

PressMimic: Pressure-Guided Motion Capture and Control for Humanoid Robot Imitation

Yi Lu^{*1}, Shenghao Ren^{*1}, Tianyu Xiong¹, Zhaoxiang Li¹, Jiaqi Li¹,
He Zhang³, Tao Yu³, Qiu Shen^{†1,2}, Xun Cao^{1,2}

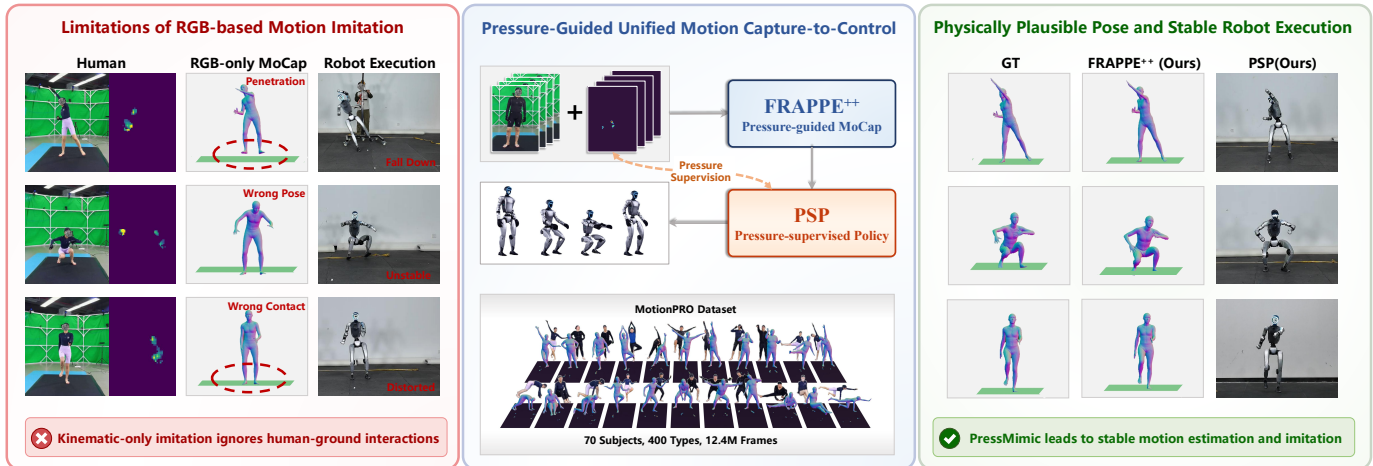


Fig. 1. We propose PressMimic, a unified framework that integrates pressure into both motion capture and motion control for humanoid motion imitation. By fusing RGB with pressure signals in the motion capture module FRAPPE⁺⁺ and supervising robot training via the pressure-guided policy PSP, PressMimic achieves physically grounded pose estimation and stable robot execution.

Abstract—Humanoid motion imitation requires not only accurate perception of human kinematics but also faithful reproduction of physical interactions with the environment. However, existing pipelines rely primarily on vision-based motion capture and kinematic imitation, largely ignoring contact dynamics, leading to artifacts such as foot sliding, floor penetration, and unstable behaviors. In this work, we revisit humanoid motion imitation from the perspective of physical grounding and leverage pressure as a unified modality across perception and control. We present PressMimic, a framework that integrates pressure into the full pipeline from motion capture to humanoid control. In the perception stage, we introduce FRAPPE⁺⁺, a multimodal model that fuses RGB and pressure to jointly estimate 3D pose and global motion, where pressure provides explicit contact and support constraints to resolve ambiguity in vision-based estimation. In the control stage, we propose a pressure-supervised policy (PSP) that incorporates pressure-derived signals into reinforcement learning, enabling physically consistent contact patterns during execution. We further construct MotionPRO, a large-scale dataset with synchronized RGB, pressure, and motion capture data. Experiments show that pressure improves motion estimation accuracy, trajectory consistency, and execution stability. These results demonstrate that pressure serves as an effective physical grounding signal, bridging perception and

control for physically consistent humanoid motion imitation. *Project Page: <https://yeelou.github.io/PressMimic/>*

Index Terms—Human Motion Capture, Dataset, Pressure Sensing, Multimodal Fusion, Humanoid Robot Control, Reinforcement Learning, Embodied Intelligence

I. INTRODUCTION

Enabling humanoid robots to faithfully replicate human motion has long been a central challenge at the intersection of computer vision and robotics. Achieving this goal demands not only accurate perception of human body kinematics, but also robust translation of such perception into physically plausible robot control. Despite remarkable progress in both human motion capture [1]–[10] and humanoid motion control [11]–[13], existing approaches remain agnostic to the physical interactions between the body and its environment, leading to failure modes including temporal jitter, floor penetration, and foot-sliding artifacts. This perceptual deficiency creates a critical bottleneck: robots driven by kinematically-estimated references frequently fail to reproduce the physical grounding of human motion, resulting in unstable locomotion, degraded performance in scenarios with frequent contact changes, and, in extreme cases, hardware damage.

Pressure naturally encodes the foot-ground interactions and contact dynamics that kinematic signals fail to capture. Previous work [14] demonstrated that pressure substantially improves MoCap quality by providing explicit grounding

^{*}Equal contribution. [†]Corresponding author. E-mail: shenqiu@nju.edu.cn
¹School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China.

²Key Laboratory of Optoelectronic Devices and Systems with Extreme Performances of MOE, Nanjing University, Nanjing 210023, China.

³BNRist, Tsinghua University, Beijing 100084, China.

Manuscript received May 24, 2026.

constraints on 3D pose and trajectory estimation. Beyond perception, pressure also offers a direct supervisory signal for robot motion control, constraining contact timing, ground reaction forces, and weight distribution that existing imitation pipelines overlook.

In this work, we present **PressMimic**, a unified framework that leverages pressure as a core modality across the full pipeline of humanoid motion imitation from human motion capture to robot motion control. To the best of our knowledge, PressMimic is the first system to systematically explore and exploit pressure signals in both the perception and control stages of humanoid motion imitation.

The proposed framework operates in two stages. In the first stage, we introduce a multimodal human motion estimation network **FRAPPE⁺⁺** that fuses RGB video with pressure maps. Both modalities are non-invasive and require no wearable instrumentation, yet capture complementary aspects of human motion: RGB encodes visual appearance and geometric structure, while pressure provides explicit contact states and ground reaction dynamics imperceptible to cameras alone. To exploit this complementarity, we extract compact pressure representations and model spatiotemporal dependencies across modalities, fusing them into a unified motion representation. We further impose orthographic similarity constraints along the camera axis and whole-body contact constraints along the vertical axis to combine precise lower-body contact and global translation cues from pressure with accurate full-body pose from RGB. In the second stage, we introduce a Pressure-Supervised motion control Policy (**PSP**), where plantar pressure maps are modeled as distributional offsets characterizing foot contact patterns and incorporated as an auxiliary reward term during reinforcement learning, explicitly encouraging the robot to reproduce the ground reaction dynamics of the human demonstrator.

To support both stages, we introduce **MotionPRO**, a large-scale multimodal dataset capturing pressure, RGB video, and optical motion capture signals from 70 volunteers performing 400 motion types, encompassing 12.4M pose frames in total. Extensive experiments on MotionPRO demonstrate that our multimodal fusion approach outperforms state-of-the-art RGB-based pose and trajectory estimation methods, maintaining physically plausible estimates even under extreme occlusions and severe vertical trajectory drift. For humanoid motion imitation, pressure supervision consistently improves task success rates and locomotion stability while preserving strong motion similarity to the reference demonstrations.

The main contributions of this work are summarized as follows:

- We present **PressMimic**, the first unified framework to systematically exploit pressure across both the perception and control stages of humanoid motion imitation.
- We propose a multimodal motion estimation network **FRAPPE⁺⁺** fusing RGB and pressure, achieving robust pose and trajectory estimation under occlusion and trajectory drift.
- We introduce a pressure-supervised control policy called **PSP** that incorporates pressure-derived distributional off-

sets as an auxiliary reward, improving task success rates and locomotion stability in humanoid motion imitation.

- We construct **MotionPRO**, a large-scale human motion dataset with pressure, RGB, and optical sensors. Extensive experiments demonstrate that our methods can achieve significant performance improvements in both motion estimation and motion imitation tasks.

II. RELATED WORK

A. Humanoid Robot Motion Imitation

Humanoid robot motion imitation aims to enable robots to reproduce human motion in a physically plausible and stable manner. Early approaches [15]–[22] relied on motion retargeting to directly map human joint configurations to robot joint angles, followed by trajectory optimization or model predictive control to generate feasible robot motions. While effective for simple periodic motions such as walking, these methods struggle to generalize to complex or contact-rich behaviors due to their reliance on accurate dynamic models and hand-crafted controllers.

