Article | Received 5 November 2025; Revised 6 February 2026; Accepted 28 February 2026; Published 20 May 2026
https://doi.org/10.55092/rl20260015

Cross-embodiment human-like behavior execution for humanoid robots

Shipeng Lyu[1,2], Fangyuan Wang[1,2], Weiwei Lin[2], Luhao Zhu[2], David Navarro-Alarcon[1,3,*] and Guodong Guo[2]

[1] Department of Mechanical Engineering, The Hong Kong Polytechnic University (PolyU), Kowloon, Hong Kong, China

[2] Ningbo Institute of Digital Twin, Eastern Institute of Technology (EIT), Ningbo, China

[3] Research Institute for Smart Ageing (RISA), PolyU, Kowloon, Hong Kong, China

* Correspondence author; E-mail: dnavar@polyu.edu.hk.

Highlights:

- A unified view of human-likeness: we define human-like robot behavior by jointly considering behavioral similarity and behavioral appropriateness.
- Morphology-aware motion generation: we generate morphology-aware motions by integrating robot state, task goal, and situational context, without structural mismatches between generation and execution.
- Cross-embodiment adaptation: we enable robust deployment across heterogeneous humanoid platforms through bone-scaling-based augmentation with millimeter-level compatibility.

Abstract: Achieving both behavioral similarity and behavioral appropriateness in human-like motion generation for humanoid robots remains an open challenge. This challenge is further compounded by the requirement of cross-embodiment adaptability. To address this problem, we propose HuBE, a morphology-aware framework for human-like behavior execution in humanoid robots. HuBE integrates robot state, goal pose, and situational context to generate behaviors that are both behaviorally similar and appropriate, while reducing 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 strong baselines. Overall, HuBE provides a transferable foundation for robust, human-like behavior execution across diverse humanoid robots.

Keywords: humanoid robot; human-like behavior; behavioral appropriateness; pose generation

Copyright©2026 by the authors. Published by ELSP. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.

1. Introduction

Humanoid robots play a pivotal role in human-robot interaction (HRI), where the human-like behavior 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 discomfort in humans. This phenomenon presents a fundamental challenge when generating human-like behavior 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. This guarantees that the generated actions are contextually meaningful and socially acceptable. In this paper, we focus on situationally grounded posture selection, rather than long-horizon social planning or affective dialogue behaviors. As illustrated in Figure 1, human-like behaviors must not only satisfy kinematic goals, but also align with human expectations in specific contextual scenarios [2]. This requires adapting actions to these situations to achieve behavioral appropriateness.

Although substantial progress has been made in improving behavioral similarity via motion retargeting (Inverse Kinematics, IK) [3–5], imitation learning [6–8], and motion synthesis [9–11], behavioral appropriateness remains relatively underexplored in the literature. Moreover, existing humanoid motion generation methods that primarily focus on behavioral similarity suffer from several limitations. First, most existing motion datasets [12] provide limited semantic descriptions of expressive human behaviors. They typically rely on 6D joint representations that lack annotations linking situational context to behavioral appropriateness. In addition, many datasets use only short and coarse textual labels (e.g., "running") to describe motion sequences, which fail to capture the rich and nuanced contextual information underlying human behaviors. Second, many existing pipelines adopt a modular design in which motion generation and retargeting are treated as two loosely coupled, independent stages. This often introduces structural mismatches between the generated poses and the robot's physical embodiment, degrading behavioral similarity and semantic consistency. For example, prior work [13] generates poses without conditioning on the current robot state, causing the outputs to violate robot-specific kinematic constraints. Consequently, additional morphology-adaptation operation at execution time is required, which can further distort the intended behaviors. Third, cross-embodiment adaptation remains largely unaddressed, limiting deployment across heterogeneous humanoid robots [14] in diverse real-world settings. This refers to the challenge of adapting behaviors across different robot embodiments without additional modifications.

To address these challenges, we propose HuBE, a bi-level morphology-aware framework for human-like behavior execution in humanoid robots. Generally, HuBE aims to generate behaviors that achieve both behavioral similarity and behavioral appropriateness, while enabling robust adaptation for cross-embodiment deployment. In this paper, cross-embodiments refer to the ability to generate human-like behaviors that are inherently compatible with different humanoid robot embodiments at the pose generation stage. This eliminates the need for additional test-time morphology adaptation to compensate for embodiment mismatches. Generally, morphology adaptation is a post-hoc motion

retargeting that compensates for embodiment mismatch after generating an embodiment-agnostic human motion. However, embodiment adaptation in HuBE is realized by conditioning the behavior generation policy on robot-specific states and training it with morphology-aware data. The retargeting process is only used to map generated Cartesian-space actions (6D poses) to robot joint-space commands (joint angles).

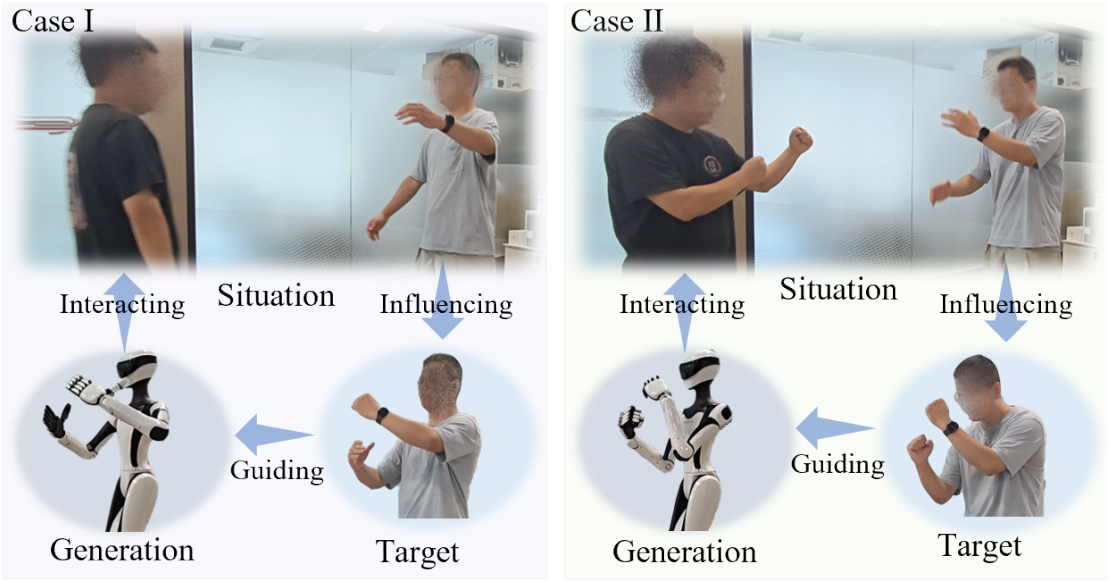


Figure 1. Situational context shapes human-like action execution by affecting behavioral appropriateness. Two examples illustrate that similar end-effector goals can require different full-body postures under different situational semantics. Case I ("hugging a friend") favors an open-arm posture with elevated elbows to convey affection, whereas Case II ("boxing with a friend") adopts guarded elbows for defensive intent. These context-dependent pose adaptations reflect behavioral appropriateness, *i.e.*, alignment between motion and situation.

To achieve HuBE, we first construct the HPose dataset by extending existing open-source motion datasets with fine-grained situational semantics through detailed language annotations. Meanwhile, we preserve human-like behaviors by representing motions using 6D poses. Second, HuBE introduces a morphology-aware mechanism through implicit skeletal parameter adaptation, allowing the integration of robot state, behavioral goals, and contextual situations into an end-to-end motion generation and execution pipeline. This design effectively avoids structural mismatches between motion generation and robot control. Furthermore, incorporating contextual situations into the behavior generation process enables HuBE to produce contextually appropriate actions while maintaining behavioral similarity. Finally, we propose a bone-scaling-based data augmentation strategy to simulate the morphological diversity of state-of-the-art humanoid robots, enabling robust generalization across heterogeneous platforms. Together, these components establish millimeter-level cross-embodiment compatibility and distinguish HuBE from prior approaches. The contributions are summarized as follows:

- We propose a new perspective on human-likeness with behavioral similarity and behavioral appropriateness, supported by the HPose dataset with fine-grained situational annotations.
- We propose a morphology–aware framework that enables humanoid robots to generate actions satisfying both behavioral similarity and behavioral appropriateness, and eliminate structural

mismatches between motion generation and execution.

- We introduce a bone-scaling-based data augmentation strategy that facilitates cross-embodiment training across heterogeneous humanoid robots, allowing the generation policy to produce embodiment-compatible poses.

Table 1 offers a comparison between HuBE and representative humanoid behavior generation approaches. Instead of reporting performance metrics, it summarizes key method-level capabilities along five dimensions: (1) Context (situational): whether generation is grounded in situational context beyond low-level control signals; (2) Embodment-aware: whether the method explicitly accounts for the target embodiment, *i.e.*, robot state; (3) Cross-embodment: whether behaviors can be deployed across embodiments without per-robot morphology adaptation; (4) Hard constrains: whether feasibility is ensured via explicitly enforced kinematic/physical constraints (e.g., joint limits); and (5) Focus: whether the primary emphasis is behavioral similarity (Sim.) and behavioral appropriateness (App.). Overall, the table highlights that HuBE jointly supports situational context grounding, morphology-aware generation, cross-embodiment deployment, and feasibility-aware execution, while explicitly targeting both similarity and appropriateness.

Table 1. Conceptual comparison between HuBE and representative humanoid behavior/motion generation approaches. Symbols in the table follow a simple convention: $\surd$ for yes, $\times$ for no, and $\triangle$ for partial support.

Method	Context (situational)	Embodment -aware	Cross -embodment	Hard constrains	Focus
IK-based retargeting [3–5]	$\times$	$\surd$	$\triangle$	$\surd$	Sim.
Text-conditioned synthesis [9–11]	$\triangle$	$\triangle$	$\times$	$\times$	Sim.
Imitation learning/ BC [6–8]	$\triangle$	$\surd$	$\times$	$\triangle$	Sim.
HuBE (Ours)	$\surd$	$\surd$	$\surd$	$\surd$	Sim. + App.

