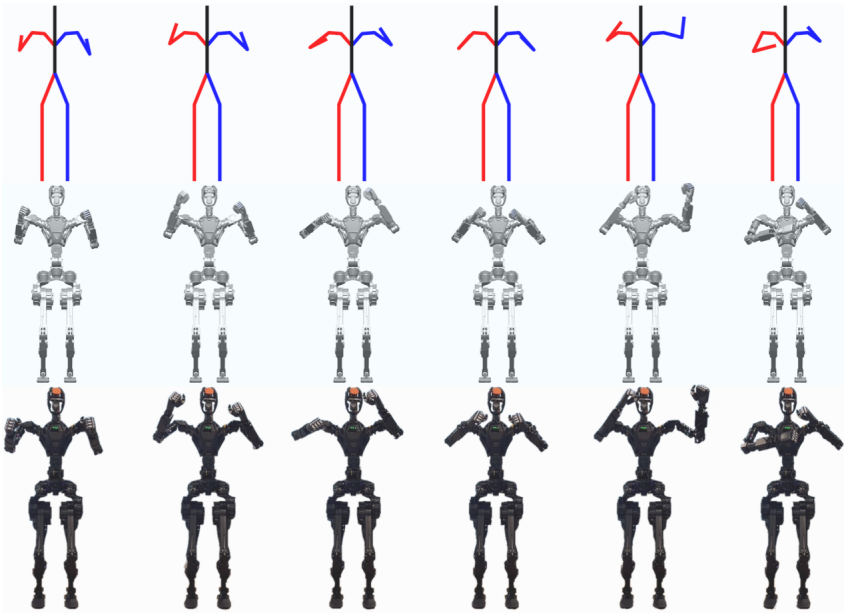

Fig. 10. Motion of the GR1 robot during the experiment. The first line depicts the human motion generated by the MoMask model, modified by our Body shape-adapting Module. The second line displays the motion mapping results from human body Cartesian space to robot joint space processed by the motion-retargeting Module. The last line indicates the reactions of the GR1 robot during the interaction process.

Table 4. Evaluation Metrics for Gesture Generation Model

Model	PMB	RM	$T_g/[s]$	$T_i/[s]$	S
Gesticulator	0.710	0.576	6.325	143	4.2
QPGesture	0.742	0.785	21.286	221	4.3
DiffuseStyleGesture	0.797	0.726	12.284	129	4.3
UnifiedGesture	0.852	0.897	10.302	108	4.6
SemanticGesticulator	0.875	0.632	7.337	139	4.1

in a poor user experience. We need to develop models with faster response speed. T_i denotes the duration required for GR1 to perform expressive actions, depending on the number of actions and the difference between consecutive actions. As the number of actions and the disparity between actions increase, the value of T_i also increases; conversely, it decreases. As *RM* is intended to assess the diversity and naturalness of robot behaviors, we interpret higher values of *RM* as indicative of greater diversity in robot actions, and consequently greater expressiveness. The value of *RM* is predominantly influenced by the model, which underscores the importance of incorporating a larger number of motion beats in the development of a humanoid behavior generation model to ensure the expressiveness of the generated behavior. S is MOS of subjective metrics which show the end-user's attitudes and evaluations of behaviors demonstrated by GR1. Upon analyzing the data S, we observed that users exhibit greater tolerance towards on-referential actions, i.e., gestures. These actions are intended to facilitate and attract the user's cognitive understanding of the robot's speech feedback rather than convey a specific meaning. As a result of the low expectations associated with

these types of actions, end-users tend to assign higher scores to them. Therefore, it is imperative to strategize on enhancing the metrics value of *PMB* and *RM*, while reducing the metrics value of T_g and T_i when developing a new human behavior generation model. This approach will lead to favorable results in the metric S, which means that the robot is executing expressive behavior.

In summary, our framework not only provides reliable algorithmic support for deploying expressive motion produced by generation models on physical robots, but also facilitates data-driven optimization of motion generation model design.