Reinforcement learning (RL) has emerged as the dominant paradigm for humanoid motion imitation, enabling controllers to be learned directly from interaction with physics simulators. DeepMimic [23] pioneered the use of reference-state initialization and motion imitation rewards to train physics-based controllers capable of reproducing a wide range of human motions. Subsequent works extended this framework to handle more diverse motion repertoires [24], [25], multi-skill transitions [26], [27], and real-world deployment on physical robots [28]–[35]. PHC [24] introduced a perpetual humanoid controller that tracks arbitrary motion sequences through curriculum-based training. More recently, BeyondMimic [36] and SONIC [30] have advanced whole-body motion imitation on full-sized humanoid platforms, demonstrating stable policy deployment on real hardware.

Despite these advances, existing methods construct their imitation rewards purely from kinematic references, including joint positions, orientations, and end-effector states, without accounting for the dynamic foot-ground interactions that underpin stable locomotion. This limitation frequently manifests as foot-sliding, floor penetration, and unstable gait patterns, particularly in contact-rich scenarios. In contrast, PressMimic incorporates plantar pressure as an explicit contact supervision signal, encouraging the robot to reproduce not only the kinematic trajectory but also the underlying contact dynamics of the human demonstrator.

B. Non-intrusive Human Motion Capture

Robotic imitation learning relies heavily on human motion capture to acquire high-quality demonstrations of human behaviors. Traditional MoCap systems, such as marker-based optical setups [37], as well as approaches based on wearable IMUs [9], [38], VR tracking devices [39], can provide precise measurements but typically require markers or sensors to be attached to the human body. This not only increases system complexity and cost but also interferes with natural human motion. To address these limitations, non-intrusive human pose

estimation has emerged as a promising alternative. It aims to recover 3D human pose and motion trajectories from external observations without requiring any markers or on-body sensors. Compared with conventional MoCap systems, non-intrusive approaches offer advantages such as flexible deployment, lower cost, and minimal interference with users, making them increasingly attractive for robotic applications [31], [33], [40], [41].

Existing non-intrusive motion capture methods mainly rely on visual perception based on RGB inputs, early approaches typically regress 3D joint positions or parametric human body models from single RGB images [1]–[7]. With the advancement of deep learning, an increasing number of methods utilize the perspectives of temporal information [42]–[46], prior knowledge of human body [1], [47]–[49], and precise camera model [4], [50]–[52] to obtain accurate human body pose. However, purely vision-based methods remain inherently ill-posed due to depth ambiguity and occlusion, which often lead to global scale inconsistency and physical artifacts such as foot skating or floor penetration. Moreover, they struggle to infer human–environment interactions, especially contact states, which are critical for physically plausible motion estimation [53]–[58].

Ground pressure sensors can directly capture the contact states and force distributions between the human body and the environment, offering unique advantages in modeling foot contact, support patterns, and center-of-mass dynamics. Early works reconstruct human pose from single-frame pressure data, mainly focusing on lying scenarios [59]–[65]. Moreover, auxiliary information derived from pressure data, such as the center of pressure (CoP), is used to constrain pose estimation and improve physical plausibility [2], [66]. Despite these advances, existing pressure-based methods still suffer from several limitations. On the one hand, CoP-based methods only work well for slow and steady movements. They fail in fast actions because they don’t account for the inertia that comes with rapid motion. [67], [68]. On the other hand, most approaches focus on foot contact and lack modeling of contact interactions involving other body parts [69], [70]. Therefore, effectively leveraging the physical priors provided by pressure signals to develop more robust and physically plausible non-intrusive human motion capture methods remains a critical challenge.

III. METHODS

A. Pressure-Guided Humanoid Motion Imitation

As illustrated in Fig. 2, the proposed PressMimic framework takes synchronized RGB video and pressure sequences as input and produces humanoid robot motion as output, operating through three successive modules: motion estimation, motion retargeting, and motion control. Pressure signals serve as a core modality throughout the pipeline, enhancing both the accuracy of human motion estimation and the physical plausibility of robot motion control.

The humanoid motion imitation task can be formulated as an optimization problem over human motion parameters and

robot joint configurations:

$$\begin{aligned} \min_{\theta, T, \mathcal{A}} \quad & \mathcal{L}_H(\mathcal{S}_H(\theta, T), \tilde{\mathcal{S}}_H) + \mathcal{L}_R(\mathcal{S}_R(\mathcal{A}), \tilde{\mathcal{S}}_R(\tilde{\mathcal{A}})) \\ \text{s.t.} \quad & \mathcal{S}_R = \mathcal{M}(\mathcal{S}_H) \\ & \tilde{\mathcal{S}}_R \subseteq C_R \end{aligned} \quad (1)$$

where $\mathcal{S}_H(\theta, T)$ denotes the full state representation of the human body model parameterized by pose θ and global translation T . $\tilde{\mathcal{S}}_H$ denotes the corresponding ground-truth human state. \mathcal{L}_H is a composite loss function measuring the discrepancy between estimated and ground-truth human states across multiple terms including pose, joint positions and foot contact states. $\mathcal{S}_R(\mathcal{A})$ denotes the reference robot state with action \mathcal{A} , and $\tilde{\mathcal{S}}_R(\tilde{\mathcal{A}})$ denotes the actually executed robot state with action $\tilde{\mathcal{A}}$. \mathcal{L}_R is a composite distance function over the robot state space, including body positions, orientations, velocities, and angular velocities. $\mathcal{M}(\cdot)$ denotes the mapping from the human body model to the humanoid robot model, and C_R denotes the physical feasibility constraints of the robot, including joint angle limits, velocity bounds, and stability conditions.

The first term of the objective minimizes the discrepancy between the estimated human body joints and their ground-truth counterparts, which is addressed in the motion estimation stage (Sec. III-B). The second term minimizes the gap between the reference robot joints, retargeted from the estimated human motion via $\mathcal{M}(\cdot)$, and the actually executed robot joints, which is addressed in the motion control stage (Sec. III-C). The constraint $\tilde{\mathcal{S}}_R(\tilde{\mathcal{A}}) \subseteq C_R$ ensures that the executed robot motion remains within the physical feasibility bounds of the robot throughout imitation. Pressure signals are incorporated into both stages as indicated by the orange dashed arrow in Fig. 2, they provide grounding constraints during motion estimation and serve as auxiliary supervision during motion control.

B. Human Motion Estimation by Fusing Pressure and RGB

In the first stage, we present a pressure-guided motion estimation framework FRAPPE⁺⁺ that integrates pressure representation learning, temporal context modeling, and cross-modal feature fusion to achieve robust estimation of human 3D pose and global motion trajectories, as illustrated in Fig. 3.

1) *Pressure-Guided Motion Estimation Pipeline*: Given a synchronized RGB video sequence and a corresponding pressure distribution sequence, our goal is to recover the human 3D pose parameters θ and the global translation trajectory T from multimodal observations. First, the input RGB images are processed by a visual encoder to extract frame-wise visual features that represent human appearance and pose structure. Meanwhile, the two-dimensional pressure distribution maps are fed into the pressure encoder, referred to as **Sparse Pressure Encoder (SPE)**. This module converts the sparse pressure distribution into a structured one-dimensional sequence representation and extracts corresponding pressure features. The resulting pressure and visual feature sequences are then fed into the **Temporal Context Aggregation Module (TCAM)** for temporal context modeling, capturing temporal

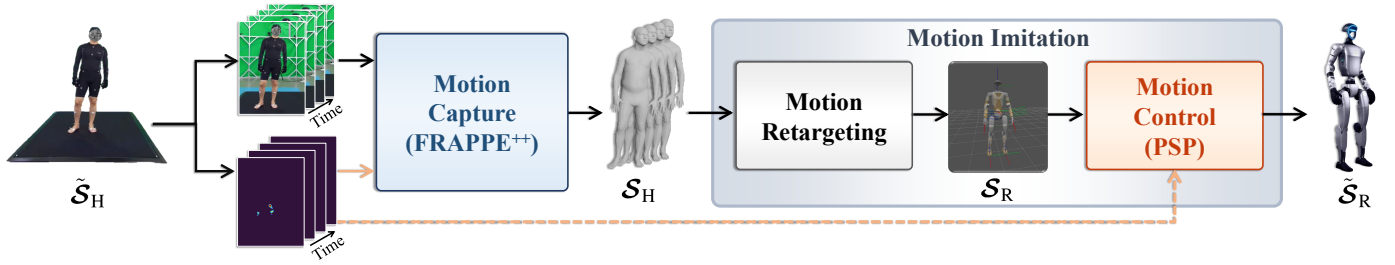


Fig. 2. **Overview of PressMimic.** Given synchronized RGB video and pressure sequences, PressMimic estimates human body motion, retargets it to the humanoid robot, and executes it via a learned control policy. Pressure signals enhance motion estimation accuracy and provide auxiliary contact supervision for the control policy, as indicated by the orange dashed arrow.

dependencies within the current window while incorporating contextual information from neighboring time windows. After temporal modeling, we introduce the **Fusion Cross-Attention Module (FCAM)** to enable cross-modal feature interaction. Finally, the temporally modeled features are fed into a regression network to predict the human pose parameters θ and the global translation T . The predicted parameters are further passed through a parametric human body model to compute the corresponding three-dimensional joint positions.