2. Related work

To ensure the behavioral similarity of humanoid robots, recent approaches focus on replicating human motion patterns through two representative paradigms: motion synthesis and imitation learning (IL). Motion synthesis methods generate human-like trajectories using auto-regressive models [15] or diffusion processes [9–11], but produce physically infeasible poses requiring post-processing. Moreover, a challenge for motion synthesis methods [16,17] is that they cannot precisely regulate joint-level control inputs introduced for robotics execution. In most cases, these methods rely on post-hoc morphology adaptation and inverse kinematics to map generated human motions to specific robot embodiments. Motivated by these issues, we focus on frame-wise, state-conditioned pose generation, which better supports controllability and feasibility during execution.

IL methods [18–20] learn control policies directly from human demonstrations, but generalizing across heterogeneous morphologies and novel constraints remains challenging. Furthermore, some studies also explore alternative observation modalities (e.g., egocentric vision) for more natural behavior learning [13,21]. However, end-to-end control policies are often tightly coupled to specific embodiments

or task settings, making cross-embodiment deployment difficult without additional adaptation. In contrast, we learn a frame-wise pose generation policy conditioned on robot state and goals, and use bone-scaling augmentation to improve cross-embodiment generalization.

Research on behavioral appropriateness, rooted in social psychology and related disciplines, examines the regulatory mechanisms through which social norms govern individual actions [22]. In HRI, empirical studies confirm that the congruence between robot behaviors and user cognitive expectations strongly affects social acceptance for robots [23,24]. Although prior work has made progress on behavioral appropriateness for mobile robots (e.g., verification models [25]), it remains challenging for humanoid robots to perform contextually appropriate behaviors. One key limitation is that situational semantics (beyond low-level control signals) are still not systematically integrated into many human-like motion generation pipelines, which constrains context-consistent behavior synthesis. To address this, we propose a semantic-task fusion framework that integrates situational semantics with robotic behavioral goals.

3. Methodology

To drive humanoid robots to perform human-like behavior while ensuring behavioral similarity and appropriateness, we propose HuBE, a human-like behavior execution framework. HuBE consists of three modules, as shown in Figure 2: data processing, behavior generation, and motion execution. The behavior generation module operates in a frame-wise, state-conditioned manner, where motion situations, the current robot state, and target poses are jointly taken as inputs to generate human-like actions. By augmenting data during training and explicitly incorporating the robot's current configuration into the generation process, HuBE achieves morphology-aware motion generation. This ensures the generated poses are robot-scale and kinematically consistent with the target humanoid robot, instead of relying on a fixed human template. The behavior execution module then maps the generated Cartesian poses to the humanoid robot, ensuring kinematic feasibility and precise execution, while preserving the intended behavioral expression.

3.1. Cross-embodiment adaptation mechanism

In this work, we consider cross-embodiment humanoids that share a comparable semantic body-part topology (*i.e.*, the same set of body frames), while allowing differences in link lengths, joint limits, and kinematic ranges. Specifically, HuBE achieves cross-embodiment adaptation by embedding embodiment awareness throughout the training, generation, and execution, rather than compensating for embodiment mismatches through test-time morphology adaptation. At the policylevel, morphology-aware augmentation exposes the policy to diverse humanoid kinematic parameters during training, preventing overfitting to a single canonical skeleton. At the generation level, the behavior generation module conditions on the robot's current configuration to preserve embodiment compatibility. At the execution level, an IK solver is used solely to enforce kinematic feasibility and robot-specific constraints, rather than to compensate for embodiment mismatch. Through this multi-level mechanism, HuBE enables humanoids with heterogeneous morphologies to generate and execute human-like behaviors in a morphology-aware manner.

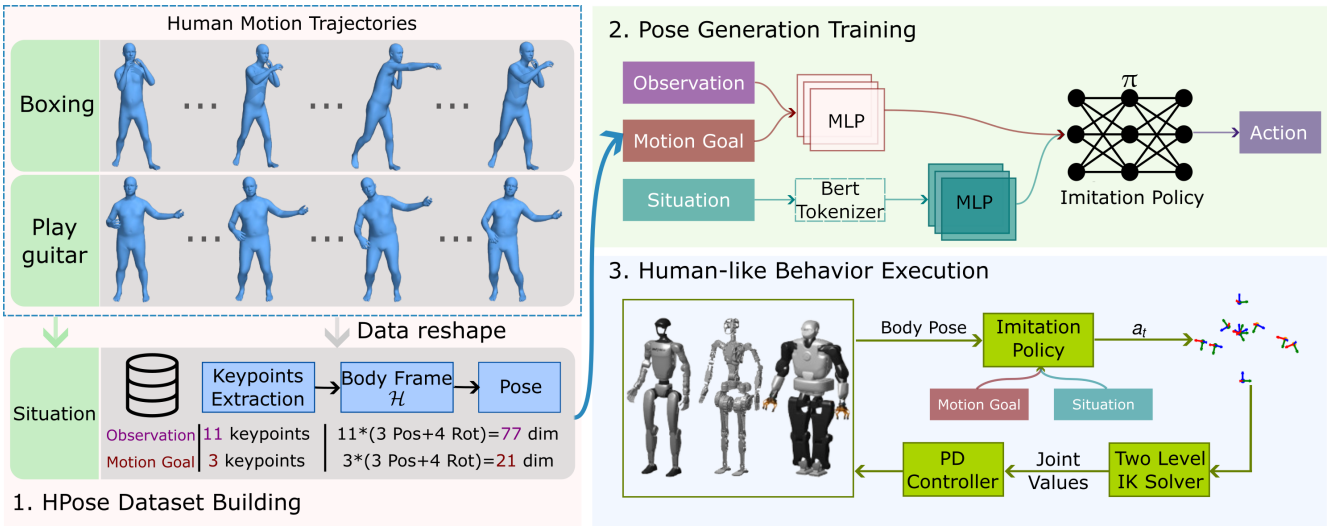


Figure 2. Overview of HuBE's pipeline, including **(i)** dataset construction, **(ii)** model training, and **(iii)** deployment/execution. During generation, the policy is explicitly conditioned on the current robot state (observation), together with the goal pose and situational description, to produce retargeting-compatible behaviors. This forms a morphology–aware generation–execution pipeline, rather than relying on a test-time motion adaptation step to compensate for embodiment mismatch.

3.2. Data specifications

We construct HPose, a morphology-aware training dataset that exposes the behavior generation policy to diverse humanoid embodiments during training, eliminating the need for adaptation to different robot configurations at test time. For this study, we construct HPose by augmenting and unifying several open-source motion datasets, including KIT [12], AMASS [26], and Motion-X [27], as summarized in Table 2. We apply the following three operations:

- Data definition: As depicted in Figure 3, we simplify the full-body skeleton to 11 main joints. Specifically, the positions of the left and right hands are defined to be identical to those of the corresponding wrists. Additionally, we introduce the rotational data that is represented by quaternions for each joint in our dataset. To enrich the situational context associated with human motions, we leverage a large language model (GPT-4o) to generate descriptive annotations for each motion sequence based on the original annotation $\hat{l}$. Specifically, we select two randomly frames from each motion sequence associated with the annotation $\hat{l}$ and provide them as inputs to the Large Language Model (LLM). After reasoning about the contextual information conveyed by these inputs, the LLM generates a rich and detailed description l.
- 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 observed state s_i. Additionally, a situational context l_i is appended to characterize human behavior scenarios.
- Data augmentation: To enable the dataset structure described above to generalize across heterogeneous humanoid embodiments, we further introduce a morphology-aware data augmentation strategy.

Table 2. Data source distribution of HPose dataset.

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

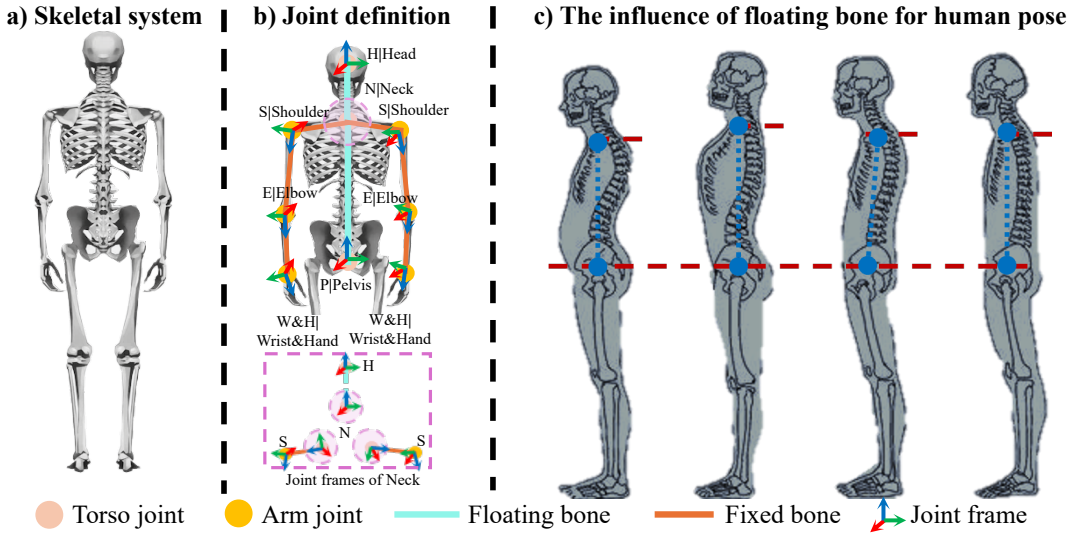


Figure 3. The representation of human skeleton. **(a)** Full-body human skeletal structure; **(b)** Simplified upper-body kinematic chain used to build HPose; **(c)** Example illustrates the effect of a floating base/torso ("floating bone"), which can change global posture even when local limb configurations are similar.

