

FARM: Frame-Accelerated Augmentation and Residual Mixture-of-Experts for Physics-Based High-Dynamic Humanoid Control

Jing Tan¹, Shiting Chen², Yangfan Li¹, Weisheng Xu¹, Renjing Xu^{1*}

¹ Hong Kong University of Science and Technology (Guangzhou), China

² Guangdong University of Technology, China
colinjing1@gmail.com, renjingxu@hkust-gz.edu.cn

Abstract

Unified physics-based humanoid controllers are pivotal for robotics and character animation, yet models that excel on gentle, everyday motions still stumble on explosive actions. We bridge this gap with FARM (Frame-Accelerated Augmentation and Residual Mixture-of-Experts), an end-to-end framework composed of frame-accelerated augmentation, a robust base controller, and a residual mixture-of-experts (MoE). Frame-accelerated augmentation exposes the model to high-velocity pose changes by widening inter-frame gaps. The base controller reliably tracks everyday low-dynamic motions, while the residual MoE adaptively allocates additional network capacity to handle challenging high-dynamic actions, significantly enhancing tracking accuracy. In the absence of a public benchmark, we curate the High-Dynamic Humanoid Motion (HDHM) dataset, comprising 3593 physically plausible clips. On HDHM, FARM reduces the tracking failure rate by 42.8% and lowers global mean per-joint position error by 14.6% relative to the baseline, while preserving near-perfect accuracy on low-dynamic motions. These results establish FARM as a new baseline for high-dynamic humanoid control and introduce the first open benchmark dedicated to this challenge.

Code and Dataset — <https://github.com/Colin-Jing/FARM>

1 Introduction

Physics-based humanoid motion imitation combines the visual fidelity of motion-capture data with the strict dynamics of rigid-body simulation, underpinning advanced robotics, digital characters, and immersive AR/VR experiences. In practice, however, each new target sequence still demands days of reward redesign and hyperparameter tuning. This heavy engineering overhead hampers large-scale adoption and highlights the need for a robust universal controller that can track a wide range of motions without per-task retraining.

The Universal Humanoid Controller (UHC) (Luo et al. 2021) first explored the challenge of universal motion tracking, achieving a 97% tracking success rate on the 10k-sequence AMASS (Mahmood et al. 2019) dataset by incorporating a hand-crafted residual force controller (RFC) (Yuan and Kitani 2020). Perpetual Humanoid Control (PHC) (Luo

et al. 2023) advanced this by removing the RFC and introducing a progressive multiplicative control architecture, elevating the tracking success rate to 98.9%. Its subsequent refinement, PHC+ (Luo et al. 2024), further improved performance through data filtering and training optimizations, achieving near-100% success. However, the incremental learning approach inherent in PHC+ complicates the training process and hampers scalability. Addressing this complexity, the Fully Constrained (FC) Controller proposed in MaskedMimic (Tessler et al. 2024) adopts a unified network architecture, simplifying training and yielding superior generalization capabilities. Despite these methodological advances, all aforementioned controllers share a critical limitation in that they are trained primarily on the AMASS dataset, dominated by simple, low-dynamic everyday motions. This restricts their ability to generalize to highly dynamic motions, leaving the robust control of explosive actions an open challenge in universal humanoid control.

In order to address this limitation, we initially considered a straightforward intuition: since the AMASS dataset predominantly comprises low-dynamic actions with small pose-to-pose variations, artificially accelerating motion clips through uniform frame dropping might enlarge the pose differences between successive frames, mimicking the abrupt transitions seen in highly dynamic motions. However, experiments (Section 4.4) revealed that fine-tuning the FC controller directly on these frame-accelerated clips provided minimal improvement for high-dynamic actions, and even degraded its previously near-perfect accuracy on low-dynamic sequences. This prompted a second key insight inspired by human motor attention: routine, everyday movements demand minimal cognitive resources, whereas explosive, high-dynamic actions require heightened attention and additional processing capacity. Analogously, we reasoned that effective high-dynamic control necessitates dynamically adjusting network capacity to match motion intensity, ensuring high accuracy without compromising low-dynamic performance.