2) *Training Objective:* To achieve accurate pose reconstruction and physically consistent motion generation, we design a set of joint loss functions to simultaneously constrain the pose parameters, spatial joint positions, and global motion trajectories. The overall training objective is defined as

$$\begin{aligned} \mathcal{L}_H = & \lambda_{\text{pose}} \mathcal{L}_{\text{pose}} + \lambda_{3d} \mathcal{L}_{3d} + \lambda_{2d} \mathcal{L}_{2d} \\ & + \lambda_{\text{trans}} \mathcal{L}_{\text{trans}} + \lambda_{\text{contact}} \mathcal{L}_{\text{contact}}, \end{aligned} \quad (2)$$

where λ_{pose} , λ_{3d} , λ_{2d} , λ_{trans} , and λ_{contact} are corresponding weights. The loss of pose parameters $\mathcal{L}_{\text{pose}}$ is the mean squared error between the predicted and ground-truth pose parameters. The 3D joint loss, \mathcal{L}_{3d} , is the mean squared error between the predicted and ground-truth joint positions, after pelvis alignment. Global translation loss $\mathcal{L}_{\text{trans}}$ is the mean squared error between predicted and ground truth translation. The ground contact loss, $\mathcal{L}_{\text{contact}}$, is the mean squared error between the predicted and the ground-truth global joints that are in contact with the ground.

3) *Sparse Pressure Encoder (SPE):* In human pose estimation, ground pressure maps are highly sparse, making direct convolution over the full 2D map computationally inefficient and prone to missing fine-grained contact structures. To address this, we propose the Sparse Pressure Encoder (SPE), which converts the 2D pressure distribution into a compact 1D sequence while preserving spatial topology. Specifically, a Hilbert space-filling curve serializes the valid pressure points, maintaining local neighborhood relationships during the 2D-to-1D mapping. Each point is represented by its spatial coordinates and pressure value, encoded via Fourier sine-cosine embeddings to capture multi-scale spatial variations. The embedded features are projected by an MLP and refined through an attention mechanism that models relationships among all pressure points, yielding a fixed-length pressure feature vector for subsequent cross-modal fusion.

4) *Temporal Context Aggregation Module (TCAM):* Accurate human motion estimation requires capturing temporal

dependencies across long sequences. In practice, video sequences are often divided into fixed-length temporal windows for processing, which may introduce temporal discontinuities at window boundaries. To ensure the stability of motion estimation while managing computational overhead, TCAM adopts a modeling strategy that operates across two distinct scales. Within each window, we integrate Gated Recurrent Units with self-attention mechanisms to capture temporal dependencies that span short durations. Outside the current window, neighboring windows are compressed into compact representations through temporal average pooling. These representations can be introduced as augmented context via cross attention, effectively extending the field of reception for the model.

We divide the input sequence into a series of temporal windows and denote the current window as $\mathbf{x}_t = \{\mathbf{x}_t^j\}_{j=1}^L$, where L is the number of frames per window, $\mathbf{x}_t^j \in \mathbb{R}^D$ is the feature vector of the j -th frame and D is the feature dimension. To extend the temporal receptive field, we incorporate N neighboring windows on each side of the current window. Each neighboring window \mathbf{x}_{t+i} ($i \in [-N, N]$, $i \neq 0$) is compressed into a global representation $\mathbf{e}_{t+i} \in \mathbb{R}^D$ via temporal average pooling:

$$\mathbf{e}_{t+i} = \frac{1}{L} \sum_{j=1}^L \mathbf{x}_{t+i}^j. \quad (3)$$

The compressed context representations are then concatenated with the current window to form an extended feature sequence:

$$\mathbf{f}_t = [\mathbf{e}_{t-N}, \dots, \mathbf{e}_{t-1}, \mathbf{x}_t, \mathbf{e}_{t+1}, \dots, \mathbf{e}_{t+N}]. \quad (4)$$

Given $\mathbf{f}_t \in \mathbb{R}^{(L+2N) \times D}$, we apply a cross-attention mechanism where the current window \mathbf{x}_t serves as the query source and the extended sequence \mathbf{f}_t provides the keys and values. Specifically, the query \mathbf{Q}_{TCAM} , key \mathbf{K}_{TCAM} , and value \mathbf{V}_{TCAM} are computed as:

$$\begin{aligned} \mathbf{Q}_{\text{TCAM}} &= \mathbf{W}_q \mathbf{x}_t, \\ \mathbf{K}_{\text{TCAM}} &= \mathbf{W}_k \mathbf{f}_t, \\ \mathbf{V}_{\text{TCAM}} &= \mathbf{W}_v \mathbf{f}_t. \end{aligned} \quad (5)$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ are learnable projection matrices. The attention output is computed as:

$$\mathbf{f}_{\text{TCAM}} = \text{Attention}(\mathbf{Q}_{\text{TCAM}}, \mathbf{K}_{\text{TCAM}}, \mathbf{V}_{\text{TCAM}}). \quad (6)$$

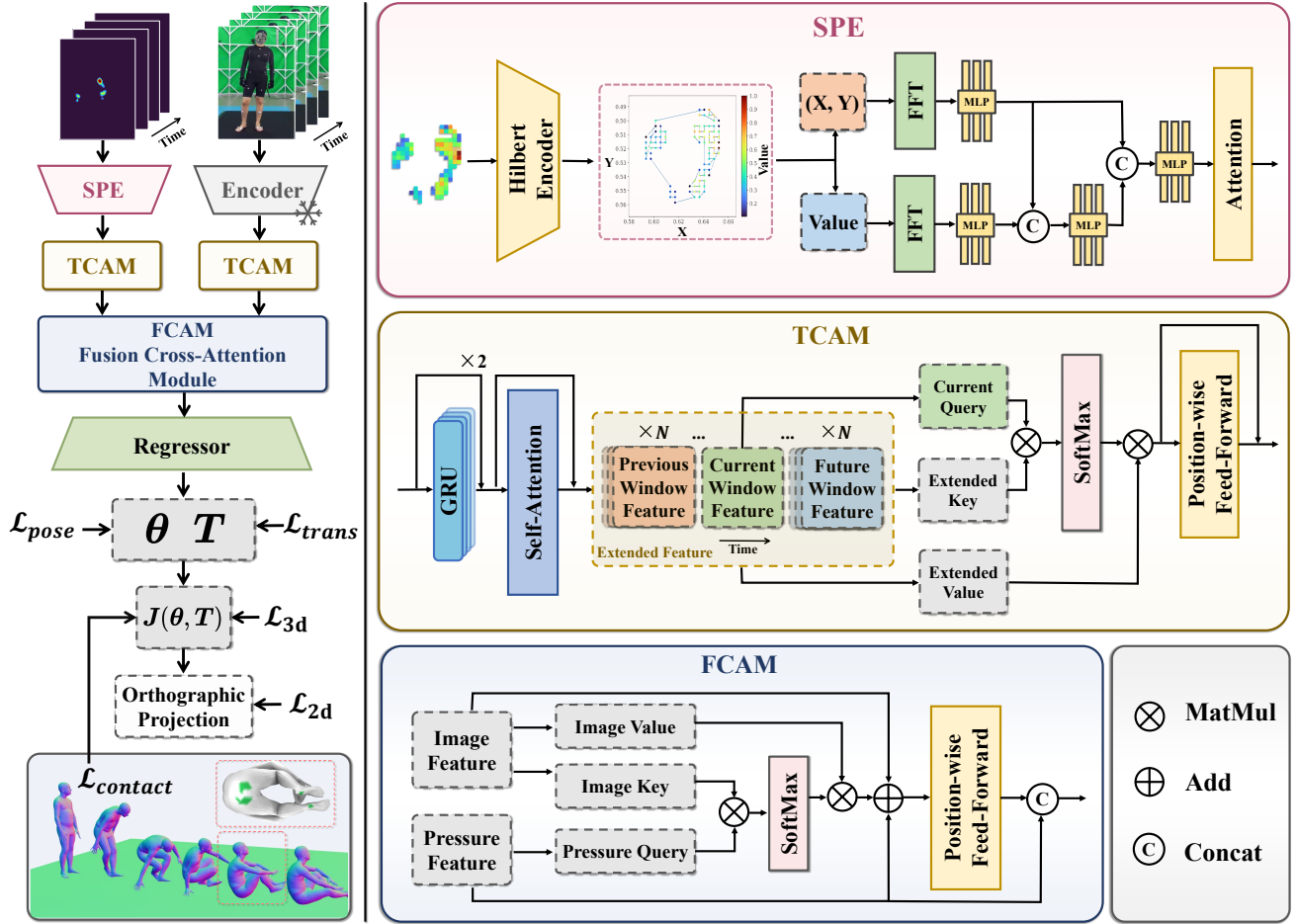


Fig. 3. **The framework of FRAPPE⁺⁺.** Pressure and RGB video are processed by the SPE and image encoder respectively, followed by TCAM to model spatiotemporal dependencies within each modality. FCAM then fuses the two modalities via cross-attention, where image features serve as key and value while pressure features serve as query. The fused representation is fed into a regressor to estimate SMPL parameters.

