arXiv:2508.19002v1 [cs.RO] 26 Aug 2025

HuBE: Cross-Embodiment *Hu*man-like *B*ehavior *E*xecution for Humanoid Robots

Shipeng Lyu†, Fangyuan Wang†, Weiwei Lin, Luhao Zhu,
David Navarro-Alarcon* and Guodong Guo

***Abstract*—Achieving both behavioral similarity and appropriateness in human-like motion generation for humanoid robot remains an open challenge, further compounded by the lack of cross-embodiment adaptability. To address this problem, we propose *HuBE*, a bi-level closed-loop framework that integrates robot state, goal poses, and contextual situations to generate human-like behaviors, ensuring both behavioral similarity and appropriateness, and eliminating structural mismatches between motion generation and execution. To support this framework, we construct HPose, a context-enriched dataset featuring fine-grained situational annotations. Furthermore, we introduce a bone scaling-based data augmentation strategy that ensures millimeter-level compatibility across heterogeneous humanoid robots. Comprehensive evaluations on multiple commercial platforms demonstrate that *HuBE* significantly improves motion similarity, behavioral appropriateness, and computational efficiency over state-of-the-art baselines, establishing a solid foundation for transferable and human-like behavior execution across diverse humanoid robots.**

***Index Terms*—Humanoid robot, human-like behavior, behavioral appropriateness, pose generation.**

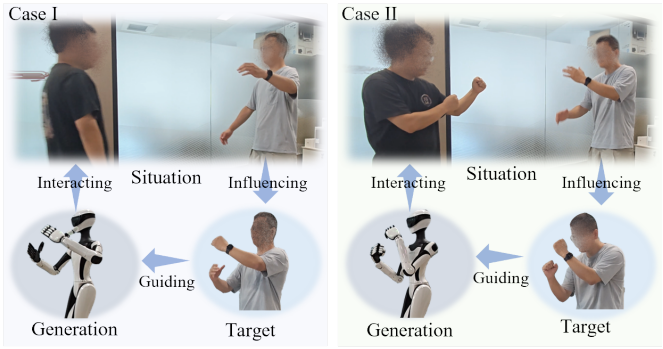

Fig. 1. **Contextual semantics dictate human-like action execution by influencing behavioral appropriateness**. Two cases to show the influence of the contextual situation for performing human-like actions. Case I ("hugging friend") requires an open-arm posture with elevated elbows to express affection, while Case II ("boxing with friend") demands guarded elbows for defensive intent, despite nearly identical end-effector positions. These full-body pose adaptations to contextual demands constitute *behavioral appropriateness*, i.e., the alignment of motions with contextual situations.

I. Introduction

Humanoid robots play an pivotal role in human-robot interaction (HRI), where the behavioral human-likeness significantly influences user perception and acceptance [1]. According to the uncanny valley theory, subtle deviations in robot behaviors that closely approximate human actions can elicit profound discomfort for humans. This phenomenon reveals a fundamental challenge when generating human-like behaviors for humanoid robots: simultaneously ensuring behavioral appropriateness while preserving motion similarity. Generally, similarity emphasizes the faithful reproduction of human kinematic and dynamic patterns, ensuring that the robot's behavior physically resembles human motion. In contrast, appropriateness emphasizes that robot behaviors must comply with situational demands and human cognitive expectations, thereby ensuring that the generated actions are contextually meaningful and socially acceptable within a given scenario. As the example shown in Fig 1, these behaviors must not only satisfy kinematic goals, which are the targets of traditional behavior planning methods for behavioral similarity, but also align with human expectations within specific scenarios [2], i.e., modifying the actions to adapt the contextual situation for achieving behavioral appropriateness.

Despite progress in behavioral similarity via various methods, e.g., motion retargeting (IK) [3]–[5] and imitation learning [6]–[8], behavioral appropriateness remains underexplored in extensive discussions. Moreover, existing humanoid motion generation methods to ensure behavior similarity suffer from several limitations. First, most of the current datasets [9] exhibit insufficient description of expressiveness of human motion, relying on 6D joint positions which lacks semantic annotations that bridge contextual situations to behavioral appropriateness. Moreover, several human demonstrated datasets just use short and simple text (e.g., 'running') to represent human motion sequence, which can not describe the entail and rich contextual situations. Second, the open-loop mechanism during pose generation and retargeting processes leads to the body structural mismatches in the two stages, which results in action human-likeness and semantics degradation. For example, [10] typically maps the generated pose to humanoids without considering the current robot state into generation module, thereby the generated pose neglecting the physical constraints of humanoids. Sequently, resulting the retargeting problem. Third, cross-embodiment adaptation remains unaddressed, impeding deployment on heterogeneous

This work is supported in part by the Research Grants Council (RGC) of Hong Kong under grant 15231023, and in part by the PolyU-EIT Collaborative PhD Training Programme under application number 220766263. *Corresponding author: David Navarro-Alarcon.*

S. Lyu, F. Wang, and D. Navarro-Alarcon are with the Department of Mechanical Engineering, The Hong Kong Polytechnic University (PolyU), Kowloon, Hong Kong. D. Navarro-Alarcon is also with the Research Institute for Smart Ageing (RISA), PolyU. (e-mail: shipeng.lyu, fangyuan.wang@connect.polyu.hk, dnavar@polyu.edu.hk)

S. Lyu, F. Wang, W. Lin, L. Zhu and G. Guo are with the Ningbo Institute of Digital Twin, Eastern Institute of Technology (EIT), China. (e-mail: gdguo@eitech.edu.cn)

† Equal contribution; * Corresponding author

robots [11] for different potential usage.