Motivated by these insights, we propose the FARM framework, which consists of three components: frame-accelerated augmentation, a base controller, and a residual mixture-of-experts (MoE). Frame-accelerated augmentation widens the pose gap between successive frames to expose the model to the abrupt transitions found in highly dynamic actions. The base controller delivers dependable baseline motion track-

*Corresponding author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

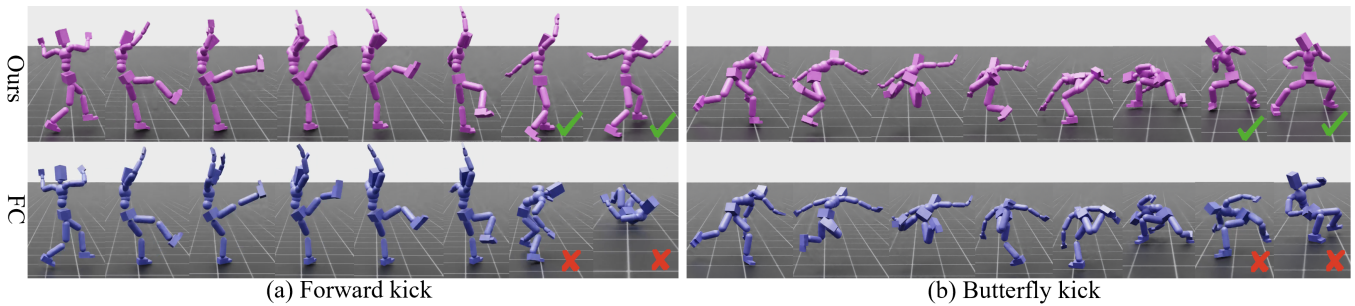


Figure 1: Comparison between FARM and the baseline FC on two high-dynamic motions. FARM accurately completes both motions (green check), while FC falls or loses balance (red cross). Frames are shown in playback order from left to right.

ing for everyday movements. Building on this foundation, the residual MoE tackles the remaining high-dynamic errors. Its speed-aware router (SAR) stratifies motions by dynamic intensity and directs each band to a dedicated subset of experts, while the dynamic expert-assignment (DEA) adaptively scales capacity, activating only the required experts for low-dynamic segments and progressively more for highly dynamic ones. Together, these components enable FARM to master challenging high-dynamic motions while preserving strong performance on low-dynamic tasks. Figure 1 illustrates FARM’s ability to accurately complete complex high-dynamic motions, where the baseline FC model fails.

Because no public dataset exists specifically for high-dynamic humanoid motions, we curate the High-Dynamic Humanoid Motion (HDHM) benchmark for controlled evaluation. HDHM comprises 3593 physics-plausible clips, each manually screened to ensure kinematic validity. On this benchmark, our FARM framework cuts the tracking failure rate by 42.8% and lowers the global mean per-joint position error (MPJPE_g) by 14.6% relative to the baseline, while preserving the near-perfect accuracy of the baseline on low-dynamic motions. Our contributions are as follows:

1. **HDHM dataset.** We release the first open, manually curated benchmark dedicated to high-dynamic humanoid control, comprising 3593 physics-plausible clips.
2. **FARM framework.** We propose a frame-accelerated residual MoE framework. Frame-accelerated augmentation allows the model to acquire high-dynamic skills from predominantly low-dynamic data, and the residual MoE adaptively allocates computational capacity according to motion intensity.
3. **Comprehensive experiments.** Extensive evaluations show that FARM significantly outperforms strong baselines on challenging high-dynamic motions while preserving the near-perfect accuracy on low-dynamic motions.

2 Related Works

2.1 Physics-based motion imitation

Physics-based motion imitation learns control policies to track motion references in simulation using reinforcement learning (Mnih et al. 2016). DeepMimic (Peng et al. 2018), AMP (Peng et al. 2021), and PHC (Luo et al. 2023) progressively improved motion diversity and generalization, with

PHC+ (Luo et al. 2024) achieving near-perfect tracking on large-scale datasets like AMASS. MaskedMimic (Tessler et al. 2024) further unified diverse conditioning forms into a single controller via masked modeling. While effective on standard mocap data, these methods struggle with highly dynamic motions due to limited exposure to fast transitions. Our approach addresses this by explicitly targeting high-dynamic regimes through data and architectural adaptations.

2.2 Mixture-of-experts in motion control

Several recent works apply MoE to improve specialization and scalability in humanoid or legged control. GMT (Chen et al. 2025) employs MoE alongside adaptive sampling to track diverse whole-body motions with a unified policy. MoRE (Wang et al. 2025) uses a mixture of latent residual experts to learn human-like gaits across complex terrains. MoE-LoCo (Huang et al. 2025) addresses multitask locomotion with dynamic selection of experts to mitigate gradient conflicts and support varied gait and terrain combinations. These methods demonstrate the benefits of MoE-based specialization. Unlike fixed expert usage in prior work, our residual MoE adapts the expert count based on motion intensity for better capacity allocation.