This allows each frame in x_t to attend to both fine-grained frame-level features within the current window and compressed global representations from neighboring windows, improving the temporal coherence of the predicted motion.

5) *Fusion Cross-Attention Module (FCAM)*: Visual information provides cues about human appearance and pose structure, whereas pressure signals capture contact states between the body and the ground. To exploit their complementary properties, FCAM performs cross-modal feature fusion. Let f_{rgb} and f_p denote the output features of the visual encoder and the SPE module, respectively. Pressure features serve as the query source to actively probe the visual representation, while visual features provide the keys and values:

$$\begin{aligned} Q_{\text{FCAM}} &= W_q f_p, \\ K_{\text{FCAM}} &= W_k f_{\text{rgb}}, \\ V_{\text{FCAM}} &= W_v f_{\text{rgb}}. \end{aligned} \quad (7)$$

The fused representation is obtained by:

$$f_{\text{FCAM}} = \text{Attention}(Q_{\text{FCAM}}, K_{\text{FCAM}}, V_{\text{FCAM}}). \quad (8)$$

f_{FCAM} is subsequently passed through a position-wise feed-forward network with residual connections and concatenated with f_p to form the final motion representation. Through this

design, pressure features guide visual attention toward human-ground contact regions, improving the accuracy and physical plausibility of motion estimation.

C. Humanoid Motion Control under Pressure Supervision

Building upon the pressure-guided motion estimation described above, we further leverage plantar pressure as an explicit supervisory signal to improve the physical consistency of humanoid motion control called PSP (Pressure-Supervised Policy), as illustrated in Fig. 4.

1) *Pressure Modeling*: To incorporate plantar pressure into the control pipeline, we first extract per-foot pressure intensity from the raw pressure maps. Taking the left foot as an example, the total pressure of the left foot ϕ_{lf} is computed by summing the pressure values of all pixels belonging to the left foot region:

$$\phi_{\text{lf}} = \sum_{i \in \text{lf}} \phi_i, \quad (9)$$

where ϕ_i denotes the pressure value of the i -th pixel. The right foot pressure ϕ_{rf} is computed analogously.

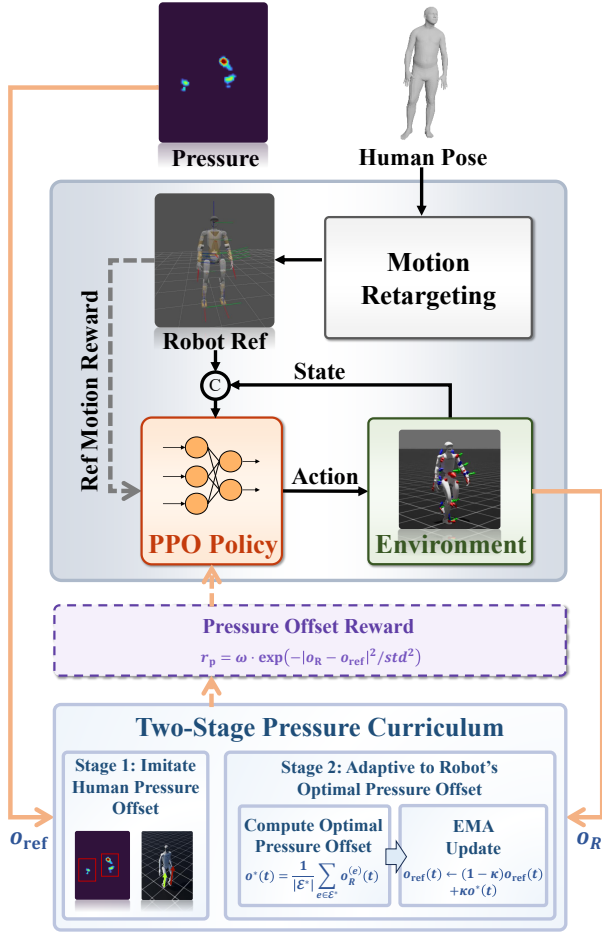


Fig. 4. **Humanoid motion imitation under pressure-supervised policy (PSP).** Human pose and plantar pressure serve as dual inputs: the pose is retargeted to robot reference motion and concatenated with the current state as policy input, while pressure drives a two-stage curriculum that first imitates the human pressure offset and then adapts to the robot’s own optimal contact distribution via EMA updates.

Building upon prior work [71], we further define a pressure offset o to characterize the bilateral weight distribution between the left and right feet:

$$o = \frac{\phi_{rf} + k \cdot \epsilon}{\phi_{lf} + \phi_{rf} + \epsilon}, \quad (10)$$

where $\epsilon > 0$ is a small constant for numerical stability. k is a scalar that we set to 2 in practice, chosen to place the airborne signature well outside the single-foot contact range $(0, 1)$ and far beyond any value reachable through normal load variation between the two feet. Because ϵ is negligibly small relative to any non-zero foot pressure, the numerical behaviour of o is determined entirely by the foot pressure magnitudes in all contact cases, and no ambiguity arises at the boundary between contact states. This formulation provides an interpretable scalar representation of foot contact state. Specifically, when both feet are in contact ($\phi_{lf} \neq 0$ and $\phi_{rf} \neq 0$), o lies within $(0, 1)$, reflecting the relative bilateral load distribution. When only the right foot is in contact ($\phi_{lf} = 0$ and $\phi_{rf} \neq 0$), o approaches 1. When only the left foot is in contact ($\phi_{lf} \neq 0$ and $\phi_{rf} = 0$), o approaches 0. When both

feet are airborne ($\phi_{lf} = 0$ and $\phi_{rf} = 0$), $o = k = 2$, providing a distinctive out-of-contact signature that is unambiguously separable from all single-foot and double-foot contact cases.

2) *Policy Training with Pressure Supervision:* As illustrated in Fig. 4, the motion control stage follows a reinforcement learning (RL) framework built upon BeyondMimic [36], trained with the Proximal Policy Optimization (PPO) algorithm [72]. The estimated human pose is first mapped to robot reference motion via motion retargeting module [73], [74], which is concatenated with the current robot state as the policy input. The policy outputs joint-level actions that are executed in a physics simulator, and the resulting state is fed back to close the control loop.

The reward function comprises standard kinematic imitation terms adopted in prior works, including body joint position, joint angle, and end-effector rewards, supplemented by a pressure offset reward. To compute the pressure offset reward, we collect the ground reaction forces at the robot’s feet from the simulator in real time, compute the robot’s pressure offset o_R following Eq. 10, and compare it against the reference pressure offset o_{ref} derived from the human demonstrator’s plantar pressure measurements. The pressure offset reward is defined as:

$$r_p = \omega \cdot \exp\left(-|o_R - o_{ref}|^2 / std^2\right), \quad (11)$$

where o_R and o_{ref} denote the pressure offsets computed from the robot’s ground reaction forces and the human demonstrator’s plantar pressure measurements, respectively, following Eq. 10. ω is a weighting coefficient that scales the contribution of the pressure reward relative to other reward terms, and std controls the sensitivity of the reward to deviations in pressure offset. This Gaussian-shaped reward provides dense and smooth supervision: it yields a high reward when the robot’s bilateral weight distribution closely matches that of the human demonstrator, and decays gracefully as the discrepancy grows, avoiding abrupt reward discontinuities that could destabilize policy training.