To address these challenges, we propose *HuBE*, a bi-level closed-loop framework, that generates human-like behaviors achieving both behavioral similarity and appropriateness. First, we propose the HPose dataset through open-sourced datasets, which incorporates contextual semantics of behavioral situations through fine-grained language annotations (e.g., "You hold the bottom of the cardboard box with both hands and struggle to raise it above your head.") and leverage 6D poses to preserve the behavioral human-likeness. Second, we introduce a closed-loop mechanism through implicit skeletal parameter adaptation in *HuBE*. Therefore, it can fuse multi-modalities, i.e., robot state, behavioral goal, and contextual situation, to enable the end-to-end integration of motion generation and robot control and avoid the structure mismatching problem. Furthermore, the introduction of contextual situation in robot action planning process guaranteed the behavioral appropriateness while keeping the behavioral similarity. Finally, we propose a bone scaling operation for data augmentation to simulate the morphological distribution of state-of-the-art commercial humanoids, addressing robust adaptation across heterogeneous platforms. This enables the trained action generation model to generalize across diverse robotic kinematic parameters, laying the foundation for the millimeter-level cross-embodiment compatibility that distinguishes our framework with other works. The contributions are summarized as follows:

- We propose a new perspective on human-likeness, defined by behavioral similarity and appropriateness, supported by HPose dataset with fine-grained annotations bridging motion semantics for context-aware pose generation.
- We propose a bi-level closed-loop framework that enables humanoid robots to generate actions satisfying both behavioral similarity and appropriateness, while eliminating structural mismatches between the motion generation and execution modules.
- We introduce a bone scaling-based data augmentation strategy that provides millimeter-level cross-embodiment compatibility, enabling robust deployment across heterogeneous humanoid robots without additional shape adaptation.

II. Related works

To ensure the behavioral similarity of humanoid robots, recent approaches focus on replicating human motion patterns through two paradigms: motion synthesis and imitation learning (IL). Motion synthesis methods generate human-like trajectories using auto-regressive models [12] or diffusion processes [13]–[15], but often produce physically infeasible poses requiring post-processing. Moreover, a challenge for motion synthesis works [16], [17] is that they cannot control well the added control signal for robot joints. To tackle this issue, we concentrate on creating the robot's pose on a frame-by-frame basis using the robot state and joint goal, rather than generating an entire motion sequence. IL methods [18]–[20] learn control policies directly from human demonstrations, yet face challenges in handling heterogeneous robot morphologies and novel task constraints. Furthermore, some attempts [21], [22] ensure that humanoids learn autonomous skills from egocentric vision instead of third-view vision, which is closer to human behavioral habits. However, this end-to-end learning method still suffers from the challenges of generalization performance for new tasks. Consequently, we focus on learning a pose generation policy rather than a task skill to enhance the generalization ability of our model. Additionally, we introduce bone scaling operation strategies to boost the model's adaptability across different humanoid robots.

Research on behavioral appropriateness, rooted in social psychology and anthropology, examines the regulatory mechanisms through which social norms govern individual actions [23]. In HRI, empirical studies confirm that the congruence between robotic behaviors and user cognitive expectations directly dictates social acceptance for robots [24], [25]. Although existing efforts have achieved several accomplishments in behavioral appropriateness for mobile robots, such as verification models [26], it is still a challenge for humanoid robotics to perform contextually appropriate behaviors. A critical limitation is the absence of systematic integration of contextual non-kinetic parameters (e.g., contextual situation) in recent human-like motion generation methods for humanoid robots, which severely constrains contextual appropriate motion synthesis. To resolve this, we propose a semantic-task fusion framework that synergizes situational semantics with robotic behavioral tasks, establishing a behavior generation paradigm compliant with human sociocognitive principles.

III. Methodology

To drive humanoid robots to perform expressive human-like poses under the requirement of behavioral similarity and appropriateness, we propose the *HuBE*, i.e., **hu**man-like **b**ehavior **e**xecution framework, which consists of three modules as shown in Fig. 2, i.e., data processing, behavior generation, and behavior execution. The behavior generation module takes motion situations, current robotic observations, and target poses as inputs and generates human-like actions to reach the target. The behavior execution module then maps the generated human-like pose sequence to humanoid robots based on the physical characteristics provided in the robot pose data.

A. Data Specifications

We enhanced a new dataset (HPose) in Table I for this study by open-sourced datasets, i.e., KIT [9], AMASS [27], and Motion-X [28] with three operations below:

Data Definition. We reduced the body joints as depicted in Fig. 3, identifying 11 main joints. Specifically, the position of the left/right hand is identical to that of the left/right wrist. Additionally, we introduced the rotational data that is represented by quaternions for each joint in our dataset. To enrich the contextual situation l of human motions, we use the LLM (GPT-4o) to generate this description of given motion sequence with the annotation $\hat{l}$ in the original dataset. In detail, we select two directed frames of one motion sequence with