Fig. 11. The robot demos the motion response based on different humanoid motion generation models in our framework. The results of generation models with physical meaning are shown in the first row, and the input dialogue is *"please show me how to play drums."* The results of generation models without physical meaning are shown in the second row, and the input dialogue is *"please give me a brief story of Snow White."*

7 Discussion

This work investigates whether LLM-driven expressive behaviors can enhance HRI beyond commonly used unimodal feedback, while maintaining communicative comprehensibility and achieving perceptual quality comparable to expert-designed strategies. We further examine whether such behaviors can improve user engagement during interaction. Specifically, we evaluate users' subjective perceptions and observable engagement behaviors through a controlled perceptual study (Study 1) and a live interaction study (Study 2). These findings help explain how semantically grounded multimodal behavior supports interpretation and enhances engagement in HRI.

Comprehensibility as a prerequisite for expressive behaviors (H0): A central concern in expressive HRI is whether increased expressiveness compromises communicative clarity. Addressing H0, Study 1 shows that LLM-driven expressive behaviors did not reduce perceived understanding, interpretation confidence, or information clarity compared to baselines. This result indicates that

expressive enhancements can be introduced without degrading communicative effectiveness when they remain aligned with verbal content. Rather than constituting a trade-off, these findings suggest that comprehensibility can serve as a prerequisite for expressiveness within our studied setting. This foundation is critical for interpreting subsequent benefits in expressiveness and engagement, as gains achieved at the expense of understanding would be of limited practical value.

Effects of multimodal expressiveness on user perception (H1, H2): Study 1 further addressed H1 and H2, showing that LLM-driven multimodal expressive behaviors significantly improved perceived expressiveness and appeal. These benefits appear to arise from coordinated multimodal behaviors rather than isolated motion or vocal variations, highlighting the importance of semantic and contextual alignment across modalities. Together with the H0, these results demonstrate that expressive behaviors can enhance user experience while preserving clarity, supporting their role as an integral component of effective human–robot communication.

Generative versus expert-designed expressive behaviors (H3): Beyond perceptual improvements, H3 examined whether generative expressive behaviors can approximate expert-designed ones. Study 1 found that generative multimodal feedback was statistically equivalent to expert-designed behaviors within predefined equivalence bounds on key perceptual dimensions. This suggests that generative approaches can achieve comparable perceived expressiveness to expert-crafted designs. This equivalence does not diminish the value of expert design but instead positions generative expressiveness as a complementary and scalable alternative, particularly in contexts where manual authoring is impractical.

User engagement in open-ended interaction (H4, H5): Extending beyond controlled evaluations, Study 2 addressed H4 and H5 by examining engagement during open-ended interactions. LLM-driven expressive behaviors led to increased behavioral engagement, reflected in longer interaction durations and more frequent turn-taking. These findings provide behavioral evidence that expressive behaviors influence not only user perception but also sustained participation over time.

Context sensitivity and boundary conditions: Despite their overall benefits, expressive behaviors exhibited context sensitivity. As interaction complexity increased, perceived expressiveness tended to decline, indicating that expressiveness should be adaptively modulated rather than uniformly maximized. These boundary conditions are essential for interpreting engagement effects and for guiding the deployment of expressive behaviors in task-oriented scenarios.

Design implications: We evaluate the usage of our proposed framework for various behavior generation models in Section 6. The results suggest that expressiveness should be treated as an integrated aspect of communication, grounded in coordinated multimodal behaviors. Adaptive mechanisms that adjust expressiveness to interaction demands are likely necessary, and generative approaches should be paired with appropriate constraints to ensure interpretability, feasibility, and safe deployment.