To further improve training stability and adaptability, we introduce a two-stage pressure curriculum. In the first stage, the policy is trained with the fixed human reference pressure o_{ref} . Once the mean episode reward exceeds a preset threshold, the second stage is triggered, in which the reference pressure is dynamically updated to reflect the robot’s own optimal contact distribution rather than the human demonstrator’s. At each simulation step, the per-frame pressure offset o_R is recorded across all parallel environments indexed by motion frame t . At the end of each episode, environments are ranked by total reward and the top n -th percentile is selected as the set of optimal environments \mathcal{E}^* , filtering out low-quality episodes that could corrupt the reference. The optimal pressure curve is obtained by averaging the per-frame pressure data across \mathcal{E}^* :

$$o^*(t) = \frac{1}{|\mathcal{E}^*|} \sum_{e \in \mathcal{E}^*} o_R^{(e)}(t), \quad (12)$$

where frames not covered by any episode in \mathcal{E}^* are excluded from the update. The reference pressure is then updated via

Methods	MPJPE ↓	PMPJPE ↓	PVE ↓	Accel ↓
VIBE [42]	59.7	40.9	82.9	19.6
CLIFF [4]	54.7	39.7	68.6	24.3
SMPLer-X [77]	51.6	32.8	72.4	437.3
TRACE [56]	61.4	43.2	81.4	14.6
PhysPT [78]	56.4	38.7	72.6	3.0
WHAM [57]	160.4	28.3	227.5	2.9
GVHMR [58]	65.0	43.5	71.5	1.6
FRAPPE [14]	41.8	30.2	58.6	3.0
FRAPPE⁺⁺	33.2	24.7	47.5	1.5

TABLE I
EVALUATION OF GLOBAL POSE ESTIMATION ON MOTIONPRO.

Methods	WMPJPE ↓	WAMPJPE ↓	RTE ↓	Jitter ↓	WBCE ↓
TRACE	141.2	92.5	1193.0	68.6	10272.4
WHAM	75.6	50.2	1023.0	9.2	1217.6
GVHMR	75.0	57.1	507.1	1.8	458.3
FRAPPE	60.8	44.6	41.6	6.0	110.2
FRAPPE⁺⁺	45.8	32.4	25.9	2.2	98.8

TABLE II
EVALUATION OF GLOBAL TRAJECTORY ON MOTIONPRO.

alone. This leads to reduced drift and more accurate long-term motion reconstruction.

In addition, our method produces smoother motion sequences with improved physical plausibility, as reflected by lower Accel, Jitter and WBCE in Tab. I and Tab. II. Such improvements are crucial for downstream applications such as humanoid motion imitation, where stability and physical plausibility are essential.

Notably, even the basic pressure fusion variant (FRAPPE) already outperforms most RGB-based baselines, while the full model (FRAPPE⁺⁺) further improves performance across all aspects. This demonstrates that pressure information serves as a strong physical prior and can be more effectively leveraged through proper feature representation and temporal modeling.

3) *Ablation Study*: We conduct ablation studies on MotionPRO to analyze the contribution of each key component. The results are shown in Tab. III.

Removing the pressure modality leads to a significant performance drop across all metrics. The degradation is particularly pronounced in trajectory estimation and lower-body pose accuracy, indicating that pressure signals provide essential information about contact and support. This is especially important for resolving foot-ground interactions, which are difficult to infer from visual inputs alone.

Temporal modeling also plays a critical role in motion estimation. Removing TCAM leads to degradation in both pose and trajectory estimation accuracy, along with noticeable increases in jitter and acceleration errors. This indicates that modeling temporal dependencies not only improves motion estimation performance, but is also essential for maintaining motion smoothness and consistency.

Pressure encoding mainly affects trajectory estimation. When SPE is removed, trajectory-related metrics degrade noticeably, while pose accuracy shows a slight improvement. This suggests a trade-off between learning strong pressure representations and optimizing pose estimation.

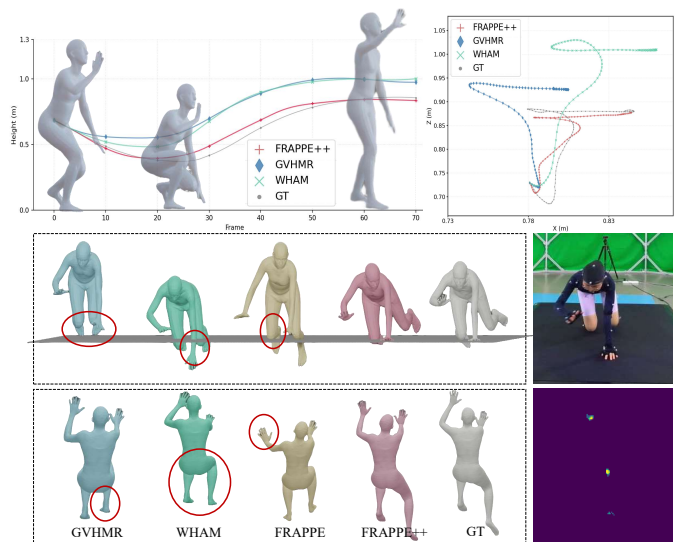


Fig. 7. Qualitative comparison with methods for human pose estimation.

C. Pressure-Guided Humanoid Motion Imitation

1) *Experiment Setup*: We deploy and evaluate our method on the Unitree G1 humanoid robot. To ensure a fair and consistent evaluation, all control policies are trained with the same number of parallel environments (4,096) and transferred to MuJoCo for unified assessment, as well as to the real Unitree G1 platform without any fine-tuning, enabling cross-simulator and sim-to-real transferability analysis. Each baseline is trained following its original recommended configuration until convergence.

2) *Metrics*: To evaluate the performance of humanoid motion imitation, we adopt the following metrics.

Success Rate (SR) measures the proportion of successful trials across all motion sequences, where each sequence is executed for 20 independent trials. A trial is considered successful if the robot does not fall during execution and the positional deviation between its root joint and the reference trajectory remains below 0.85 m throughout. **Motion Completion (MC)** computes the ratio of successfully completed frames to the total number of reference frames. These two metrics jointly assess the stability of the control policy.

Foot Contact Accuracy (FCA) measures the percentage of frames in which the robot’s foot contact states match the reference contact labels derived from pressure measurements recorded during human demonstration.

To evaluate end-to-end trajectory-level accuracy, we use **Root Position Error (RPE, m)** and **Root Velocity Error (RVE, m/s)** to quantify the mean Euclidean distance and velocity difference at the robot’s root joint during execution, respectively, where RPE is computed after aligning the executed and reference trajectories at the first frame to remove initial positional offset. **World-frame Mean Per-Joint Position Error (WMPJPE, m)** measures joint-level positional deviations in the world coordinate frame, similarly computed after first-frame alignment to remove initial positional offset.

To evaluate the similarity between the robot’s local pose and that of the human demonstrator, we adopt **Mean Joint Angle Error (MJAE, rad)** to measure joint angular deviation, and

Models	MPJPE ↓	PMPJPE ↓	LPM.	GTraj ↓	GMPJPE ↓	Jitter ↓↓	Accel ↓	FS ↓
w/o Pressure	65.3	49.4	46.2	209.6	213.7	4.3	2.3	5.4
w/o SPE	31.9	24.0	23.5	64.0	64.0	2.4	1.5	2.1
w/o TCAM	39.0	27.1	26.8	67.5	70.6	6.3	2.4	3.3
Full Model (Ours)	33.2	24.7	24.4	61.3	64.0	2.2	1.5	2.1

TABLE III
ABLATION STUDY ON MOTIONPRO.

Reference	Control Policy	SR ↑	MC ↑	FCA ↑	RPE ↓	RVE ↓	WMPJPE ↓	MJAE ↓	PMPJPE ↓
Optical MoCap	BeyondMimic	0.9278	0.9727	0.7008	0.2706	0.2594	0.5092	0.5533	0.0830
WHAM	BeyondMimic	0.0333	0.0616	0.0244	1.5763	9.1591	6.9643	3.3961	1.7045
GVHMR	BeyondMimic	0.2278	0.4644	0.2885	0.5286	0.8417	1.0913	0.9464	0.1996
FRAPPE ⁺⁺ (Ours)	BeyondMimic	0.9056	0.9746	0.6883	0.2150	0.2295	0.4796	0.4766	0.0875

TABLE IV
QUANTITATIVE EVALUATION OF THE IMPACT OF HUMAN MoCAP QUALITY ON MOTION IMITATION PERFORMANCE.