Unlike conventional data augmentation that aims to improve robustness by increasing data diversity, our augmentation strategy is specifically designed for cross-embodiment training by explicitly modeling humanoid morphological variations. Due to differences in body morphology between humans and humanoid robots, such as arm length, it is necessary to adjust the skeletal structure of the original motion data to generate poses compatible with humanoids of various configurations. The human upper-body skeleton is modeled as a rooted kinematic tree, where each directed joint pair (J_i, J_k) corresponds to a parent-child relationship, and the pelvis is treated as the root joint. We obtain the augmented dataset $\mathscr{H}'$ (pseudo ground truth, Pseudo GT) through a bone-scaling operation by updating the bone size with collected humanoid robot body parameters $\mathscr{R} = \{r_i, r_k\}$ as shown in Equation (1).

$$J'_k = J'_i + \frac{J_k - J_i}{\|J_k - J_i\|} \cdot \|r_k - r_i\| \cdot \alpha_{ik}. \tag{1}$$

where each directed joint pair (J_i, J_k) is defined according to the kinematic tree structure, and α_{ik} denotes a posture-dependent scaling factor. The robot skeletal parameters $\mathscr{R}$ are collected from nine representative humanoid robots, such as Unitree H1 and PAL's TALOS. Through this morphology-aware augmentation, the behavior generation policy is exposed to a distribution of humanoid skeletal parameters during training. This design enables embodiment adaptation to be addressed at train time, rather than relying on test-time retargeting or post-hoc kinematic compensation.

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. Fixed bones are skeletal segments where the Euclidean distances between adjacent joints remain approximately constant across different poses in human motion data. These typically correspond to limb segments connected (e.g., upper arms and forearms). In contrast, floating bones correspond to skeletal segments whose effective lengths vary with body posture due to the accumulation of multiple joint rotations, trunk bending, and articulation effects, which is commonly observed in spine. The classification of fixed and floating bones is predefined based on skeletal topology and remains consistent across all datasets and humanoid robot platforms. Based on this categorization, we adopt a unified bone-scaling formulation as Algorithm 1. For fixed bones, a constant scaling factor is applied to directly match the corresponding robot bone length. For floating bones, an additional posture-dependent scaling ratio is introduced, computed as the ratio between the bone length in the current pose and that in a canonical human T-pose provided in the original dataset. This design preserves posture-induced variations of floating bones while ensuring geometric compatibility with heterogeneous humanoid morphologies, without introducing separate algorithmic branches.

Algorithm 1: Bone scaling for data augmentation

The data: Human skeleton $\mathscr{H} = \{J_i\}$, robot skeleton $\mathscr{R} = \{r_i\}$, Human T-pose skeleton $\mathscr{H}^T = \{J_i^T\}$
The result: Augmented human skeleton $\mathscr{H}'$

Initialize root joint position $J'_0 = J_0$
Construct directed joint pair set $D = \{(J_i, J_k)\}$ from the kinematic tree (parent–child)
Sort D according to tree traversal order from root to leaf
foreach $(J_i, J_k) \in D$ **do**
    Compute bone direction $\hat{d}_{ik} = \frac{J_k - J_i}{\|J_k - J_i\|}$
    Compute robot bone length $l_{ik} = \|r_k - r_i\|$
    if (J_i, J_k) *is a fixed bone* **then**
        $\alpha_{ik} \leftarrow 1$
    else
        $\alpha_{ik} \leftarrow \frac{\|J_k - J_i\|}{\|J_k^T - J_i^T\|}$
    $J'_k \leftarrow J'_i + \hat{d}_{ik} \cdot l_{ik} \cdot \alpha_{ik}$
return $\mathscr{H}'$

3.3. Pose generation

We aim to drive the humanoid robot to perform human-like poses based on situational context l, the current state s, and a motion target g. By conditioning the policy on the robot state s, the generation process is tightly coupled with the robot's kinematic structure and current configuration. This design enables HuBE to generate motions within the robot's feasible motion space in a morphology-aware manner. We formulate the humanoid robot motion generation problem as a Markov Decision Process [28] without a specific reward function [29], denoted as $(s, a, T, g, l, \rho_0, \gamma)$. s is the state space containing all keypoint joint poses of the upper body of the robot, which implicitly encodes embodiment-specific structural information, such as bone lengths. g is the goal space that includes the target poses of end effectors, *i.e.*, both hands and head. 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. $T : s \times a \to s$ 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, $\pi_\theta : s, g, l \to a$,

which specifies the action that should be taken in the robot's current state (Equation (2)). Our objective is to learn an imitation policy π_θ from a 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:

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

While the above formulation enables imitation-based pose generation, it does not guarantee behavioral appropriateness or generalization to unseen target poses under varying situational contexts. This issue is exacerbated by the common practice of learning from fixed human motion sequences without explicitly modeling situational constraints. To overcome these challenges, we propose input alignment and hindsight training strategies for stable generation of human-like behaviors.

Input alignment: To encourage behavioral appropriateness under varying situational contexts, we use the situational description l_i to guide motion generation. Given l_i, we encode it with a BERT text encoder ($BERT(\cdot)$) to obtain a fixed-dimensional language vector. In parallel, an MLP ($MLP(\cdot)$) maps humanoid proprioception (s_i, g_i) to the same dimensionality. The two embeddings are then fused by concatenation, as shown in Figure 2 and Equation (3).

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

Hindsight training: As illustrated in Figure 4a, we implement a goal relabeling mechanism to exploit long-horizon dependencies within trajectories. Therefore, based on the idea of hindsight experience replay [30], which takes the achieved state in the same trajectory τ as the final goal, we augment the dataset $\mathscr{D}$ by sampling a future offset $k \sim \{1, \ldots, H\}$ within a horizon of H frames and using the corresponding future pose as the relabeled target (Equation (4)). Specifically, we randomly sample a future time step $t+k$ for the current state s_t. The achieved future state s_{t+k} plays two roles: (i) it provides a hindsight goal g_{t+k} that replaces the original goal to condition the policy input; (ii) it serves as supervision target. This mechanism augments training by treating reached states as successful outcomes for their corresponding relabeled goals.

$$\mathscr{D}_{\text{new}} = \{(s_t,\ g_{t+k},\ l_t,\ a_{t+k}) \mid \tau \in \mathscr{D},\ t = 0, \ldots, T-1\}, \quad k \sim \{1, \ldots, H\}. \tag{4}$$

Policy network: Our policy network contains two implementation structures, *i.e.*, transformer-based structure and MLP-based structure. The transformer-based policy shown in Figure 4b embeds text tokens (128-dimensional) and the concatenated state/goal vector via a linear layer. The fused sequence is processed by positional encoding and a 3-layer, 4-head Transformer encoder. The result is processed by the MLP head to output action a. Transformer-style policies are well-suited to language-conditioned control because attention layers can fuse textual and proprioceptive signals into a unified policy representation and directly map them to control commands. The MLP-based policy in Figure 4c is a feedforward policy that maps state and a language embedding to actions through stacked fully connected layers. These MLP policies serve as a standard and effective baseline in robot imitation and reinforcement learning, offering a lightweight controller that can be trained for continuous control and locomotion tasks. We train the policy

network with MSE on next observation prediction, using AdamW (lr = 1e−4, weight decay: 1e−4), batch size 64, and 100 epochs; validation occurs every $10k$ steps and checkpoints every $100k$.

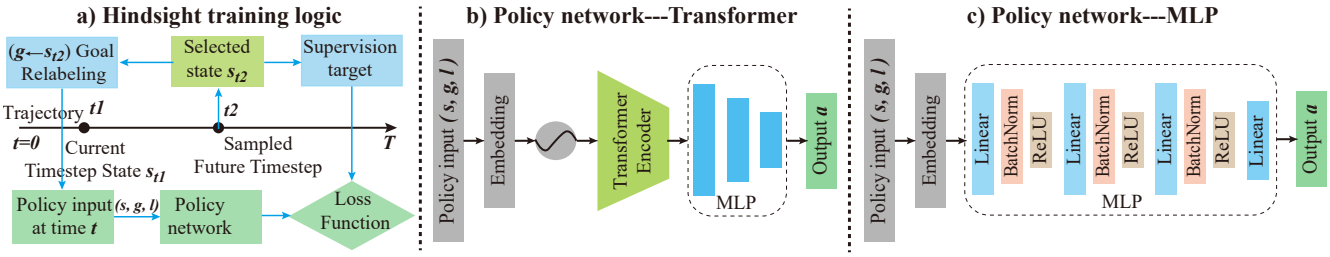


Figure 4. The details of pose generation training. **(a)** Hindsight training logic. For a given state s_{t_1}, a future state s_{t_2} is randomly sampled from the trajectory. This future state serves dual purposes: it is extracted as a relabeled goal g to condition the policy input at time t, and simultaneously acts as the supervision target for predicting the future outcome; **(b)** The transformer-based structure of policy network; **(c)** The MLP-based structure of policy network.

3.4. Motion execution

Our IK module is a constrained realization step that maps already embodiment-consistent Cartesian targets a_t into joint commands. It does not need to perform additional embodiment adaptation or motion warping. To achieve this Cartesian-to-joint mapping, a key challenge is preserving human-like motion, particularly for floating-bone regions (e.g., torso). In human motion, the effective lengths of these regions vary with posture, but on a humanoid robot, they correspond to fixed link lengths.

Multi-constrained IK formulation: We formulate the Cartesian-to-joint mapping as a weighted least-squares IK problem with multiple task objectives and physical constraints. Specifically, we jointly consider: (i) end-effector pose tracking (soft objectives), (ii) torso orientation consistency to capture floating-bone posture variability (soft, high-priority objectives), (iii) joint limits as hard constraints, and (iv) temporal regularization for smoothness.