Limitations and future work: Despite the effectiveness and promise of the proposed approach, several limitations remain, primarily stemming from the capabilities of the current system platform. First, the initial deployment of our framework requires a relatively long loading time, which affects overall deployment efficiency and limits its practicality in time-sensitive scenarios. Second, to ensure motion stability and safety, the lower-body movements of the humanoid robot were simplified in the current implementation and restricted to three fixed gait patterns. While this design choice supports reliable system operation, it also constrains the richness of whole-body expressiveness, resulting in an experimental focus on upper-body expressive behaviors. In addition, although strong synchronization between speech and actions can be achieved in animation and simulation environments, action responses lag behind speech responses in real-world robot experiments. We employed a content alignment algorithm to improve action–speech coordination; however, its

effectiveness remains limited, indicating that achieving both high responsiveness and stable motion in dynamic interaction settings continues to be a challenge.

Future work will build directly on these limitations by advancing the proposed framework toward humanoid robots with higher dynamic capabilities. On the one hand, we will focus on improving real-time performance by reducing system loading and response latency. On the other hand, we plan to extend the framework to support full-body expressive behaviors, enabling more immersive and physically grounded human–robot interactions. As a concrete direction, we aim to integrate lower-body motion generation through the development of stable and safe dynamic gait control strategies based on model predictive control (MPC) and reinforcement learning (RL). Owing to the modular design of our framework, we expect these extensions to be achieved with minimal architectural changes, facilitating the deployment of LLM-driven expressive behaviors in highly dynamic, real-world human–robot interaction systems.

8 Conclusion

This work investigated the effectiveness of LLM-driven expressive behaviors that integrate generated motion and voice feedback for enhancing human–robot interaction. Through a controlled video-based study and a live interaction study, we showed that semantically aligned multimodal expressiveness improves users' perceived expressiveness, appeal, and engagement without compromising comprehension. Across both studies, generated expressive behaviors, when paired with generated speech, achieved perceptual quality comparable to expert-designed motion. Moreover, the live interaction results demonstrate that expressive feedback significantly influences interaction dynamics, leading to longer interactions, more frequent user-initiated turns, and higher engagement than voice-only or fixed-gesture strategies. These findings suggest that expressive behaviors serve as integral components of interaction, rather than merely acting as supplementary feedback modalities. Overall, this work highlights the importance of behavioral comprehensibility, semantic grounding, and cross-modal coordination in expressive humanoid robot design. The results support a shift from fixed motion libraries toward scalable generative frameworks that integrate language understanding with embodied action. Moreover, our work demonstrates that aligning robot behaviors with human cognitive expectations and social norms through technical advancements such as LLM-driven frameworks is pivotal for enhancing their acceptability. This underscores the necessity of continued research into human-like, expressive behavior generation as a critical pathway to bridging the gap between technological potential and user acceptance for humanoid robots.

References

[1] Malin Andtfolk, Linda Nyholm, Hilde Eide, and Lisbeth Fagerström. 2022. Humanoid robots in the care of older persons: A scoping review. *Assistive Technology* 34, 5 (2022), 518–526.

[2] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. 2022. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–19.

[3] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. 2023. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18000–18010.

[4] Xuxin Cheng, Yandong Ji, Junming Chen, Ruihan Yang, Ge Yang, and Xiaolong Wang. 2024. Expressive Whole-Body Control for Humanoid Robots. *arXiv preprint arXiv:2402.16796* (2024).

[5] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. 2024. Motionlcm: Real-time controllable motion generation via latent consistency model. In *European Conference on Computer Vision*. Springer, 390–408.

[6] Claudia Di Napoli, Giovanni Ercolano, and Silvia Rossi. 2023. Personalized home-care support for the elderly: a field experience with a social robot at home. *User Modeling and User-Adapted Interaction* 33, 2 (2023), 405–440.

[7] Yan Ding, Xiaohan Zhang, Saeid Amiri, Nieqing Cao, Hao Yang, Andy Kaminski, Chad Esselink, and Shiqi Zhang. 2023. Integrating action knowledge and LLMs for task planning and situation handling in open worlds. *Autonomous Robots* 47, 8 (2023), 981–997.

[8] Vishnu Sashank Dorbala, James F Mullen Jr, and Dinesh Manocha. 2023. Can an embodied agent find your "cat-shaped mug"? llm-based zero-shot object navigation. *IEEE Robotics and Automation Letters* (2023).