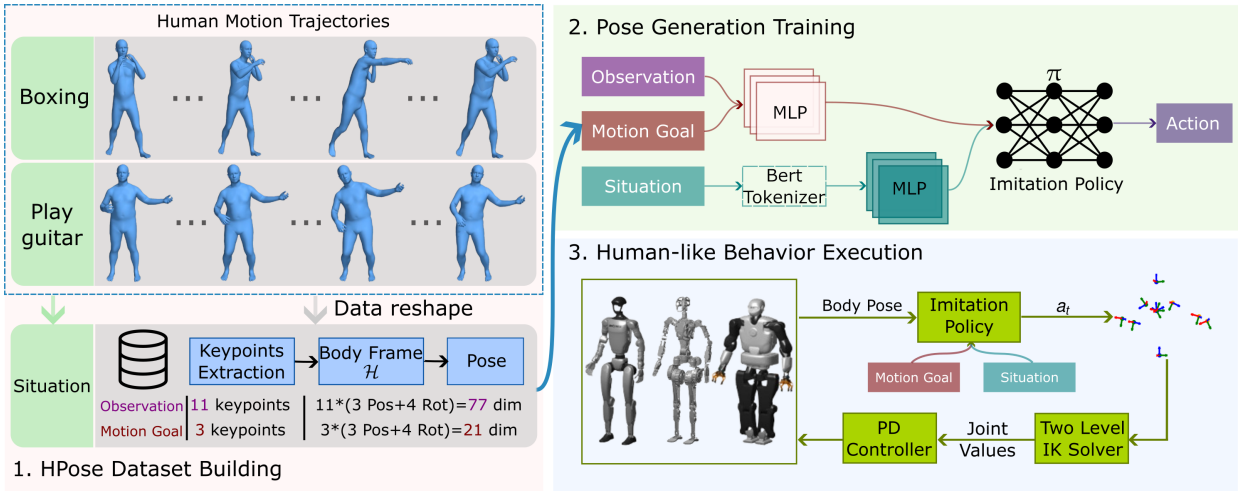

Fig. 2. Overview of the whole algorithm. This algorithm includes three parts, i.e., building dataset, model training and algorithm implementation.

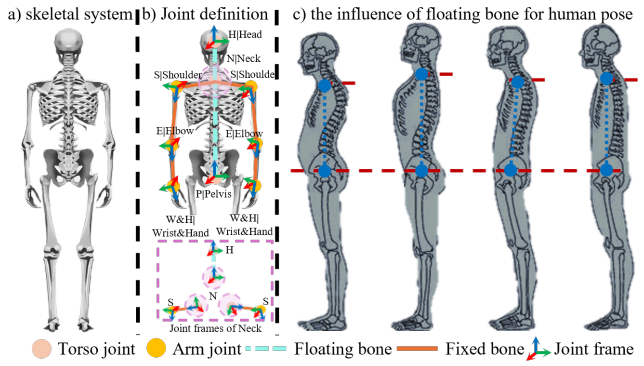

Fig. 3. The human skeletal system introduction. a) The composition of the human body's skeleton. b) Simplified definition of human upper body chain $\mathcal{H}$ for expressive human pose dataset (HPose). c) A simple example of the influence caused by floating bone.

annotation $\hat{l}$ (i.e., a person is drinking water) and put them into LLM. After reasoning the situation of these materials, the LLM generates a rich and detailed l as *"a human is drinking water while the cup is closely on the table in his left side. He is attempting to cause the cup with his right hand."*

Dataset Structure. We reformat motion data in each frame into the configuration $\{(s_i, g_i, l_i, a_i)\}$ to align with policy training needs. Initially, we extract the 11 joint poses of the current frame, defining these as the apparent state s_i; Subsequently, we incorporate the poses of the three end-effectors, i.e., both hands and the head, from the subsequent frame as the goal state g_i. Ultimately, the 11 joint poses from the subsequent frame are utilized as the action a_i to fulfill the goal g_i based on the apparent state s_i. Additionally, a situational context l_i is appended to characterize human behavior scenarios.

Data Augmentation. Due to the differences in body mechanisms between humans and humanoids, such as arm length, it is necessary to adjust the body shape of the model in the original dataset to generate poses for humanoids with various configurations. We get the augmented dataset $\mathcal{H}'$ (pseudo ground truth, **Pseudo GT**) through a bone scaling operation by updating the bone size with collected humanoid robot body parameters $R = \{r_i, r_k\}$ (same definition as Fig. 3) with Eq. 1 and Algorithm 1.

$$\mathcal{H}' = \bigcup_{(J_i, J_k) \in \mathcal{H}} \left(J'_k = J'_i + \frac{\overrightarrow{J_k - J_i}}{\|\overrightarrow{J_k - J_i}\|} \cdot \|\overrightarrow{r_k - r_i}\| \right) \quad (1)$$

Where the $\{(J_i, J_k)\}$ is the directed joint pair in the original dataset $\mathcal{H}$ (ground truth, **GT**), while the $\{(r_i, r_k)\}$ is the one in robot's body parameters $\mathcal{R}$. Furthermore, $\mathcal{R}$ are collected from 9 typical humanoid robots, such as Unitree H1 and PAL's TALOS.

Algorithm 1: Bone scaling for data augmentation

The data: Human body chain $\mathcal{H} = \{J_i\}$, robot body chain $\mathcal{R} = \{r_i\}$
The result: Augmented human body chain $\mathcal{H}'$
Generating directed joint pair set $D = \{(J_i, J_k)\}$ and joint frame set $\mathcal{H}' = \{J'_0\}$
while $D \neq \emptyset$ **do**
  Randomly select (J_i, J_k)
  if $(J_i, J_k) \in \mathcal{H}$ **then**
    $V_k = \frac{J_k - J_i}{norm(J_k - J_i)} norm(r_k - r_i)$
    if $J'_i \in \mathcal{H}'$ **then**
      $J'_k = V_k + J'_i$
      $\mathcal{H}' = \mathcal{H}' \cup J'_k$
      $D = D - \{(J_i, J_k)\}$
Return $\mathcal{H}'$

To enable effective data augmentation for cross-embodiment compatibility, we simplify the skeletal system by categorizing bones into two types, i.e., fixed and floating bones, based on their size variability during human movement. As shown in Fig. 3-c, floating bones (e.g., the spine) exhibit length changes with body poses, while fixed bones remain stable; this distinction is critical for adapting human motions to humanoid robots via bone scaling operations.