Let ${}^{w}T_i(J) \in SE(3)$ denote the forward kinematics of body frame i under joint configuration J, and ${}^{w}T_i^{\text{target}}$ be the target transform extracted from the generated pose. We define the orientation tracking error for frame i as:

$$\phi_i(J) = \log\left(R_i^{\text{target}\top} R_i(J)\right) \in \mathbb{R}^3. \tag{5}$$

where $p_i(\cdot)$ and $R_i(\cdot)$ are the translation and rotation of ${}^{w}T_i(\cdot)$, and $\log(\cdot)$ is the logarithmic log map on $SO(3)$ returning the axis-angle vector in $\mathbb{R}^3$. The IK problem is written as:

$$\begin{cases} J^* = \arg\min\limits_{J} & \sum\limits_{i\in\mathscr{E}} \left(w_i^{\text{pos}} \| p_i(J) - p_i^{\text{target}} \|^2 + w_i^{\text{rot}} \| \phi_i(J) \|^2 \right) + \lambda \| J - J_{\text{prev}} \|^2, \\ \text{s.t.} & J_{\min} \le J \le J_{\max}. \end{cases} \tag{6}$$

where $J_{\min}, J_{\max}$ are joint-limit bounds enforced as hard constraints, and the regularization term encourages temporal continuity.

Modeling and constraining floating-bone variability: Since robot link lengths are fixed, posture-dependent variability observed in human floating-bone regions is expressed on the robot primarily through torso joint

rotations. Therefore, we explicitly constrain multiple torso-related joints (e.g., pelvis and neck) by assigning higher rotational weights for floating-bone related joints:

$$w_i^{\text{rot}} \gg w_i^{\text{pos}}, \quad \forall i \in \mathscr{E}_{\text{torso}}. \tag{7}$$

This prioritizes torso orientation consistency (capturing posture-induced variability) during the Cartesian-to-joint mapping, while still tracking end-effector goals.

Two-stage solving and conflict handling: Executing human-like actions on floating-base humanoid robots is challenging, as single-level whole-body inverse kinematics formulations often yield under-constrained or ill-conditioned solutions due to strong torso–limb coupling. This leads to torso drift and unstable global postures. Therefore, we split the full joint vector as $J = [J_{\text{torso}}^\top, J_{\text{arms}}^\top]^\top$ and solve IK in two sequential stages to resolve conflicts between torso posture preservation and arm reachability. First, we compute J_{torso}^* by optimizing Equation 6 with respect to J_{torso} with $\mathscr{E} = \mathscr{E}_{\text{torso}}$. Second, we lock torso joints (fix $J_{\text{torso}} = J_{\text{torso}}^*$) to obtain a reduced model and solve for arm joints J_{arms}^* with $\mathscr{E} = \mathscr{E}_{\text{arms}}$. This stage-wise strategy enforces a clear priority: torso (floating-bone posture) is preserved first, then arms are solved conditioned on the fixed torso.

4. Experimental results

To evaluate the effectiveness of the proposed method, we employ a set of quantitative metrics to assess both behavioral similarity and behavioral appropriateness of the generated humanoid motions.

Behavioral similarity: We consider two complementary aspects to evaluate the similarity between generated robot motions and their corresponding human demonstrations:

- End-effector Accuracy (E-A): measures the positional and rotational accuracy of key end-effectors, including both hands and the head.
- Human Similarity (H-S): evaluates overall pose consistency by considering all selected body markers.

For both aspect, we report the Mean Per Joint Position Error (MPJPE) [31] and the Mean Per Joint Orientation Error (MPJOE), which quantify positional and rotational discrepancies, respectively.

Behavioral appropriateness: To assess whether the generated motions are semantically appropriate under given situational contexts, we employ the following three metrics:

- Fréchet Motion Distance (FMD): measures the distributional similarity between generated motions and authentic human motions in a learned feature space, analogous to the FID [32]. This metric is based on the assumption that human-demonstrated motions in the dataset are semantically appropriate.
- Multimodal Distance (MM-Dist): calculates average distances between situational context and generated motions.
- R-Precision [32]: quantifies the semantic discriminability of generated motions by establishing a text-to-motion cross-modal retrieval task.

4.1. Annotation quality and consistency analysis

To evaluate the quality, consistency, and potential bias of the GPT-4o-generated situational annotations in HPose, we conducted a lightweight human evaluation. Three evaluators with backgrounds in robotics and human motion analysis independently assessed 100 randomly sampled motion clips. For each clip, evaluators

were shown a motion visualization together with the corresponding generated description and rated it on a 5-point scale (1: strongly disagree, 5: strongly agree) along three dimensions of annotation quality:

- Accuracy: evaluating whether the description correctly reflects the observed motion;
- Context appropriateness: assessing whether the inferred situational context is reasonable and avoids over-interpretation;
- Disambiguation usefulness: measuring whether the description provides additional information beyond the original short label to distinguish behaviors under similar end-effector goals.

As summarized in Table 3, the GPT-4o-generated situational annotations achieved consistently high scores across all three evaluation dimensions. In particular, the mean scores for accuracy, contextual appropriateness, and disambiguation usefulness are all close to 4.0. Moreover, the inter-rater agreement exceeds 0.75 for all metrics, indicating good agreement among evaluators. These results suggest that the machine-generated annotations are reliable and suitable for conditioning context-aware behavior generation.

Table 3. Human evaluation of GPT-4o-generated situational annotations in HPose.

Metric	Mean $\pm$ Std	Inter-rater Agreement
Accuracy	4.12 ± 0.63	ICC(2,k) = 0.79
Context appropriateness	3.97 ± 0.71	ICC(2,k) = 0.76
Disambiguation usefulness	4.25 ± 0.58	ICC(2,k) = 0.81

The GPT-4o-generated situational annotations are designed to encode high-level action intent and contextual information rather than fine-grained kinematic details. As illustrated in Figure 5, the annotations accurately describe the overall action context, while their inherent granularity limits prevent explicit modeling of subtle hand, finger, or pose-level variations. These limitations are expected and have limited impact on policy learning, as precise motion details are handled by the behavior generation and execution modules. We further inspected typical failure modes noted by evaluators. The most common issues include (i) hallucinated scene/object details that are not observable from the motion visualization, and (ii) sporadic left/right or body-part confusion in complex reaching sequences. To mitigate these biases, we constrained the annotation prompt to only describe observable action intent and physical interactions, and applied a lightweight post-filter that triggers an automatic rewrite pass when such phrases are detected.

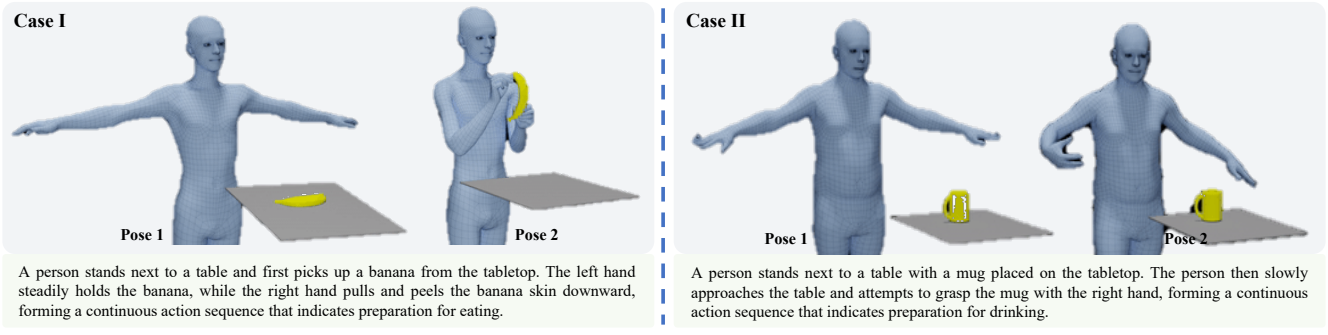


Figure 5. Examples of GPT-4o-generated situational annotations for HPose. The figure highlights both typical successful cases (descriptions that correctly reflect the observed motion and context) and common limitations, such as descriptions that capture high-level intent but omit fine-grained kinematic or temporal details.

4.2. Cross-embodiment testing

We evaluate cross-embodiment capability by deploying the same trained policy on multiple humanoid robots without retraining or test-time adaptation. Specifically, we evaluate bone-size stability to determine whether the generation module produces actions consistent with the skeletal structures of different humanoid robots. Additionally, retargeting accuracy is evaluated to assess the performance of execution module.

Bone-size stability: Bone-size stability measures whether the generated pose respects the target embodiment's bone proportions, *i.e.*, whether the policy avoids producing embodiment-incompatible kinematic structures. By comparing models trained on ground-truth and pseudo ground-truth data (with morphology-aware augmentation), we can directly assess the effect of train-time embodiment exposure on cross-embodiment consistency. Specifically, we examine the generated results across three representative humanoid robots under diverse human motion scenarios. The stability metric is computed as the average absolute error between the bone lengths of the input robot pose and those of the generated poses.

The results in Table 4 show that the bone-size stability in the GT dataset is poor for both bone types of all the selected humanoid robots. This is because the trained policy has not been exposed to the target humanoid skeletal structure during training. As a result, the generated poses may exhibit incorrect bone lengths. Although morphology adaptation is commonly used to solve this mismatching problem, it suffers from inherent limitations, such as degraded human-likeness during adaptation, as discussed in [13]. To address this problem, we aim to produce human-like poses for humanoids while adhering to their skeletal system constraints. Therefore, we augment the training data with the pseudo ground-truth samples and retrain the policy. As a result, the generation error is reduced to the millimeter level, indicating acceptable bone-size stability across different humanoid robots. In addition, floating bones exhibit greater sensitivity than fixed bones in both datasets, as shown in Table 4 and Figure 6. This increased sensitivity arises because the effective lengths of floating bones vary with body posture, even when the underlying skeletal configuration remains unchanged.

Table 4. Bone-size stability results in two datasets. The unit of the results is 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**

Retargeting accuracy: We evaluate retargeting accuracy by deploying the generated motions on multiple humanoid robots, such as GR1, under the situational context "a person starts dancing with a partner". As shown in Figure 7, the quantitative results demonstrate that our framework is effectively applicable to the selected humanoid robots, maintaining action accuracy and human similarity. This is reflected by low MPJPE and MPJOE for both E-A and H-S. However, the MPJPE for H-S is higher than that for E-A, as greater emphasis is placed on preserving rotational consistency during pose mapping process for floating bones. Overall, the retargeting module maintains reliable accuracy, achieving consistently low MPJOE and acceptable MPJPE for both E-A and H-S. A qualitative demonstration is

presented in Figure 8, where two consecutive frames from a planned human-like motion under a given situational context are shown. For each motion frame shown in different colored boxes, the light-colored pose represents the initial state, while the dark-colored pose indicates the planned target pose under the specified conditions (e.g., goals). These results indicate that our framework exhibits strong adaptability for action planning and execution across diverse humanoid robots.