Reference	Control Policy	SR ↑	MC ↑	FCA ↑	RPE ↓	RVE ↓	WMPJPE ↓	MJAE ↓	PMPJPE ↓
Optical MoCap	PBHC	0.0381	0.1710	0.0829	1.0474	3.300	1.1561	1.7099	0.7000
Optical MoCap	BeyondMimic	0.9278	0.9727	0.7008	0.2706	0.2594	0.5092	0.5533	0.0830
Optical MoCap	PSP (Ours)	0.9222	0.9725	0.7262	0.2592	0.2486	0.4950	0.5534	0.0854
GVHMR	PBHC	0.0452	0.1727	0.0880	1.0301	3.2131	1.1210	1.7336	0.7041
GVHMR	BeyondMimic	0.2278	0.4644	0.2885	0.5286	0.8417	1.0913	0.9464	0.1996
GVHMR	PSP (Ours)	0.5778	0.7142	0.4726	0.3988	0.5000	0.7415	0.6606	0.1395
FRAPPE ⁺⁺ (Ours)	PBHC	0.0667	0.1644	0.0940	1.1107	3.3978	1.1807	1.8327	0.7263
FRAPPE ⁺⁺ (Ours)	BeyondMimic	0.9056	0.9746	0.6883	0.2150	0.2295	0.4796	0.4766	0.0875
FRAPPE ⁺⁺ (Ours)	PSP (Ours)	0.9444	0.9846	0.6918	0.2120	0.2144	0.4737	0.4756	0.0873

TABLE V
QUANTITATIVE EVALUATION OF THE IMPACT OF MOTION CONTROL ON MOTION IMITATION PERFORMANCE.

Procrustes-aligned Mean Per-Joint Position Error (**PMPJPE**, m) to evaluate local joint position similarity after removing global translational, rotational, and scale discrepancies via Procrustes alignment.

Each metric is averaged across all 20 trials per motion sequence. Since methods with lower MC only execute a fraction of the motion sequence, their error metrics are computed over fewer and typically easier frames, leading to artificially low values. To correct for this bias, all error metrics (RPE, RVE, WMPJPE, MJAE, and PMPJPE) are normalized by MC before comparison.

3) *Effect of Reference Motion Quality*: Tab. IV presents quantitative results of humanoid motion imitation under different reference motion sources. We selected BeyondMimic [36], a baseline method with good control performance, as the control policy.

We find that reference quality directly governs imitation performance. Optical MoCap and FRAPPE⁺⁺ achieve strong SR of 0.9278 and 0.9056, MC of 0.9727 and 0.9746, and FCA of 0.7008 and 0.6883 respectively, while GVHMR [58] degrades substantially (SR: 0.2278) and WHAM [57] nearly collapses (SR: 0.0333). After MC normalization, WHAM exhibits severely inflated error metrics (RPE: 1.5763, RVE: 9.1591, WMPJPE: 6.9643), reflecting early termination rather than genuine motion execution. Notably, FRAPPE⁺⁺ achieves

comparable performance to Optical MoCap across all metrics, demonstrating that our RGB and pressure fusion-based mocap method produces reference motions of sufficient quality to serve as a viable alternative to optical motion capture. As shown in Fig. 8, GVHMR and WHAM references introduce severe kinematic distortions and incorrect foot contact in real-world execution, leading to postural instability and even falls, whereas FRAPPE⁺⁺ enables the robot to maintain stable locomotion and accurate posture throughout execution.

The lower RPE, RVE, WMPJPE, and MJAE of FRAPPE⁺⁺ compared to Optical MoCap indicate that the control policy tracks the FRAPPE⁺⁺ reference more accurately. We attribute this to the inherently smoother kinematics of the FRAPPE⁺⁺ reference: the RGB and pressure fusion-based reconstruction attenuates high-frequency motion details, yielding reference trajectories that better conform to the robot’s dynamic constraints and are therefore more amenable to precise tracking.

4) *Effect of Pressure-Supervised Control*: Tab. V presents quantitative results under three control policies across all reference sources. As WHAM-based references consistently fail to produce stable locomotion across all control policies (SR < 0.05), we exclude it from further analysis and focus on the remaining reference sources. PBHC [32] fails to produce stable locomotion regardless of reference quality, with SR below 0.07 in all conditions, indicating that a capable base



Fig. 8. **Real-world humanoid motion imitation results.** Rows show execution results under different reference sources and control policies. Low-quality references (GVHMR, WHAM) lead to twisted postures and instability, while FRAPPE⁺⁺ with BeyondMimic achieves stable but conservative foot placement. FRAPPE⁺⁺ with PSP (Ours) produces stable execution with more natural foot contact dynamics (white circles), validating sim-to-real transfer of pressure supervision.

policy is a prerequisite for effective motion imitation. PSP consistently outperforms BeyondMimic across all estimated reference sources (GVHMR and FRAPPE⁺⁺), demonstrating that pressure supervision provides effective stabilizing contact constraints throughout the motion imitation pipeline.

Under the FRAPPE⁺⁺ reference, PSP achieves the best overall performance: SR improves from 0.9056 to 0.9444 and MC from 0.9746 to 0.9846, marginally surpassing the Optical MoCap-based PSP counterpart (SR: 0.9222, MC: 0.9725). We attribute this to the unified pressure-guided design that runs through both the motion capture and imitation stages: the pressure signal shapes the FRAPPE⁺⁺ reference with physically grounded contact constraints during capture, and PSP reinforces the same contact dynamics during policy training, resulting in natural alignment between the two stages and leading to more complete and stable execution.

Under the Optical MoCap reference, PSP yields a slight decrease in SR and MC compared to BeyondMimic, while improving FCA and all error metrics. We attribute the marginal

SR drop to a cross-modal misalignment between the optical motion capture system and the pressure sensing system. When both references are of exceptionally high fidelity, subtle spatial and temporal misalignments between the two modalities introduce conflicting supervision signals during policy training, occasionally destabilizing execution near support leg transitions. This phenomenon is unique to the Optical MoCap setting. Under FRAPPE⁺⁺, the reference inherently carries pressure-informed contact constraints through the fusion process, resulting in better natural alignment between the two modalities and allowing PSP to consistently improve performance.

Real-world results further corroborate these findings: as shown in Fig. 8, FRAPPE⁺⁺ with BeyondMimic achieves stable execution but exhibits conservative foot placement that deviates from the reference contact pattern, whereas FRAPPE⁺⁺ with PSP produces more natural foot contact dynamics and weight transfer, as highlighted by the white circles.

V. CONCLUSION

In this paper, we investigate the role of pressure as a physical grounding signal for humanoid motion imitation and presented PressMimic, a unified framework that integrates pressure into both motion perception and control. By augmenting vision-based motion capture with pressure, our approach improves the accuracy and physical plausibility of whole-body human motion estimation. Furthermore, by incorporating pressure-derived supervision into policy learning, we enable humanoid robots to better reproduce contact dynamics, leading to more stable and reliable execution.

Our study highlights that pressure provides complementary information to visual inputs by explicitly encoding human–environment interactions, which are otherwise difficult to infer from images alone. This additional modality not only reduces ambiguity in motion estimation but also serves as an effective supervisory signal for control, bridging the long-standing gap between kinematic imitation and physically consistent execution.

Despite these advances, several limitations remain. First, the pressure signals are currently limited to ground interactions and may not fully capture complex multi-contact scenarios involving interactions with external objects. Second, the integration of pressure in robot control is still mainly applied to bipedal contact-related scenarios, leaving room for more rigorous whole-body contact dynamics modeling, mapping, and learning. Future work will explore richer physical representations, improved generalization beyond instrumented environments, and tighter coupling between perception and control. We believe that incorporating physical sensing modalities such as pressure is a crucial step toward achieving robust and physically grounded embodied intelligence.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant U25B2046 and 62231002.