[9] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.

[10] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. 2024. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1900–1910.

[11] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. 2022. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*. Springer, 580–597.

[12] Jan-Gerrit Habekost, Connor Gäde, Philipp Allgeuer, and Stefan Wermter. 2024. Inverse kinematics for neuro-robotic grasping with humanoid embodied agents. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7315–7322.

[13] Jan-Gerrit Habekost, Erik Strahl, Philipp Allgeuer, Matthias Kerzel, and Stefan Wermter. 2023. CycleIK: Neuro-inspired Inverse Kinematics. In *International Conference on Artificial Neural Networks*. Springer, 457–470.

[14] Peide Huang, Yuhan Hu, Nataliya Nechyporenko, Daehwa Kim, Walter Talbott, and Jian Zhang. 2025. Emotion: Expressive motion sequence generation for humanoid robots with in-context learning. *IEEE Robotics and Automation Letters* (2025).

[15] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. 2023. Inner Monologue: Embodied Reasoning through Planning with Language Models. In *Conference on Robot Learning*. PMLR, 1769–1782.

[16] Martin Inderbitzin, Aleksander Väljamäe, Jose Maria Blanco Calvo, Paul FMJ Verschure, and Ulysses Bernardet. 2011. Expression of emotional states during locomotion based on canonical parameters. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 809–814.

[17] Matthias Kerzel, Philipp Allgeuer, Erik Strahl, Nicolas Frick, Jan-Gerrit Habekost, Manfred Eppe, and Stefan Wermter. 2023. Nicol: A neuro-inspired collaborative semi-humanoid robot that bridges social interaction and reliable manipulation. *IEEE access* 11 (2023), 123531–123542.

[18] Taras Kucherenko, Patrik Jonell, Sanne Van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 international conference on multimodal interaction*. 242–250.

[19] Minae Kwon, Hengyuan Hu, Vivek Myers, Siddharth Karamcheti, Anca Dragan, and Dorsa Sadigh. 2024. Toward grounded commonsense reasoning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5463–5470.

[20] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European conference on computer vision*. Springer, 612–630.

[21] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

[22] Shipeng Lyu, Fangyuan Wang, Weiwei Lin, Luhao Zhu, David Navarro-Alarcon, and Guodong Guo. 2025. HuBE: Cross-Embodiment Human-like Behavior Execution for Humanoid Robots. *arXiv preprint arXiv:2508.19002* (2025).

[23] Karthik Mahadevan, Jonathan Chien, Noah Brown, Zhuo Xu, Carolina Parada, Fei Xia, Andy Zeng, Leila Takayama, and Dorsa Sadigh. 2024. Generative expressive robot behaviors using large language models. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 482–491.

[24] Teresa Onorati, Álvaro Castro-González, Javier Cruz del Valle, Paloma Díaz, and José Carlos Castillo. 2023. Creating Personalized Verbal Human-Robot Interactions Using LLM with the Robot Mini. In *International Conference on Ubiquitous Computing and Ambient Intelligence*. Springer, 148–159.

[25] Eun-A Park, Ae-Ri Jung, and Kyoung-A Lee. 2021. The humanoid robot Sil-Bot in a cognitive training program for community-dwelling elderly people with mild cognitive impairment during the COVID-19 pandemic: a randomized controlled trial. *International journal of environmental research and public health* 18, 15 (2021), 8198.

[26] Mathis Petrovich, Michael J. Black, and Gül Varol. 2023. TMR: Text-to-Motion Retrieval Using Contrastive 3D Human Motion Synthesis. In *International Conference on Computer Vision (ICCV)*.

[27] Peteris Racinskis, Janis Arents, and Modris Greitans. 2022. A motion capture and imitation learning based approach to robot control. *Applied Sciences* 12, 14 (2022), 7186.

[28] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.