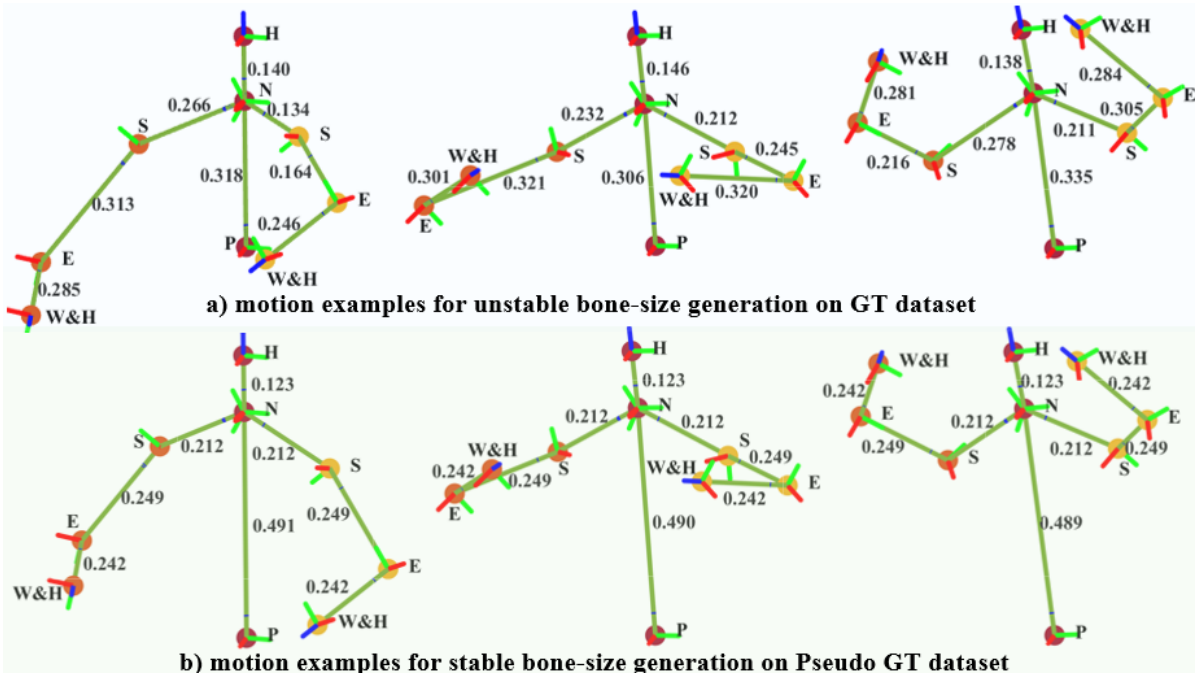


Figure 6. Bone-size stability under cross-embodiment generation. The top row shows actions generated by a model trained on the original ground-truth data (GT), while the bottom row uses the pseudo-GT produced with bone-scaling augmentation. The input pose is captured from GR1 and the situational description is "Lift a heavy box over your head with both hands." Numbers indicate the corresponding bone-length estimates (units: meter), showing that pseudo-GT training yields more consistent limb proportions.

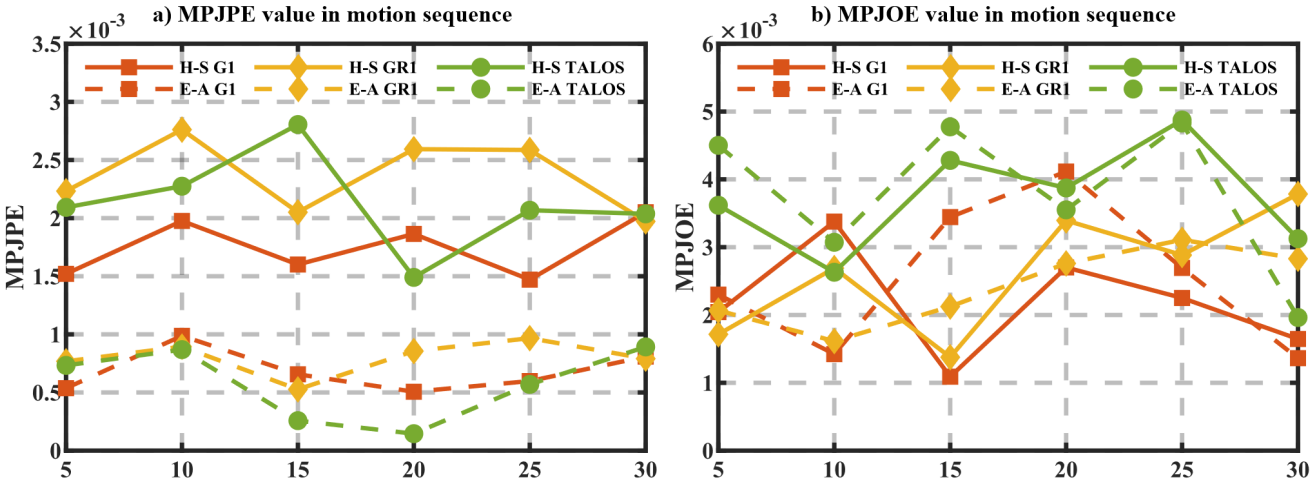


Figure 7. Retargeting accuracy of the generated motions during execution on three humanoid platforms (GR1, G1, and TALOS). We report MPJPE and MPJOE over time. Curves compare two evaluation settings (E-A and H-S). Lower values indicate better kinematic agreement between the generated motion and the robot-executed motion.

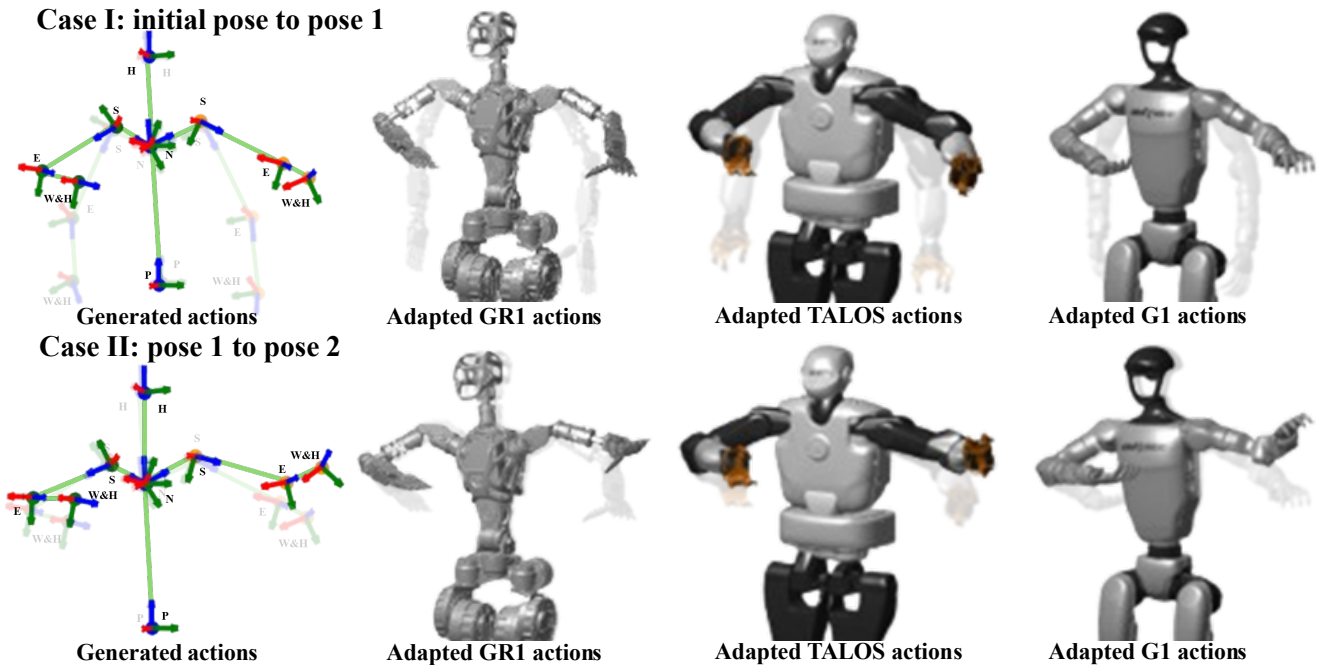


Figure 8. Implementation and deployment of HuBE on three humanoid robots (from left to right: Fourier GR1, Unitree G1, and PAL TALOS). The same generation–execution pipeline is applied across platforms, while the embodiment-specific components (e.g., kinematic mapping) ensure feasibility under each robot's constraints.

4.3. Ablation study

We conduct ablation studies to disentangle the contributions of key design choices in HuBE, covering: (i) conditioning with situational context l; (ii) target sampling strategy in training (fixed-frequency *vs.* hindsight); and (iii) policy architecture (MLP *vs.* Transformer). Experiments are performed on KIT_ML, whose trajectories are uniformly sampled at 240 Hz. This property enables controlled comparisons when varying the target sampling strategy. Additionally, we analyze robustness under partial observability by ablating observation completeness and further examine the architectural role of robot state by fully removing robot observations from the generator.

Ablation study on data frequency: To evaluate the impact of data frequency on the performance, we train our model under four frequency settings, *i.e.*, 240 Hz (original), 60 Hz, 15 Hz and a random frequency (hindsight training). We perform the evaluation on KIT_ML and randomly sample target poses from the test set during evaluation. As shown in Table 5, reducing the fixed sampling frequency consistently improves both behavioral similarity and appropriateness metrics. Specifically, MPJPE(E-A) decreases from 0.135 (240 Hz) to 0.091 (60 Hz) and further to 0.056 (15 Hz), with a similar trend for H-S. The FMD (MM-Dist) also drops from 3.56 to 1.95 (10.4 to 6.74), and R-Precision increases from 0.317 to 0.384. This is because, under high fixed sampling frequencies, adjacent training frames are temporally too close to exhibit meaningful differences. This is supported by the consistent performance trends across all evaluation metrics as the sampling frequency decreases. However, since the motion speed of human actions is not uniformly distributed over time, simply reducing the sampling frequency (e.g., to 15 Hz) does not yield optimal performance. In contrast, the hindsight training strategy randomly selects future poses along the same trajectory as training targets, effectively performing goal relabeling. This broadens

the temporal/distance span and increases goal diversity. This expanded supervision improves kinematic agreement in both E-A and H-S, while also yielding a closer match to the distribution of human behaviors (lower FMD/MM-Dist) and better text-conditioned appropriateness (higher R-Precision).