REFERENCES

- [1] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, “End-to-end recovery of human shape and pose,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7122–7131.
- [2] S. Tripathi, L. Müller, C.-H. P. Huang, O. Taheri, M. J. Black, and D. Tzionas, “3d human pose estimation via intuitive physics,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 4713–4725.
- [3] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, “Learning to reconstruct 3d human pose and shape via model-fitting in the loop,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2252–2261.
- [4] Z. Li, J. Liu, Z. Zhang, S. Xu, and Y. Yan, “Cliff: Carrying location information in full frames into human pose and shape estimation,” in *European Conference on Computer Vision*. Springer, 2022, pp. 590–606.
- [5] Q. Fang, Q. Shuai, J. Dong, H. Bao, and X. Zhou, “Reconstructing 3d human pose by watching humans in the mirror,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 814–12 823.
- [6] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, “Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3383–3393.
- [7] S. Goel, G. Pavlakos, J. Rajasegaran, A. Kanazawa, and J. Malik, “Humans in 4d: Reconstructing and tracking humans with transformers,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14 783–14 794.
- [8] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, and G. Pons-Moll, “Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time,” *ACM Transactions on Graphics (TOG)*, vol. 37, no. 6, pp. 1–15, 2018.
- [9] X. Yi, Y. Zhou, M. Habermann, S. Shimada, V. Golyanik, C. Theobalt, and F. Xu, “Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13 167–13 178.
- [10] Y. Du, R. Kips, A. Pumarola, S. Starke, A. Thabet, and A. Sanakoyeu, “Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 481–490.
- [11] Z. Gu, J. Li, W. Shen, W. Yu, Z. Xie, S. McCrory, X. Cheng, A. Shamsah, R. Griffin, C. K. Liu *et al.*, “Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning,” *IEEE/ASME Transactions on Mechatronics*, vol. 31, no. 2, pp. 2300–2330, 2026.
- [12] Y. Xie, B. Lou, A. Xie, and D. Zhang, “A review: Robust locomotion for biped humanoid robots,” in *Journal of Physics: Conference Series*, vol. 1487, no. 1. IOP Publishing, 2020, p. 012048.
- [13] L. Cao, “Humanoid robots and humanoid ai: Review, perspectives and directions,” *ACM Computing Surveys*, vol. 58, no. 4, pp. 1–37, 2025.
- [14] S. Ren, Y. Lu, J. Huang, J. Zhao, H. Zhang, T. Yu, Q. Shen, and X. Cao, “Motionpro: exploring the role of pressure in human mocap and beyond,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 27 760–27 770.
- [15] J. Koenemann, F. Burget, and M. Bennewitz, “Real-time imitation of human whole-body motions by humanoids,” in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 2806–2812.
- [16] F. Abi-Farrajl, B. Henze, A. Werner, M. Panzirsch, C. Ott, and M. A. Roa, “Humanoid teleoperation using task-relevant haptic feedback,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 5010–5017.
- [17] M. Elobaid, Y. Hu, G. Romualdi, S. Dafarra, J. Babic, and D. Pucci, “Telexistence and teleoperation for walking humanoid robots,” in *Proceedings of SAI Intelligent Systems Conference*. Springer, 2019, pp. 1106–1121.
- [18] K. Ayusawa and E. Yoshida, “Motion retargeting for humanoid robots based on simultaneous morphing parameter identification and motion optimization,” *IEEE Transactions on Robotics*, vol. 33, no. 6, pp. 1343–1357, 2017.
- [19] K. Otani and K. Bouyarmane, “Adaptive whole-body manipulation in human-to-humanoid multi-contact motion retargeting,” in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*. IEEE, 2017, pp. 446–453.
- [20] L. Penco, B. Clément, V. Modugno, E. M. Hoffman, G. Nava, D. Pucci, N. G. Tsagarakis, J.-B. Mouret, and S. Ivaldi, “Robust real-time whole-body motion retargeting from human to humanoid,” in *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2018, pp. 425–432.
- [21] K. Darvish, Y. Tirupachuri, G. Romualdi, L. Rapetti, D. Ferigo, F. J. A. Chavez, and D. Pucci, “Whole-body geometric retargeting for humanoid robots,” in *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2019, pp. 679–686.
- [22] S. Dafarra, U. Pattacini, G. Romualdi, L. Rapetti, R. Grieco, K. Darvish, G. Milani, E. Valli, I. Sorrentino, P. M. Viceconte *et al.*, “icub3 avatar system: Enabling remote fully immersive embodiment of humanoid robots,” *Science Robotics*, vol. 9, no. 86, p. eadh3834, 2024.
- [23] X. B. Peng, P. Abbeel, S. Levine, and M. van de Panne, “Deepmimic: example-guided deep reinforcement learning of physics-based character skills,” *ACM Trans. Graph.*, vol. 37, no. 4, p. 143, 2018.
- [24] Z. Luo, J. Cao, K. Kitani, W. Xu *et al.*, “Perpetual humanoid control for real-time simulated avatars,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 895–10 904.
- [25] Z. Luo, J. Cao, J. Merel, A. Winkler, J. Huang, K. Kitani, and W. Xu, “Universal humanoid motion representations for physics-based control,” in *International Conference on Learning Representations*, vol. 2024, 2024, pp. 56 766–56 782.
- [26] C. Tessler, Y. Guo, O. Nabati, G. Chechik, and X. B. Peng, “Masked-mimic: Unified physics-based character control through masked motion

- inpainting,” *ACM Transactions On Graphics (TOG)*, vol. 43, no. 6, pp. 1–21, 2024.
- [27] Z. Luo, J. Cao, S. Christen, A. Winkler, K. Kitani, and W. Xu, “Omni-grasp: Grasping diverse objects with simulated humanoids,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 2161–2184, 2024.
- [28] L. Yang, X. Huang, Z. Wu, A. Kanazawa, P. Abbeel, C. Sferrazza, C. K. Liu, R. Duan, and G. Shi, “Omni-2target: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction,” *arXiv preprint arXiv:2509.26633*, 2025.
- [29] Z. Zhang, J. Guo, C. Chen, J. Wang, C. Lin, Y. Lian, H. Xue, Z. Wang, M. Liu, J. Lyu *et al.*, “Track any motions under any disturbances,” *arXiv preprint arXiv:2509.13833*, 2025.
- [30] Z. Luo, Y. Yuan, T. Wang, C. Li, S. Chen, F. Castaneda, Z.-A. Cao, J. Li, D. Minor, Q. Ben *et al.*, “Sonic: Supersizing motion tracking for natural humanoid whole-body control,” *arXiv preprint arXiv:2511.07820*, 2025.
- [31] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. M. Kitani, C. Liu, and G. Shi, “Omni2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning,” in *Conference on Robot Learning*. PMLR, 2025, pp. 1516–1540.
- [32] W. Xie, J. Han, J. Zheng, H. Li, X. Liu, J. Shi, W. Zhang, C. Bai, and X. Li, “Kungfubot: Physics-based humanoid whole-body control for learning highly-dynamic skills,” *Advances in Neural Information Processing Systems*, vol. 38, pp. 62406–62433, 2026.
- [33] T. He, J. Gao, W. Xiao, Y. Zhang, Z. Wang, J. Wang, Z. Luo, G. He, N. Sobanbab, C. Pan *et al.*, “Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills,” *arXiv preprint arXiv:2502.01143*, 2025.
- [34] Y. Li, Z. Luo, T. Zhang, C. Dai, A. Kanervisto, A. Tirinzoni, H. Weng, K. Kitani, M. Guzek, A. Touati *et al.*, “Bfm-zero: A promptable behavioral foundation model for humanoid control using unsupervised reinforcement learning,” *arXiv preprint arXiv:2511.04131*, 2025.
- [35] Y. Ze, S. Zhao, W. Wang, A. Kanazawa, R. Duan, P. Abbeel, G. Shi, J. Wu, and C. K. Liu, “Twist2: Scalable, portable, and holistic humanoid data collection system,” *arXiv preprint arXiv:2511.02832*, 2025.
- [36] Q. Liao, T. E. Truong, X. Huang, Y. Gao, G. Tevet, K. Sreenath, and C. K. Liu, “Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion,” *arXiv preprint arXiv:2508.08241*, 2025.
- [37] P. Merriaux, Y. Dupuis, R. Boutteau, P. Vasseur, and X. Savatier, “A study of vicon system positioning performance,” *Sensors*, vol. 17, no. 7, p. 1591, 2017.
- [38] I. A. Faisal, T. W. Purboyo, and A. S. R. Ansori, “A review of accelerometer sensor and gyroscope sensor in imu sensors on motion capture,” *J. Eng. Appl. Sci.*, vol. 15, no. 3, pp. 826–829, 2019.
- [39] A. Winkler, J. Won, and Y. Ye, “Questsim: Human motion tracking from sparse sensors with simulated avatars,” in *SIGGRAPH Asia 2022 conference papers*, 2022, pp. 1–8.
- [40] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, “Humanplus: Humanoid shadowing and imitation from humans,” *arXiv preprint arXiv:2406.10454*, 2024.
- [41] M. Ji, X. Peng, F. Liu, J. Li, G. Yang, X. Cheng, and X. Wang, “Exbody2: Advanced expressive humanoid whole-body control,” *arXiv preprint arXiv:2412.13196*, 2024.
- [42] M. Kocabas, N. Athanasiou, and M. J. Black, “Vibe: Video inference for human body pose and shape estimation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5253–5263.
- [43] H. Choi, G. Moon, J. Y. Chang, and K. M. Lee, “Beyond static features for temporally consistent 3d human pose and shape from a video,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1964–1973.
- [44] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, “Learning 3d human dynamics from video,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5614–5623.
- [45] Y. Sun, Y. Ye, W. Liu, W. Gao, Y. Fu, and T. Mei, “Human mesh recovery from monocular images via a skeleton-disentangled representation,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5349–5358.
- [46] A. Arnab, C. Doersch, and A. Zisserman, “Exploiting temporal context for 3d human pose estimation in the wild,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3395–3404.
- [47] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, “Keep it smpl: Automatic estimation of 3d human pose and shape from a single image,” in *European conference on computer vision*. Springer, 2016, pp. 561–578.
- [48] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, “Expressive body capture: 3d hands, face, and body from a single image,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10975–10985.
- [49] A. Zanfir, E. G. Bazavan, H. Xu, W. T. Freeman, R. Sukthankar, and C. Sminchisescu, “Weakly supervised 3d human pose and shape reconstruction with normalizing flows,” in *European Conference on Computer Vision*. Springer, 2020, pp. 465–481.
- [50] I. Kissos, L. Fritz, M. Goldman, O. Meir, E. Oks, and M. Klinger, “Beyond weak perspective for monocular 3d human pose estimation,” in *European Conference on Computer Vision*. Springer, 2020, pp. 541–554.
- [51] M. Kocabas, C.-H. P. Huang, J. Tesch, L. Müller, O. Hilliges, and M. J. Black, “Spec: Seeing people in the wild with an estimated camera,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11035–11045.
- [52] W. Wang, Y. Ge, H. Mei, Z. Cai, Q. Sun, Y. Wang, C. Shen, L. Yang, and T. Komura, “Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 3925–3935.
- [53] Y. Yuan, U. Iqbal, P. Molchanov, K. Kitani, and J. Kautz, “Glamr: Global occlusion-aware human mesh recovery with dynamic cameras,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11038–11049.
- [54] V. Ye, G. Pavlakos, J. Malik, and A. Kanazawa, “Decoupling human and camera motion from videos in the wild,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 21222–21232.
- [55] Y. Wang, Z. Wang, L. Liu, and K. Daniilidis, “Tram: Global trajectory and motion of 3d humans from in-the-wild videos,” in *European Conference on Computer Vision*. Springer, 2024, pp. 467–487.
- [56] Y. Sun, Q. Bao, W. Liu, T. Mei, and M. J. Black, “Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8856–8866.
- [57] S. Shin, J. Kim, E. Halilaj, and M. J. Black, “Wham: Reconstructing world-grounded humans with accurate 3d motion,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2070–2080.
- [58] Z. Shen, H. Pi, Y. Xia, Z. Cen, S. Peng, Z. Hu, H. Bao, R. Hu, and X. Zhou, “World-grounded human motion recovery via gravity-view coordinates,” in *SIGGRAPH Asia 2024 Conference Papers*, 2024, pp. 1–11.
- [59] H. M. Clever, A. Kapusta, D. Park, Z. Erickson, Y. Chitalia, and C. C. Kemp, “3d human pose estimation on a configurable bed from a pressure image,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 54–61.
- [60] H. M. Clever, Z. Erickson, A. Kapusta, G. Turk, K. Liu, and C. C. Kemp, “Bodies at rest: 3d human pose and shape estimation from a pressure image using synthetic data,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6215–6224.
- [61] Y. Yin, J. P. Robinson, and Y. Fu, “Multimodal in-bed pose and shape estimation under the blankets,” in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2411–2419.
- [62] A. Tandon, A. Goyal, H. M. Clever, and Z. Erickson, “Bodymap-jointly predicting body mesh and 3d applied pressure map for people in bed,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2480–2489.
- [63] S. Liu, X. Huang, N. Fu, C. Li, Z. Su, and S. Ostadabbas, “Simultaneously-collected multimodal lying pose dataset: Enabling in-bed human pose monitoring,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 1106–1118, 2022.
- [64] H. M. Clever, P. L. Grady, G. Turk, and C. C. Kemp, “Bodypressure - inferring body pose and contact pressure from a depth image,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 137–153, 2023.
- [65] Z. Wu, F. Xie, Y. Fang, Z. Liang, Q. Wan, Y. Xiong, and X. Cai, “Seeing through the tactile: 3d human shape estimation from temporal in-bed pressure images,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 2, pp. 86:1–86:39, 2024.
- [66] H. Zhang, S. Ren, H. Yuan, J. Zhao, F. Li, S. Sun, Z. Liang, T. Yu, Q. Shen, and X. Cao, “Mmvp: A multimodal mocap dataset with vision and pressure sensors,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21842–21852.
- [67] X. Han, B. Senderling, S. To, D. Kumar, E. Whiting, and J. Saito, “Groundlink: A dataset unifying human body movement and ground reaction dynamics,” in *SIGGRAPH Asia 2023 Conference Papers*, 2023, pp. 1–10.