TABLE I
DATA SOURCE DISTRIBUTION OF HPOSE DATASET

Dataset	KIT	AMASS	Motion-X	HPose
Motion	3912	5600	22413	31925
Frame	2308K	12457K	47339K	62104K

B. Pose Generation

We aim to drive the humanoid robot to perform human-like poses based on a situational context l, current state s and a motion target g. We formulate the humanoid robot motion generation problem as a Markov Decision Process [29] without a specific reward function [30], which is denoted by $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{G}, l, \rho_0, \gamma)$. $\mathcal{S}$ is the state space containing all keypoint joint poses of the upper body of the robot, which implicitly reflects the body structure information, such as bone length. $\mathcal{G}$ is the goal space that includes the target poses of end effectors, i.e., both hands and head. $\mathcal{A}$ is the action space including all keypoint joint poses that the robot needs to take to reach the goal. We use ρ_0 to denote the distribution of the initial state. $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the transition function denoting the transition probability from the current state to the next state after taking an action. γ is the discount factor. The motion generator of the robot can be specified by a policy with parameters θ, $\pi_\theta : \mathcal{S}, \mathcal{G}, l \to \mathcal{A}$, which specify the action that should be taken in the robot's current state (Eq. 2). The problem is to determine an imitation policy π_θ, using a provided set of expert demonstrations $\{\xi_1, \xi_2, \cdots\}$. Therefore, we collect all the episodes in ξ into a dataset $\{(s_i, g_i, l_i, a_i)\}$, and the objective is to maximize the probability of the demonstrated action, which can be formulated as follows,

$$\theta^* = \arg\max_\theta \prod_{i=0}^{N} \pi_\theta(a_i | s_i, g_i, l_i), \tag{2}$$

To achieve this imitation policy, two classical network architectures are used, i.e., MLP and transformer.

Considering the strategy of current motion generation methods which typically replicate fixed human motion sequences to achieve target poses, ignoring contextual situation constraints, several problems remain to be solved for our policy training process. First, how to guarantee the behavioral appropriateness due to the exist of similar end-effector poses under varying situations. Second, how to accurately generate the target behaviors which is not seen in human motion sequences (dataset). To overcome these challenges, we propose **Input Alignment** and **Hindsight Training** strategies for stable human-like behaviors generation for humanoids. Furthermore, the cross-platform compatibility of our policy is guaranteed by training with the augmented data.

Inputs Alignment. To solve the first problem, we use l to bridge the generated motion to appropriateness. By utilizing the l of current motions, we first use the Bert tokenizer ($BERT(\cdot)$) to tokenize the l_i into a fixed vector, then use an MLP ($MLP(\cdot)$) to map the humanoid proprioception (s_i, g_i) to the same size as the language vector, as shown in Fig. 2-part 2 and Eq. 3.

$$v_{alig} = MLP(s_i \oplus g_i) \oplus MLP(BERT(l_i)) \tag{3}$$

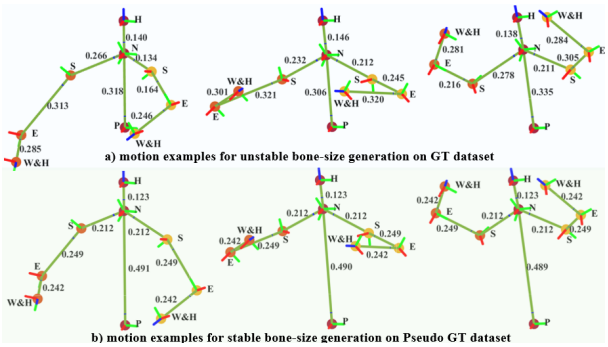

Fig. 4. The generation results for bone-size stability. The first row is the generated action based on the model trained on GT, while the second row is the results on the Pseudo GT. The input body pose is captured from GR1, while the situational context is *"Lift a heavy box over your head with both hands"*.

Hindsight Training. To generate an arbitrary motion pose for second problem, it is not efficient to directly select the next frame pose in the motion sequence as the target pose during the training. Therefore, based on the idea of hindsight experience replay buffer [31], which takes the state accomplished in the same trajectory τ as the final goal, we augment current datasets $\mathcal{D}$ by randomly taking the future poses in the frame window H as the target pose we want the robot to achieve (shown in Eq. 4).

$$\mathcal{D}_{\text{new}} = \bigcup_{\tau \in \mathcal{D}} \bigcup_{t=0}^{T} \{(s_t,\ g_{t+k},\ l_t,\ a_{t+k})\}_{k=1}^{H} \tag{4}$$

C. Behavior Execution

In this section, we aim to adopt the generated human-like actions into humanoid robots, i.e., converting actions from Cartesian space into the robot's joint space. The challenge in this module is how to keep the human-likeness during the converting process, especially involving the movement of floating bones.

To solve this problem, we propose a multi-constrained IK solver, i.e., a Closed-loop Inverse Kinematics (CLIK) algorithm based on the Pinocchio toolbox, to determine the robot joint control values corresponding to the target body pose. Specifically, a two-step solver is implemented during the IK solving process. First, we calculate the robot torso joint angles J^*_{torso}, including joints of the neck and head ($\mathcal{E}_{\text{torso}}$). In this step, we emphasize the rotational pose of the target joint because the connected bone length in the torso (see Fig. 3-floating bone) varies with different body poses. Second, we create a reduced model by locking the torso joints using the Pinocchio toolbox and then calculating the dual-arm joint angles J^*_{arms}, such as wrists ($\mathcal{E}_{\text{arms}}$). To enhance the stability of this IK solver, all the joint poses are processed by an SE(3) group filter, which is developed with Pinocchio's SE(3) interpolation algorithm.

$$\begin{cases} J^*_{\text{torso}} = \arg\min_{J_{\text{torso}}} \sum_{i \in \mathcal{E}_{\text{torso}}} \left\| f^i_{\text{FK}}(J_{\text{torso}}) - p^{\text{target}}_i \right\|^2 \\ J^*_{\text{arms}} = \arg\min_{J_{\text{arms}}} \sum_{j \in \mathcal{E}_{\text{arms}}} \left\| f^j_{\text{FK}}(J^*_{\text{torso}}, J_{\text{arms}}) - p^{\text{target}}_j \right\|^2 \end{cases} \tag{5}$$

TABLE II
BONE-SIZE STABILITY RESULTS [METER]

Dataset	Bone Type	GR1	TALOS	G1
GT	Fixed	0.18	0.14	0.23
	Floating	0.31	0.26	0.283
GT & Pseudo GT	Fixed	**1.3e-4**	**2.4e-3**	**4.4e-4**
	Floating	**4.1e-4**	**6.7e-3**	**6.8e-4**

IV. RESULTS

To validate the effectiveness of our methodology, corresponding evaluation metrics are provided. Specifically, two evaluation methods are proposed to assess the behavioral similarity.