Ablation study on situational context: We evaluate how situational context l affects contextual appropriateness in human-like motion generation. As reported in Table 5, we compare models trained with l (w) and without l (w/o) under both MLP and Transformer backbones. Notably, appropriateness-oriented metrics (FMD, MM-Dist, and R-Precision) are only meaningful when l is provided, since they quantify the alignment between generated behaviors and the given situation semantics. With l enabled, both backbones achieve substantially better appropriateness (lower FMD/MM-Dist and higher R-Precision), indicating that the generated behaviors are more consistent with the intended situational description.

Table 5. Ablations on target sampling strategy (frequency *vs.* random/hindsight), context conditioning l, and architecture (MLP/Transformer), evaluated by similarity and appropriateness metrics. "-" indicates not applicable without l. ↑ indicates higher is better (R-Precision), and ↓ indicates lower is better (all other metrics).

Model	Frequency	MPJPE		MPJOE		FMD↓	MM-Dist↓	R-Precision ↑
		E-A↓	H-S↓	E-A↓	H-S↓			
MLP(w)	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)	Random	0.031	0.045	0.121	0.084	–	–	–
MLP(w)	Random	0.015	0.026	0.044	0.043	0.951	3.22	0.492
Transformer(w/o)	Random	0.023	0.031	0.071	0.058	–	–	–
Transformer(w)	Random	**0.008**	**0.019**	**0.039**	**0.036**	**0.737**	**3.07**	**0.508**

In contrast, the impact of l on kinematic similarity is secondary. Similarity can still be learned to some extent from motion targets and pose dynamics alone. However, without l, the policy is forced to average over multiple plausible postures that share similar end-effector goals but differ in social or scene semantics. Therefore, providing l reduces this ambiguity by guiding context-dependent posture selection (e.g., "lifting a box while avoiding collisions with a table on the left"). This improves appropriateness primarily, with a modest additional gain in similarity as a by-product.

Ablation study on model architecture: We compare two widely used policy architectures, namely an MLP-based policy and a Transformer-based policy, with results reported in Table 5. Under the same training setting, the Transformer achieves consistently better behavioral similarity and contextual appropriateness than the MLP. We attribute this gap to the Transformer's self-attention mechanism, which explicitly models interactions among the goal, robot observations, and situational tokens. This enables context-dependent feature weighting and reduces the averaging effect over multimodal plausible motions that MLPs tend to produce.

Ablation study on observation completeness: The results in Figure 9 indicate that removing observations of the pelvis, neck, shoulder, and elbow has a relatively small impact on E-A, as measured by MPJPE and MPJOE. However, this significantly degrades both human similarity (H-S) and behavioral appropriateness. This is because missing observations of these joints break the integrity of the kinematic chain, impairing motion generation. In contrast, removing observations of the hand and wrist has a smaller

impact on both behavioral similarity and appropriateness, as end-effector poses are explicitly included in the training targets. Therefore, maintaining complete joint observations is essential for stable and human-like motion generation.

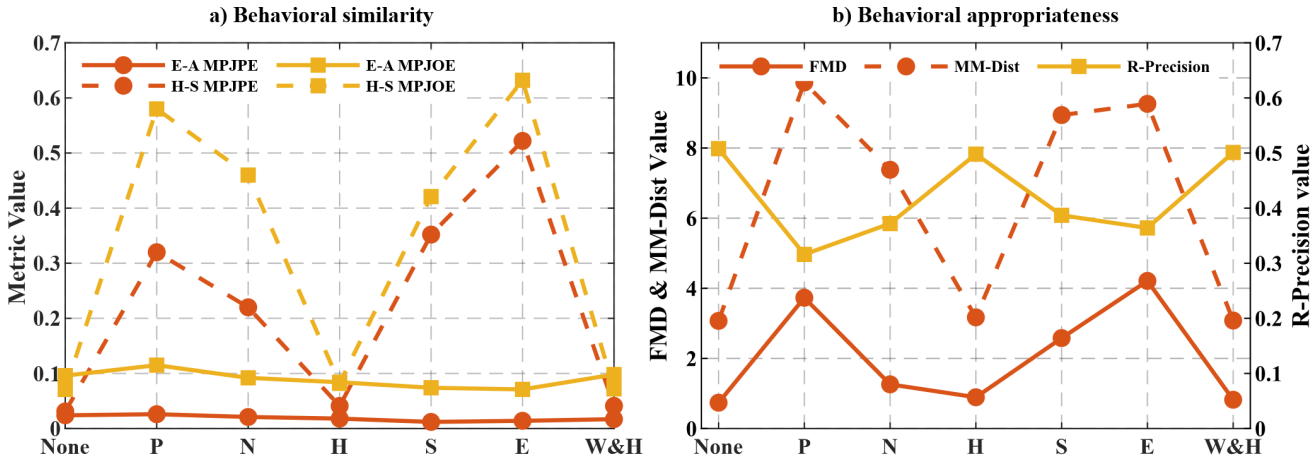


Figure 9. Ablation study on the observation input. Each x-axis label indicates which observation component is masked/removed during inference (e.g., None: full observation; N: neck-related observation removed; other labels follow the same convention). The y-axis reports the corresponding performance metric, showing that removing key observation components degrades generation quality.

The above results analyze robustness under partial observation, where only subsets of the robot state are removed. To further examine the role of robot observation at the architectural level, we consider a more extreme ablation by completely removing the robot state s from the generator. Since s is the only robot-specific conditioning signal in HuBE, this ablation not only removes state information, but also degenerates the tightly coupled generation–execution process into a decoupled pipeline that relies on post-hoc IK retargeting.

To evaluate the impact of this degeneration on execution stability, we additionally report the IK failure rate and solving cost, as summarized in Table 6. Specifically, an execution is regarded as failed if the IK solver does not converge within a predefined maximum number of iterations (1000), or if the final execution error exceeds an acceptable threshold. The solving cost is measured as the average per-frame IK solving time in seconds.

Table 6 reports the effect of completely removing robot observation s on both motion quality and execution stability. Compared with the full HuBE model, removing robot state leads to a degradation in motion quality. In particular, the MPJPE and MPJOE metrics for both E-A and H-S increase significantly, indicating that the generated motions become less accurate and less human-like. More importantly, the impact of robot-state removal is also evident at the execution level. The IK failure rate increases from 5% to 17%, and the average per-frame solving time increases from 0.028 s to 0.157 s. These results indicate that, in the absence of robot observation, the execution module must perform substantially larger corrective adjustments and faces a more difficult and unstable optimization problem. Taken together, these results demonstrate the dual role of robot state in HuBE. On the one hand, robot observation provides critical contextual information for generating behaviorally appropriate and human-like motions. On the

other hand, it is essential for maintaining a tightly coupled generation–execution process that reduces structural mismatch and ensures stable and efficient execution.

Table 6. The motion quality and execution stability results for ablation study on robot observation s. ↓ indicates lower is better.

Model	MPJPE↓		MPJOE↓		Fail Rate↓	Solving Time↓
	E-A	H-S	E-A	H-S		
HuBE (w/o s)	0.0763	0.116	0.972	1.38	0.17	0.157
HuBE (full)	0.0194	0.0217	0.0415	0.0432	0.05	0.028

HuBE (w/o s) denotes the condition that robot observation is removed from the generator, resulting in an open-loop generation process.

4.4. Comparative results

To evaluate the effectiveness of our framework, we compare it with several representative baselines. Specifically, we compare HuBE against a standard IK solver, siMLPe [31], and HumanPlus [21].

Method comparison: For behavioral similarity, the results in Table 7 show that HuBE achieves lower MPJPE and MPJOE than siMLPe and HumanPlus across all evaluated actions. A possible reason is that these baselines mainly model human motion sequences without explicitly conditioning on robot embodiment or situational context, which can limit robustness when the test motions differ from the training distribution. In contrast, HuBE explicitly conditions motion generation on robot state and situational context, enabling better generalization to unseen behaviors.

Compared with the IK solver, HuBE yields slightly higher errors in E-A (MPJPE and MPJOE), since the IK baseline directly optimizes end-effector tracking. Nevertheless, HuBE achieves substantially better H-S than IK, indicating closer whole-body pose consistency with human demonstrations. In addition, HuBE is significantly more efficient, with an average inference time of 0.0207 s per frame, compared to 0.164 s for IK, which relies on iterative optimization. In terms of behavioral appropriateness, HuBE achieves lower FMD and MM-Dist, indicating a closer match to the distribution of human motions than the learning-based baselines. Moreover, higher R-Precision indicates better semantic alignment between generated motions and their corresponding situational descriptions. Overall, these results suggest that HuBE generates motions that are both human-like and contextually appropriate across diverse scenarios.

Action-wise results: We further evaluate our method across different action categories, with action-wise results reported in Figure 10 and Table 7. Across categories, HuBE maintains consistently competitive similarity and appropriateness metrics, suggesting good generalization across diverse actions. Compared with the baselines, HuBE produces more human-like and context-consistent poses across a wide range of tasks.

User study: To further validate behavioral appropriateness and similarity from a human perceptual perspective, we conducted a user study comparing HuBE with IK, HumanPlus, and siMLPe. We designed an online video-based survey consisting of three contextually grounded scenarios. For each scenario, participants were shown four videos demonstrating robot behaviors generated by the different methods under identical task settings. A video-based protocol was adopted to ensure controllability and repeatability across conditions, allowing participants to focus on perceptual evaluation rather than interaction variability. The presentation order of methods was counterbalanced using a balanced Latin square design.