- [68] L. Mourot, L. Hoyet, F. L. Clerc, and P. Hellier, "Underpressure: Deep learning for foot contact detection, ground reaction force estimation and footskate cleanup," in *Computer Graphics Forum*, vol. 41, no. 8. Wiley Online Library, 2022, pp. 195–206.
- [69] J. Scott, B. Ravichandran, C. Funk, R. T. Collins, and Y. Liu, "From image to stability: Learning dynamics from human pose," in *European conference on computer vision*. Springer, 2020, pp. 536–554.
- [70] Y. Luo, Y. Li, M. Foshey, W. Shou, P. Sharma, T. Palacios, A. Torralba, and W. Matusik, "Intelligent carpet: Inferring 3d human pose from tactile signals," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 255–11 265.
- [71] Y. Lu, S. Ren, Q. Shen, and X. Cao, "Leveraging rgb-pressure for whole-body human-to-humanoid motion imitation," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 8932–8941.
- [72] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [73] J. P. Araujo, Y. Ze, P. Xu, J. Wu, and C. K. Liu, "Retargeting matters: General motion retargeting for humanoid motion tracking," *arXiv preprint arXiv:2510.02252*, 2025.
- [74] Q. Zhao, K. Yang, X. Wang, S. Zhao, Y. Lu, X. Zhang, Q. Shen, X.-X. Long, and X. Cao, "Make tracking easy: Neural motion retargeting for humanoid whole-body control," *arXiv preprint arXiv:2603.22201*, 2026.
- [75] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 851–866.
- [76] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5442–5451.
- [77] Z. Cai, W. Yin, A. Zeng, C. Wei, Q. Sun, W. Yanjun, H. E. Pang, H. Mei, M. Zhang, L. Zhang *et al.*, "Smpl-x: Scaling up expressive human pose and shape estimation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 11 454–11 468, 2023.
- [78] Y. Zhang, J. O. Kephart, Z. Cui, and Q. Ji, "Physpt: Physics-aware pretrained transformer for estimating human dynamics from monocular videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2305–2317.



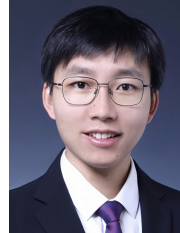
Zhaoxiang Li is an undergraduate student at the School of Electronic Science and Engineering, Nanjing University, China. His research interests include reinforcement learning and embodied intelligence.



Jiaqi Li is an undergraduate student at the School of Electronic Science and Engineering, Nanjing University, China. Her research interests include humanoid robot motion control and evaluation.



He Zhang received the M.S. and Ph.D. degrees from Beihang University in 2018 and 2024, respectively. He is currently a postdoc at BNRist, Tsinghua University, Beijing, China. His current research focuses on motion capture, biomechanics, and behaviour science.



Tao Yu is an associate researcher at BNRist, Tsinghua University. He received the B.S. degree in measurement and control from the Hefei University of Technology, in 2012, and the Ph.D degree from Beihang University, in 2019. His current research interests include 3D vision, AI, and computer graphics.



Qiu Shen received the B.S. and Ph.D. degree from the University of Science and Technology of China, in 2004 and 2009 respectively. From 2009 to 2016, he has been with Huawei 12 Lab and Nanjing University of Aeronautics and Astronautics. She is an associate professor at the School of Electronic Science and Engineering School, Nanjing University, China. Her current research focuses on next-generation video coding, collaborative video compression and analysis, embodied intelligence, and vision model for virtual reality.



Xun Cao received the B.S. degree from Nanjing University, Nanjing, China, in 2006, and the Ph.D. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2012. He held visiting positions with Philips Research, Aachen, Germany, in 2008 and Microsoft Research Asia, Beijing, from 2009 to 2010. He was a Visiting Scholar with the University of Texas at Austin, Austin, TX, USA, from 2010 to 2011. He is a Professor at the School of Electronic Science and Engineering, Nanjing University. His current research interests

include computational photography and image-based modeling and rendering.



Yi Lu is a graduate student for Ph.D degree at the School of Electronic Science and Engineering, Nanjing University, China. He received his B.S. degree from Nanjing University of Posts and Telecommunications in 2021. His research interests include embodied intelligence, humanoid robot motion control, and human motion capture.



Shenghao Ren is a postgraduate student at the School of Electronic Science and Engineering, Nanjing University, China. He received a B.S. degree from Nanjing University in 2023. His research interests include computer vision and human mesh recovery.



Tianyu Xiong is an undergraduate student at the School of Electronic Science and Engineering, Nanjing University, China. His research interests include embodied intelligence and human motion capture.