- End-effector accuracy (E-A) is designed to measure the pose accuracy of the end-effectors, including both hands and head.
- Human similarity (H-S) aims to describe the pose consistency by considering all selected body markers.

For each evaluation method, we provide two metrics called Mean Per Joint Position Error (MPJPE) [32] and Mean Per Joint Orientation Error (MPJOE) to measure the error of position and orientation, respectively. To evaluate the behavioral appropriateness, three metrics are proposed as below:

- Fréchet Motion Distance (FMD): measure the distributional similarity between generated motions and authentic human motions in the feature space similar as FID metric [33]. This metric works based on the premise that the human demonstrated motion in dataset is appropriate.
- Multimodal Distance (MM-Dist): calculate average distances between situation context and generated motions.
- R-Precision [33]: quantify the semantic discriminability of generated motions by establishing a text-to-motion cross-modal retrieval task.

A. Multi-Embodiment Testing

To evaluate the cross-embodiment capability of our works, two experiments are provided. Specifically, the bone-size stability is evaluated to determine if the generation module can generate actions satisfying skeletal systems for various humanoids. Additionally, a retargeting accuracy testing is implemented to evaluate performance of execution module.

Bone-size stability. Bone size is a fundamental characteristic of the human body and significantly influences the expressiveness of behaviors. To verify its stability, we examine the generated results of two different bone types (fixed and floating) for three classical humanoid robots in various human motions. The stability parameter is calculated by the average absolute error between the bone lengths of the input robot pose and the generated poses.

The results in Table II show that the bone-size stability in the GT dataset is poor for both bone types of all the selected humanoid robots. The reason is that the trained policy is not familiar with the input humanoid bone structure; thereby, the generated pose may be with wrong bone size (see Fig. 4). Although the motion retargeting strategy is widely used to adapt unmatched poses to humanoid robots, it is still faced with insurmountable defects, i.e., human-likeness loss during the motion retargeting processing as discussed in [10]. To address this problem, we aim to produce human-like poses for humanoids while adhering to their skeletal system constrains. Therefore, we add the augmented the pseudo GT to re-train the policy. Fortunately, the generation error is optimized to the millimeter level, which means the bone-size stability for various humanoid robots is acceptable. Additionally, in both datasets, the floating bone shows greater sensitivity compared to the fixed bone as shown in Table II and Fig. 4. This increased sensitivity arises because the floating bone's size fluctuates with different movements, even when the body configuration remains constant.

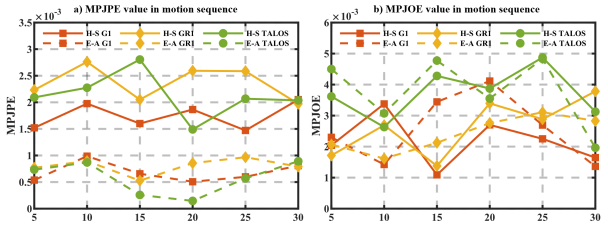

Fig. 5. The retargeting accuracy between generated actions and the robot's actions for three humanoids in different frames.

Retargeting accuracy. We testing the retargeting accuracy by implementing it in various humanoids, e.g., GR1, with generated motion in the situational context *"a person is start to dancing with partner."*. As indicated in Fig. 5, the quantitative results proved that our framework can be adopt to the selected humanoid robots while keeping action accuracy and human similarity with small value of MPJPE and MPJOE for both E-A and H-S. However, the MPJPE value of H-S is bigger than that of E-A as we emphasize the rotational pose in pose retargeting process for floating bones due to its unique character. Fortunately, the accuracy of the retargeting module is guaranteed by keep small&similar MPJOE and acceptable MPJPE in both E-A and H-S. A qualitative demonstration is presented in Fig. 6, where two continuing frames in a planned human-like motion with given situation are selected. For each motion frame in different color boxes, the pose with light color is the initial state while the pose with dark color is the planned pose based on given conditions, i.e., goals. It is proved that our framework has good adaptability for action planning and driving most humanoid robots.

B. Ablation Study

We implement our framework using two classical network architectures, MLP-based and Transformer-based policies, to assess its performance. The KIT_ML dataset was selected for this evaluation due to its consistent data frequency of 240 Hz.

Ablation study of frequency. To evaluate the influence of data frequency on our framework performance, we train our network with four types of frequency, i.e., 240Hz (original), 60Hz, 15Hz and random frequency. In the testing process, a target pose was randomly sampled in the test dataset. The testing results in Table III indicate that the random sampling method (hindsight training) achieved the best values of all

TABLE III
THE QUALITATIVE ANALYSIS RESULTS

Model	Frequency	MPJPE		MPJOE		FMD↓	MM-Dist↓	R-Precision ↑
		E-A↓	H-S↓	E-A↓	H-S↓			
MLP(w)[0]	240	0.135	0.181	0.279	0.237	3.56	10.4	0.317
MLP(w)	60	0.091	0.127	0.219	0.178	3.02	9.07	0.341
MLP(w)	15	0.056	0.077	0.172	0.131	1.95	6.74	0.384
MLP(w/o)[1]	Random	0.031	0.45	0.121	0.084	–	–	–
MLP(w)	Random	0.015	0.026	0.044	0.043	0.951	3.22	0.492
Transformer(w/o)[2]	Random	0.023	0.031	0.071	0.058	–	–	–
Transformer(w)[3]	Random	**0.008**	**0.019**	**0.039**	**0.036**	**0.737**	**3.07**	**0.508**