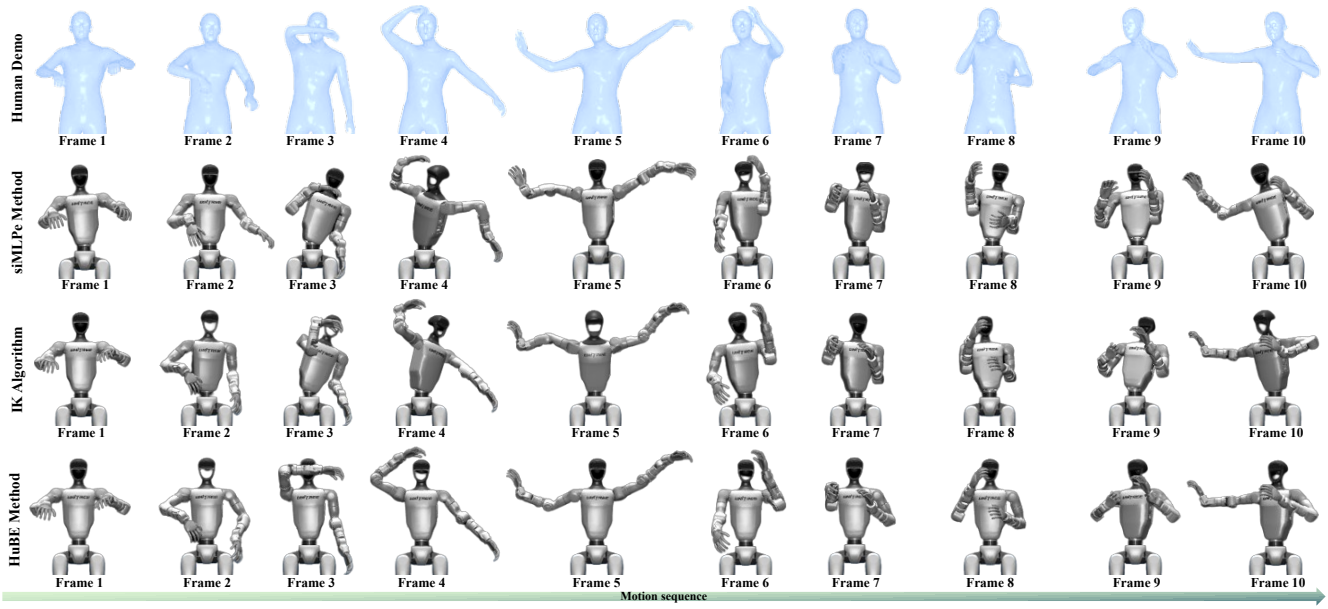


Figure 10. An example of action-wise results ("Playing Kung Fu") performed by the humanoid robot G1. Motion sequences are shown from left to right. From top to bottom: (1) human demonstration data from our dataset, (2) results generated by siMLPe, (3) motions planned by the IK algorithm with target pose 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.

Table 7. The action-wise results for different methods. ↑ indicates higher is better (R-Precision), and ↓ indicates lower is better (all other metrics). All values are reported with three significant figures.

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
	HuBE	**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
	HuBE	**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	2.94	–	–	–
	siMLPe	0.0736	0.101	2.62	3.37	1.94	7.28	0.317
	HuBE	**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
	HuBE	**0.0197**	**0.0295**	**0.0441**	**0.0403**	**0.715**	**3.27**	**0.531**

Seventeen participants (10 male, 7 female), aged between 18 and 56, were recruited. All participants provided informed consent prior to the study, and the experiment was approved by the institutional ethics review board. After viewing each set of videos, participants rated the robot behaviors on 5-point scales

assessing behavioral appropriateness, similarity, and social acceptability. Specifically, they were asked to rate: (Q1) whether the robot's behavior was appropriate for the given task and context; (Q2) whether the behavior appeared human-like and natural; and (Q3) whether the behavior was socially acceptable and comfortable to observe, with scores ranging from 1 (strongly disagree) to 5 (strongly agree). We hypothesized that HuBE would outperform the baseline methods in perceived behavioral appropriateness, similarity, and social acceptability. Given the within-subject design and ordinal scale data, statistical significance was assessed using one-tailed Wilcoxon signed-rank tests.

The results in Table 8 show that HuBE consistently outperforms IK and HumanPlus across all three perceptual dimensions. Specifically, HuBE achieves significantly higher ratings in behavioral appropriateness and similarity than both IK and HumanPlus ($p < 0.01$), and also demonstrates improved social acceptability ($p < 0.05$). Compared with siMLPe, HuBE still obtains significantly higher scores in behavioral appropriateness and similarity ($p < 0.05$), while no significant difference is observed in social acceptability ($p > 0.05$). Overall, these results indicate that HuBE enhances perceived behavioral appropriateness and similarity without compromising social acceptability, providing perceptual support for the effectiveness of the proposed framework.

Table 8. User study results on perceived behavioral appropriateness, similarity, and social acceptability using 5-point scales. p-values are from one-tailed Wilcoxon signed-rank tests comparing HuBE with each baseline. We report participant-level averages across the three scenarios (*i.e.*, each participant contributes one score per method per question), and compute mean $\pm$ SD over participants.

Method	Behavioral appropriateness			Behavioral similarity			Social acceptability		
	Mean	SD	p	Mean	SD	p	Mean	SD	p
IK	3.896	0.876	< 0.01	3.851	0.897	< 0.01	4.232	0.614	< 0.05
HuBE	4.822	0.946		4.787	0.931		4.684	0.649	
HumanPlus	4.045	0.964	< 0.01	4.116	0.914	< 0.01	4.295	0.584	< 0.05
HuBE	4.822	0.946		4.787	0.931		4.684	0.649	
siMLPe	4.325	0.973	< 0.05	4.665	0.953	< 0.05	4.624	0.704	> 0.05
HuBE	4.822	0.946		4.787	0.931		4.684	0.649	

4.5. Real-robot evaluation

Compared with simulation, the physical GR1 platform exhibits sensing noise, actuation delays, limited control bandwidth, and hardware constraints, which introduce modeling discrepancies and execution uncertainty. Therefore, we further evaluate HuBE on a physical humanoid robot, the Fourier GR1, to examine whether the generated behaviors preserve execution reliability and behavioral appropriateness under real-world embodiment constraints.

As shown in Figure 11, we evaluate HuBE on GR1 under multiple situational descriptions, including drinking water, pretending to play an electric guitar, and lifting a box. Across trials for each scenario, the robot executes the generated upper-body behaviors reliably, exhibiting clear semantic distinctions in posture and motion patterns. Despite real-world uncertainty, the executed behaviors remain qualitatively consistent with those generated in simulation, in terms of end-effector intent and overall upper-body expressiveness.

To provide a more comprehensive evaluation, we further conduct a quantitative analysis of execution

performance. Specifically, we measure three key metrics across multiple trials for each scenario:

- Consistency: calculated by finding the Root Mean Square (RMS) of the differences between the robot's position and orientation at each time step across trials. A lower RMS Deviation (RMSD) indicates that the robot's behavior is more consistent.
- Stability: calculated by evaluating the variation in the robot's end-effector position or orientation over time, based on the deviation from the ground truth trial implemented in robot. A lower Standard Deviation means that the robot's movements are more stable during the trial.
- Repeatability: computed as the global standard deviation of the end-effector trajectories across all trials and time steps. A lower execution consistency (EC) indicates higher repeatability, meaning that the robot performs the same task more consistently across different trials.

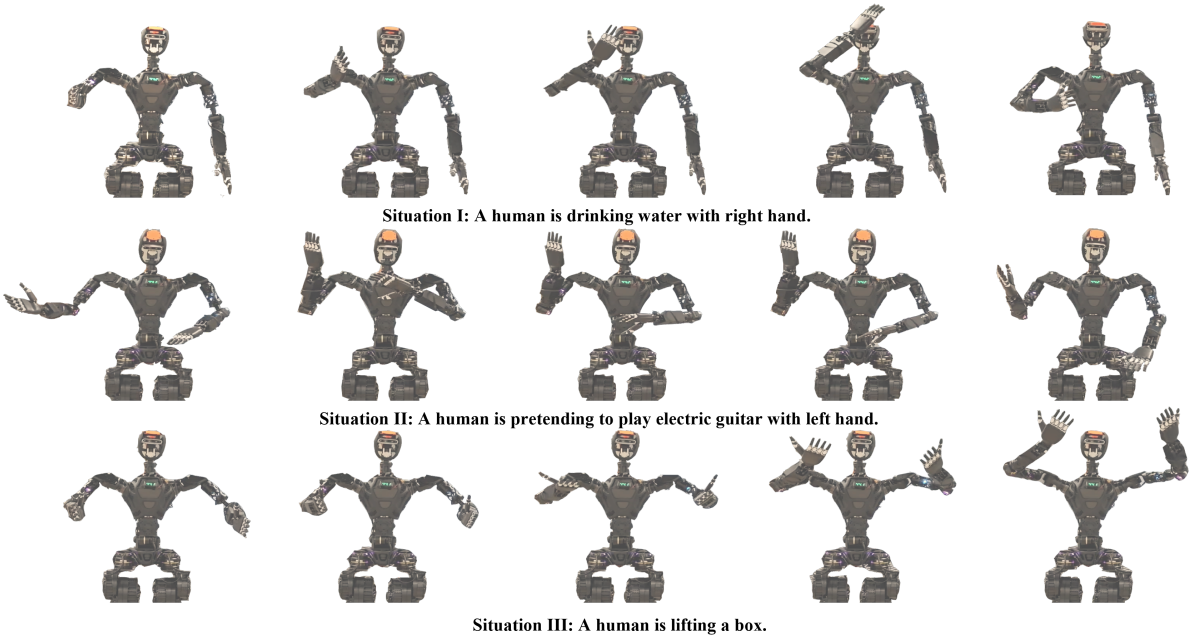


Figure 11. Real-robot evaluation on Fourier GR1. HuBE generates upper-body behaviors under different situational descriptions (drinking water, pretending to play an electric guitar, and lifting a box). The figure shows representative execution snapshots; despite sensing noise and actuation delays, GR1 executes stable motions with clear semantic differences across scenarios. Pos. refers to position, ori. refers to orientation.