[29] Roger Andre Søraa, Gunhild Tøndel, Mark W Kharas, and J Artur Serrano. 2023. What do older adults want from social robots? a qualitative research approach to human-robot interaction (HRI) studies. *International journal of social robotics* 15, 3 (2023), 411–424.

[30] Daniel Szafir, Bilge Mutlu, and Terrence Fong. 2014. Communication of intent in assistive free flyers. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. 358–365.

[31] Frank Thomas and Ollie Johnston. 1981. Disney animation: The illusion of life. *(No Title)* (1981).

[32] Lorenzo Torresani, Peggy Hackney, and Christoph Bregler. 2006. Learning to synthesize motion styles. In *Proceedings of the Snowbird Learning Workshop*. Citeseer.

[33] Rik van den Brule, Ron Dotsch, Gijsbert Bijlstra, Daniel HJ Wigboldus, and Pim Haselager. 2014. Do robot performance and behavioral style affect human trust? A multi-method approach. *International journal of social robotics* 6 (2014), 519–531.

[34] Rik van den Brule, Ron Dotsch, Gijsbert Bijlstra, Daniel HJ Wigboldus, and Pim Haselager. 2014. Do robot performance and behavioral style affect human trust? A multi-method approach. *International journal of social robotics* 6 (2014), 519–531.

[35] Gentiane Venture and Dana Kulić. 2019. Robot expressive motions: a survey of generation and evaluation methods. *ACM Transactions on Human-Robot Interaction (THRI)* 8, 4 (2019), 1–17.

[36] Benedikte Wallace, Marieke Van Otterdijk, Yuchong Zhang, Nona Rajabi, Diego Marin-Bucio, Danica Kragic, and Jim Torresen. 2024. Imitation or Innovation? Translating Features of Expressive Motion from Humans to Robots. In *Proceedings of the 12th International Conference on Human-Agent Interaction*. 296–304.

[37] Zining Wang, Paul Reisert, Eric Nichols, and Randy Gomez. 2024. Ain't Misbehavin'-Using LLMs to Generate Expressive Robot Behavior in Conversations with the Tabletop Robot Haru. In *Companion of the 2024 ACM/IEEE international conference on human-robot interaction*. 1105–1109.

[38] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. 2023. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots* 47, 8 (2023), 1087–1102.

[39] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. 2020. GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6184–6193.

[40] Sicheng Yang, Zilin Wang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Qiaochu Huang, Lei Hao, Songcen Xu, Xiaofei Wu, Changpeng Yang, et al. 2023. Unifiedgesture: A unified gesture synthesis model for multiple skeletons. In *Proceedings of the 31st ACM International Conference on Multimedia*. 1033–1044.

[41] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. 2023. DiffuseStyleGesture: stylized audio-driven co-speech gesture generation with diffusion models. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. 5860–5868.

[42] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, and Haolin Zhuang. 2023. QPGesture: Quantization-Based and Phase-Guided Motion Matching for Natural Speech-Driven Gesture Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE, 2321–2330.

[43] Takahide Yoshida, Atsushi Masumori, and Takashi Ikegami. 2025. From text to motion: grounding gpt-4 in a humanoid robot "alter3". *Frontiers in Robotics and AI* 12 (2025), 1581110.

[44] Liang Zhang, Zhihao Cheng, Yixin Gan, Guangming Zhu, Peiyi Shen, and Juan Song. 2016. Fast human whole body motion imitation algorithm for humanoid robots. *2016 IEEE International Conference on Robotics and Biomimetics (ROBIO)* 46, 6 (2016), 1430–1435.

[45] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2024. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence* 46, 6 (2024), 4115–4128.

[46] Zeyi Zhang, Tenglong Ao, Yuyao Zhang, Qingzhe Gao, Chuan Lin, Baoquan Chen, and Libin Liu. 2024. Semantic gesticulator: Semantics-aware co-speech gesture synthesis. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–17.

Received xx-xx-xxxx; revised xx-xx-xxxx; accepted xx-xx-xxxx