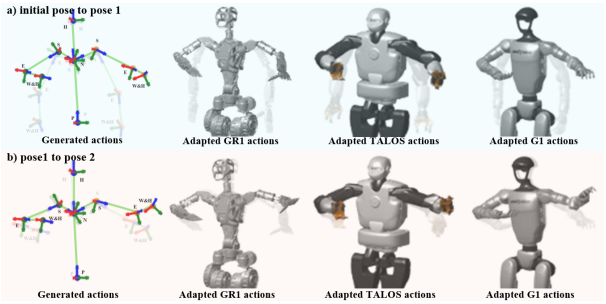

Fig. 6. The implementation of our framework for expressive human-like behaviors in different humanoid robots. From left to right, the three popular humanoid robots are Fourier's GR1, Unitree's G1, and PAL's TALOS.

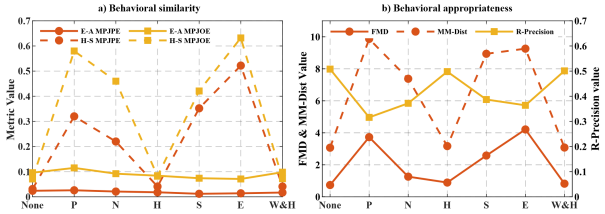

Fig. 7. The results of ablation for observation input in our framework. The X-axis means the locking data, e.g., the "None" means that we provide whole observation, while "N" means the Neck data in observation is missing.

evaluation metrics for behavioral similarity and appropriateness compared to others with fixed frequency. The reason is that the two training frames for fixed frequency groups are too close to identify the difference between the frames when the dataset frequency is too high. The inference is supported by the tendency of experimental results for all the evaluation metrics while the dataset frequency is decreased. As the motion speed of demonstrated human action is not uniformly distributed in the time domain, it still cannot achieve optimal framework performance even with small recording frequencies, such as 15Hz. Due to the random frequency strategy enlarging the dataset and enriching the data range by randomly selecting goals from the trajectory, the performance of E-A and H-S is improved. Furthermore, the hindsight training method also enriched the motion semantics of human-demonstrated behavior. Therefore, the contextual appropriateness is enhanced, which is indicated by the metrics of FMD, MM-Dist, and R-Precision.

Ablation study of motion situation. To explore the effects of the motion situation l for human-like motion generation, an ablation study with two architectures is carried out, and the results are shown in Table III. Flag (w) means the model is trained by the dataset with motion situation l. Otherwise, it is marked as (w/o). The results show that the model performance for both behavioral similarity and appropriateness trained with motion situation l is better than that without motion situation. The reason for this promotion is that the appropriate human action pose is related to the contextual situation and motion target, such as *"lifting a box while avoiding collisions with the table on your left"*. Unfortunately, this information of contextual situation cannot be explicitly represented by human actions data. Thereby, the dataset with situation context l is more essential for the human-like behavior planning task for humanoid robots.

Ablation study of observation. The results in Fig. 7 indicate that the missing observation of "pelvis, neck, shoulder, and elbow" influence E-A in both MPJPE and MPJOP is less. However, it is hugely damaged the H-S and behavioral appropriateness. The reason is that lack of observation of these joints would destroy the body chain which will suffer the generation performance. In contract, the missing observation of "Hand and Wrist&hand" has less impact on both behavior similarity and appropriateness, due to the end-effectors' human-like poses along with the part of our training target, i.e., generating a pose to meeting the goals. Therefore, it is necessary to ensure the integrity of the observations.

C. Comparative Results

To prove the efficiency of our method, we compare our framework with other state-of-the-art works. Three classical methods are selected in the comparison study, i.e., IK algorithm, siMLPe [32], and HumanPlus [22].

Methods comparison. For behavioral similarity, the results shown in Table IV demonstrated that our approach achieves better results on both the MPJOE and MPJPE metrics compared to the siMLPe and HumanPlus models. This outcome arises because the selected model just considers the behavioral traits in human-demonstrated sequential motion, which are also widely accepted by motion synthesis. Consequently, these methods do not perform as expected when dealing with unseen motion traits in training dataset, which are common in daily tasks. Fortunately, our approach has the capability to address this challenge due to its innovative mechanism. By comparing