The quantitative evaluation in Table 9 of HuBE implementation on the Fourier GR1 across three scenarios demonstrates high consistency, stability, and repeatability in the robot's performance. Specifically, the RMSD for position are consistently low, below 0.01 m, with the Lifting a Box scenario showing the smallest deviation at 0.005 m. Orientation RMSD is also low, ranging from 0.97° for Drinking Water to 1.5° for Playing Guitar, indicating consistent execution across trials. The SD reflect small variation, with position SD below 0.01 m and orientation SD below 1.7°, further confirming stable movement during each trial. Finally, EC values for position (below 0.002 m) and orientation (below 0.36°) highlight the robot's ability to reliably repeat behaviors across trials with minimal deviation. These results show that HuBE enables the robot to perform tasks consistently and stably across multiple trials, making it highly reliable for real-world applications.

Table 9. Quantitative results for real-robot evaluation on the Fourier GR1. The RMS deviation, standard deviation, and execution consistency reflect the accuracy, stability, and repeatability of the robot's behavior across multiple trials. Position is measured in meters, and orientation is measured in degrees.

Scenario	Consistency		Stability		Repeatability	
	Pos.	Ori.	Pos.	Ori.	Pos.	Ori.
Drinking Water	0.0052	0.97	0.0054	1.1	0.0013	0.29
Playing Guitar	0.0054	1.5	0.0059	1.7	0.0016	0.35
Lifting a Box	0.0050	1.0	0.0056	1.1	0.0014	0.31

5. Conclusion

In this paper, we present HuBE, a morphology-aware framework for human-like behavior execution in humanoid robots. By integrating robot state, goal poses, and situational context, HuBE generates behaviors that are both behaviorally similar and appropriate. We further introduce HPose, a context-enriched dataset with 6D joint pose representations and situational annotations, together with a bone-scaling–based augmentation strategy that enables millimeter-level cross-embodiment compatibility. Extensive experiments on multiple commercial humanoid platforms demonstrate that HuBE consistently outperforms strong baseline methods, highlighting its potential as a transferable foundation for scalable and socially acceptable human-like behavior execution. A limitation of HuBE is that it focuses on expressive upper-body behaviors, while lower-body motions also play a critical role in shaping human actions. Extending HuBE to lower-limb or full-body behaviors requires addressing stability-critical whole-body control, which remains challenging across diverse platforms and motion types. A natural direction for future work is to hierarchically integrate HuBE with lower-level balance or locomotion controllers to enable full-body expressive behaviors. In addition, real-world deployment inevitably involves sensing noise and state estimation errors. Addressing sensing uncertainty and improving execution robustness in real-world environments is another important direction for future work.

Data availability statement

No supplementary or additional data were generated in this study.

Declaration of generative AI and AI assisted technologies

During the preparation of this work, the authors used a large language model, GPT-4o, to generate descriptive annotations for human motion sequences based on the original annotations, as described in the Data Specifications section. The purpose of using GPT-4o was to enrich the situational context associated with the human motion data. The generated annotations were reviewed, checked, and curated by the authors before being used in the study.

Acknowledgments

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.

Authors' contribution

Conceptualization, Shipeng Lyu; methodology, Shipeng Lyu and Fangyuan Wang; data processing, Shipeng Lyu and Weiwei Lin; code programming, Shipeng Lyu, Fangyuan Wang and Luhao Zhu; experiments, Weiwei Lin; writing—original draft preparation, Shipeng Lyu, Fangyuan Wang, Luhao Zhu and Weiwei Lin; writing—review and editing, Guodong Guo and David Navarro-Alarcon; project administration, David Navarro-Alarcon. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

David Navarro-Alarcon holds the position of Associate Editor for *Robot Learning* and has not peer reviewed or made any editorial decisions for this paper.

References

[1] Gulletta G, Silva E, Erlhagen W, Meulenbroek R, Costa M, *et al.* A human-like upper-limb motion planner: generating naturalistic movements for humanoid robots. *Int. J. Adv. Rob. Syst.* 2021, 18(2):1729881421998585.

[2] Hill G. An unwillingness to act: behavioral appropriateness, situational constraint, and self-efficacy in shyness. *J. Pers.* 1989, 57(4):871–890.

[3] Habekost JG, Gäde C, Allgeuer P, Wermter S. Inverse kinematics for neuro-robotic grasping with humanoid embodied agents. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Abu Dhabi, UAE, October 14–18, 2024, pp. 7315–7322.

[4] Rakita D, Mutlu B, Gleicher M. An analysis of RelaxedIK: an optimization-based framework for generating accurate and feasible robot arm motions. *Auton. Rob.* 2020, 44(7):1341–1358.

[5] Zhang H, Li W, Liu J, Chen Z, Cui Y, *et al.* Kinematic motion retargeting via neural latent optimization for learning sign language. *IEEE Rob. Autom. Lett.* 2022, 7(2):4582–4589.

[6] Jang E, Irpan A, Khansari M, Kappler D, Ebert F, *et al.* Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, Auckland, New Zealand, December 14–18, 2022, pp. 991–1002.

[7] Annabi L, Ma Z, Nguyen S. Unsupervised motion retargeting for human-robot imitation. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, Boulder, USA, March 11–15, 2024, pp. 204–208.

[8] He T, Luo Z, Xiao W, Zhang C, Kitani K, *et al.* Learning human-to-humanoid real-time whole-body teleoperation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Abu Dhabi, UAE, October 14–18, 2024, pp. 8944–8951.

[9] Xie Y, Jampani V, Zhong L, Sun D, Jiang H. OmniControl: control any joint at any time for human motion generation. In *The Twelfth International Conference on Learning Representations*, Vienna, Austria, May 7–11, 2024.

[10] Petrovich M, Black M, Varol G. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, October 10–17, 2021, pp. 10985–10995.

[11] Chen X, Jiang B, Liu W, Huang Z, Fu B, *et al.* Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, June 18–22, 2023, pp. 18000–18010.

[12] Plappert M, Mandery C, Asfour T. The kit motion-language dataset. *Big Data* 2016, 4(4):236–252.

[13] Cheng X, Ji Y, Chen J, Yang R, Yang G, *et al.* Expressive whole-body control for humanoid robots. *arXiv* 2024, arXiv:2402.16796.

[14] Dou R, Yu S, Li W, Chen P, Xia P, *et al.* Inverse kinematics for a 7-DOF humanoid robotic arm with joint limit and end pose coupling. *Mech. Mach. Theory* 2022, 169:104637.

[15] Rempe D, Birdal T, Hertzmann A, Yang J, Sridhar S, *et al.* HuMoR: 3D human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, Montreal, Canada, October 11–17, 2021.

[16] Shafir Y, Tevet G, Kapon R, Bermano A. Human motion diffusion as a generative prior. In *The Twelfth International Conference on Learning Representations*, Vienna, Austria, May 7–11, 2024.

[17] Karunratanakul K, Preechakul K, Suwajanakorn S, Tang S. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France, October 1–6, 2023, pp. 2151–2162.

[18] Zhang Q, Cui P, Yan D, Sun J, Duan Y, *et al.* Whole-body humanoid robot locomotion with human reference. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Abu Dhabi, United Arab Emirates, October 14–18, 2024, pp. 11225–11231.

[19] Yan Y, Mascaro EV, Egle T, Lee D. I-CTRL: imitation to control humanoid robots through bounded residual reinforcement learning: a new framework. *IEEE Robot. Autom. Mag.* 2025.

[20] Li J, Zhu Y, Xie Y, Jiang Z, Seo M, *et al.* OKAMI: teaching humanoid robots manipulation skills through single video imitation. In *8th Annual Conference on Robot Learning (CoRL)*, Munich, Germany, November 6–9, 2024.

[21] Fu Z, Zhao Q, Wu Q, Wetzstein G, Finn C. HumanPlus: humanoid shadowing and imitation from humans. In *8th Annual Conference on Robot Learning (CoRL)*, Munich, Germany, November 6–9, 2024.

[22] Loveland K, Pearson D, Tunali-Kotoski B, Ortegon J, Gibbs M. Judgments of social appropriateness by children and adolescents with autism. *J. Autism Dev. Disord.* 2001, 31:367–376.

[23] Gao Y, Yang F, Frisk M, Hemandez D, Peters C, *et al.* 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)*, New Delhi, India, October 14–18, 2019, pp. 1–8.

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

[25] Zhou Y. 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*, Stockholm, Sweden, March 13–16, 2023, pp. 781–783.

[26] Mahmood N, Ghorbani N, Troje N, Pons-Moll G, Black M. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, Republic of Korea, October 27–November 2, 2019, pp. 5442–5451.

[27] Lin J, Zeng A, Lu S, Cai Y, Zhang R, *et al.* Motion-X: a large-scale 3D expressive whole-body human motion dataset. *Adv. Neural Inf. Process. Syst.* 2023, 36:25268–25280.

[28] Sutton R. *Reinforcement learning: An introduction*, 2nd ed. London: MIT Press, 2018.

[29] Torabi F, Warnell G, Stone P. Behavioral cloning from observation. In *Proceedings of the Twenty-Seventh* International Joint Conference on Artificial Intelligence, Stockholm, Sweden, July 13–19, 2018, pp. 4950–4957.

[30] Andrychowicz M, Wolski F, Ray A, Schneider J, Fong R, *et al.* Hindsight experience replay. *arXiv* 2017, arXiv:1707.01495.

[31] Guo W, Du Y, Shen X, Lepetit V, Alameda-Pineda X, *et al.* Back to mlp: a simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, Waikoloa, USA, January 2–7, 2023, pp. 4809–4819.

[32] Zhang M, Cai Z, Pan L, Hong F, Guo X, *et al.* Motiondiffuse: text-driven human motion generation with diffusion model. *IEEE Trans. Pattern Anal. Mach. Intell.* 2024, 46(6):4115–4128.