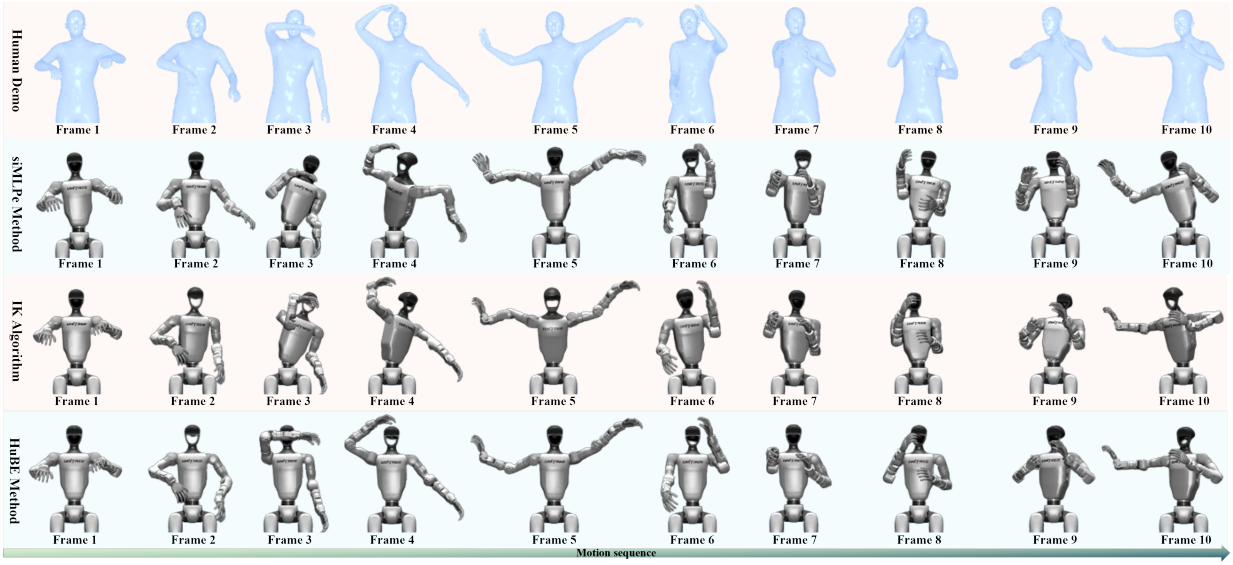

Fig. 8. An example of action-wise results ("Playing Kung Fu") performed by the humanoid robot G1. Motion sequences are shown from left to right. From top to bottom: (1) human demonstration data from our dataset, (2) results generated by siMLPe, (3) motions planned by the IK algorithm with target pose g, and (4) results produced by our proposed method (HuBE) conditioned on g and contextual situation l. Due to the physical constraints of G1, the target g refers to the wrist poses of both arms.

with IK, our framework gets poorer but acceptable results in E-A with both MPJRE and MPJOE metrics as the E-A is the only target to optimize for the IK solver. Specially, the error of E-A in our method is mainly introduced in generation process. However, we can achieve better performance in H-S than IK, which ensures that the generated pose for humanoid robots is more similar to humans. Another advantage of our method (0.0207s) is time-saving when compared with IK (0.3618s), which is based on a searching mechanism. Therefore, our method is a good choice for human-like pose generation tasks. For behavioral appropriateness, our method can achieve smaller FMD and MM-Dist values, which means that our method can generate higher quality and closer distribution of real actions than the selected method. Furthermore, the higher the R-Precision value indicates that our method can get better semantic matching accuracy between the generated action and the situation description. Therefore, the behavioral appropriateness of the generated action in our module is guaranteed.

Action-wise. We also evaluated our method with different actions, and the action-wise results are also shown in Fig. 8 and Table IV. By analyzing the results, we know that our framework is well adapted to various human actions as the value of evaluation metrics for different actions is stable and small. Consequently, our framework can effectively generate stable human-like poses for humanoids over various tasks when compared with other works.

V. Conclusion

In this paper, we presented *HuBE*, a bi-level closed-loop framework for human-like behavior execution in humanoid robots. By integrating robot state, goal poses, and contextual situations, our framework enables the generation of behaviors that satisfy both behavioral similarity and appropriateness. To support this framework, we developed HPose, a context-enriched dataset with 6D joint pose representation and situational annotations, and introduced a bone scaling-based data augmentation strategy to ensure millimeter-level cross-embodiment compatibility across heterogeneous humanoid robots. Extensive experiments on multiple commercial platforms demonstrated that *HuBE* significantly outperforms state-of-the-art methods. These results highlight the potential of *HuBE* to serve as a transferable foundation for scalable and socially acceptable human-like behavior execution. One constraint of *HuBE* is that our planner primarily emphasizes the human-like movements of the robot's upper body. However, the movements of the lower body are also significant in influencing human actions. Therefore, our upcoming research will focus increasingly on designing full-body expressive behaviors that are feasible and mimic human movements.

References

[1] G. Gulletta, E. Silva, W. Erlhagen, R. Meulenbroek, M. Costa, and E. Bicho, "A human-like upper-limb motion planner: Generating naturalistic movements for humanoid robots," *International Journal of Advanced Robotic Systems*, vol. 18, no. 2, p. 1729881421998585, 2021.

[2] G. Hill, "An unwillingness to act: behavioral appropriateness, situational constraint, and self-efficacy in shyness," *Journal of Personality*, vol. 57, no. 4, pp. 871–890, 1989.

[3] Z. Jiang, Y. Xie, K. Lin, Z. Xu, W. Wan, A. Mandlekar, L. Fan, and Y. Zhu, "Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning," 2024.

[4] S. Aisyah, T. Eugene, E. Nurhayati *et al.*, "Tantangan linguistik dalam pengimplementasian big data berbahasa indonesia pada robot humanoid: Tinjauan dan rekomendasi," *Jurnal Pendidikan Bahasa dan Sastra Indonesia*, vol. 1, no. 1, pp. 9–9, 2025.

[5] H. Zhang, W. Li, J. Liu, Z. Chen, Y. Cui, Y. Wang, and R. Xiong, "Kinematic motion retargeting via neural latent optimization for learning sign language," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4582–4589, 2022.

[6] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "Bc-z: Zero-shot task generalization with robotic imitation learning," in *Conference on Robot Learning*. PMLR, 2022, pp. 991–1002.

TABLE IV
THE ACTION-WISE RESULTS FOR DIFFERENT METHODS

Action	Model	MPJPE		MPJOE		FMD↓	MM-Dist↓	R-Precision↑
		E-A	H-S↓	E-A	H-S↓			
Dancing	IK	0.00832	0.324	0.000853	1.62	–	–	–
	HumanPlus	0.109	0.184	1.73	3.18	–	–	–
	siMLPe	0.0841	0.138	1.42	3.21	1.73	6.25	0.361
	Ous	**0.0196**	**0.0337**	**0.0743**	**0.0543**	**0.749**	**3.71**	**0.484**
Exercise	IK	0.00785	0.425	0.000742	2.24	–	–	–
	HumanPlus	0.139	0.195	2.52	2.65	–	–	–
	siMLPe	0.0728	0.114	2.12	2.64	1.68	6.82	0.337
	Ous	**0.0176**	**0.0297**	**0.0476**	**0.0403**	**0.763**	**3.28**	**0.492**
Playing	IK	0.00983	0.524	0.000947	1.93	–	–	–
	HumanPlus	0.0972	0.115	2.48	294	–	–	–
	siMLPe	0.0736	0.101	2.62	3.37	1.94	7.28	0.317
	Ous	**0.0188**	**0.0299**	**0.0553**	**0.0427**	**0.823**	**4.01**	**0.466**
Running	IK	0.00941	0.183	0.000318	2.12	–	–	–
	HumanPlus	0.0862	0.103	1.80	1.75	–	–	–
	siMLPe	0.0761	0.129	1.65	1.96	1.69	6.04	0.493
	Ous	**0.0197**	**0.0295**	**0.0441**	**0.0403**	**0.715**	**3.27**	**0.531**

[7] L. Annabi, Z. Ma, and S. Nguyen, "Unsupervised motion retargeting for human-robot imitation," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 204–208.

[8] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi, "Learning human-to-humanoid real-time whole-body teleoperation," *arXiv preprint arXiv:2403.04436*, 2024.

[9] M. Plappert, C. Mandery, and T. Asfour, "The kit motion-language dataset," *Big data*, vol. 4, no. 4, pp. 236–252, 2016.

[10] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive whole-body control for humanoid robots," *arXiv preprint arXiv:2402.16796*, 2024.

[11] R. Dou, S. Yu, W. Li, P. Chen, P. Xia, F. Zhai, H. Yokoi, and Y. Jiang, "Inverse kinematics for a 7-dof humanoid robotic arm with joint limit and end pose coupling," *Mechanism and Machine Theory*, vol. 169, p. 104637, 2022.

[12] D. Rempe, T. Birdal, A. Hertzmann, J. Yang, S. Sridhar, and L. Guibas, "Humor: 3d human motion model for robust pose estimation," in *International Conference on Computer Vision (ICCV)*, 2021.

[13] Y. Xie, V. Jampani, L. Zhong, D. Sun, and H. Jiang, "Omnicontrol: Control any joint at any time for human motion generation," *arXiv e-prints*, pp. arXiv–2310, 2023.

[14] M. Petrovich, M. Black, and G. Varol, "Action-conditioned 3d human motion synthesis with transformer vae," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 985–10 995.

[15] X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, and G. Yu, "Executing your commands via motion diffusion in latent space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 000–18 010.

[16] Y. Shafir, G. Tevet, R. Kapon, and A. Bermano, "Human motion diffusion as a generative prior," in *The Twelfth International Conference on Learning Representations*, 2024.

[17] K. Karunratanakul, K. Preechakul, S. Suwajanakorn, and S. Tang, "Guided motion diffusion for controllable human motion synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2151–2162.

[18] Q. Zhang, P. Cui, D. Yan, J. Sun, Y. Duan, G. Han, W. Zhao, W. Zhang, Y. Guo, A. Zhang *et al.*, "Whole-body humanoid robot locomotion with human reference," *arXiv preprint arXiv:2402.18294*, 2024.

[19] Y. Yan, E. Mascaro, T. Egle, and D. Lee, "I-ctrl: Imitation to control humanoid robots through constrained reinforcement learning," *arXiv preprint arXiv:2405.08726*, 2024.

[20] J. Li, Y. Zhu, Y. Xie, Z. Jiang, M. Seo, G. Pavlakos, and Y. Zhu, "Okami: Teaching humanoid robots manipulation skills through single video imitation," *arXiv preprint arXiv:2410.11792*, 2024.

[21] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang, "Open-television: Teleoperation with immersive active visual feedback," *arXiv preprint arXiv:2407.01512*, 2024.

[22] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," *arXiv preprint arXiv:2406.10454*, 2024.

[23] K. Loveland, D. Pearson, B. Tunali-Kotoski, J. Ortegon, and M. Gibbs, "Judgments of social appropriateness by children and adolescents with autism," *Journal of Autism and Developmental Disorders*, vol. 31, pp. 367–376, 2001.

[24] Y. Gao, F. Yang, M. Frisk, D. Hemandez, C. Peters, and G. Castellano, "Learning socially appropriate robot approaching behavior toward groups using deep reinforcement learning," in *2019 28th IEEE international conference on robot and human interactive communication (RO-MAN)*. IEEE, 2019, pp. 1–8.

[25] S. Ivanov and C. Webster, "Perceived appropriateness and intention to use service robots in tourism," in *Information and Communication Technologies in Tourism 2019: Proceedings of the International Conference in Nicosia, Cyprus, January 30–February 1, 2019*. Springer, 2019, pp. 237–248.

[26] Y. Zhou, "Perceived appropriateness: A novel view for remediating perceived inappropriate robot navigation behaviors," in *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, pp. 781–783.

[27] N. Mahmood, N. Ghorbani, N. Troje, G. Pons-Moll, and M. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5442–5451.

[28] J. Lin, A. Zeng, S. Lu, Y. Cai, R. Zhang, H. Wang, and L. Zhang, "Motion-x: A large-scale 3d expressive whole-body human motion dataset," *Advances in Neural Information Processing Systems*, vol. 36, pp. 25 268–25 280, 2023.

[29] R. Sutton, "Reinforcement learning: An introduction," *A Bradford Book*, 2018.

[30] F. Torabi, G. Warnell, and P. Stone, "Behavioral cloning from observation," *arXiv preprint arXiv:1805.01954*, 2018.

[31] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, P. Abbeel, and W. Zaremba, "Hindsight experience replay," *Advances in neural information processing systems*, vol. 30, 2017.

[32] W. Guo, Y. Du, X. Shen, V. Lepetit, X. Alameda-Pineda, and F. Moreno-Noguer, "Back to mlp: A simple baseline for human motion prediction," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 4809–4819.

[33] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, "Motiondiffuse: Text-driven human motion generation with diffusion model," *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 6, pp. 4115–4128, 2024.