3 Method

This section presents the details of our proposed FARM framework. Section 3.1 outlines the overall structure of FARM. Section 3.2 describes the frame-accelerated augmentation technique employed to expose the model to high-velocity pose transitions. Section 3.3 elaborates on the residual MoE architecture, which adaptively enhances the network’s capacity to handle highly dynamic motions.

3.1 Overall framework

As illustrated in Figure 2, the training pipeline of the FARM framework first mines hard samples from AMASS, and then trains the FARM model on these samples. The trained policy is evaluated on HDHM to assess generalization to unseen high-dynamic motions.

Prior work (Luo, Yuan, and Kitani 2022; Zhu et al. 2023) has shown that focusing on challenging samples can significantly improve model performance, and we observe similar benefits in our setting (see Section 4.4). To mine such samples, we uniformly accelerate the entire AMASS dataset by a

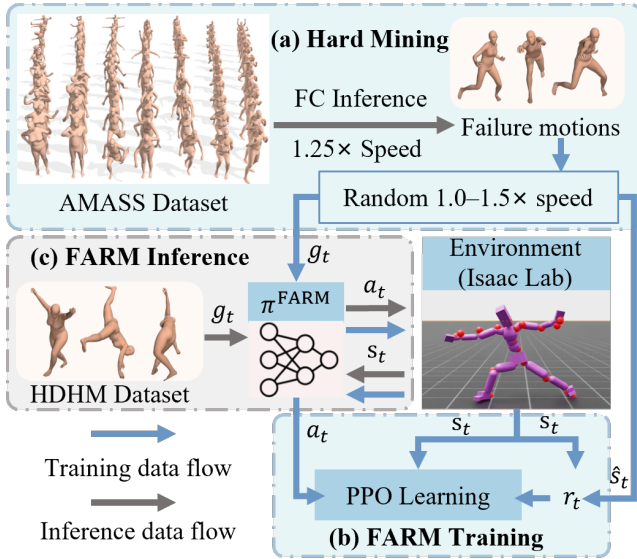


Figure 2: Overview of the FARM pipeline. Failure cases are mined by applying the FC model to the AMASS dataset at 1.25 \times speed. These hard samples are augmented with random 1.0–1.5 \times acceleration and used for FARM training. The learned policy is evaluated on the HDHM dataset to assess performance under high-dynamic motions.

factor of 1.25 and apply the base FC (Tessler et al. 2024) controller π^{FC} to perform inference on the accelerated motion sequences. Motions that the controller fails to track under this setting are collected as hard samples. A motion is considered a failure if the global mean per-joint position error exceeds 0.5 meters at any frame (Luo et al. 2021). During training, they are augmented by randomly applying frame acceleration in the range of 1.0 \times to 1.5 \times (See Section 3.2).

To train our controller for the motion imitation task, we adopt the framework of goal-conditioned reinforcement learning (GCRL). At each time step t , the agent receives the current observation s_t along with an imitation goal signal g_t , and generates an action $a_t \sim \pi^{FARM}(a_t | s_t, g_t)$. After executing the action in the physics-based simulation environment (Isaac Lab (Mittal et al. 2023)), a transition to the next state s_{t+1} occurs, and the agent receives a reward r_t . The objective is to optimize the policy π^{FARM} to maximize the expected discounted return $J = \mathbb{E} \left[\sum_{t=0}^T \gamma^t r_t \right]$, where $\gamma \in [0, 1]$ is a discount factor. We use proximal policy gradient (PPO (Schulman et al. 2017)) to optimize π^{FARM} .

The state $s_t = [f_t^{\text{char}}, f_t^{\text{scene}}]$ includes the character’s proprioception and terrain information. Although evaluation is conducted on flat terrain, introducing irregular terrain during training improves the robustness of the policy when the reference motion contains slight ground penetration or floating artifacts. The goal g_t encodes a sequence of future target poses from the reference motion. The action a_t represents the proportional-derivative (PD) target. The reward r_t encourages the agent to follow the reference motion while maintaining smooth and physically plausible behavior. It is defined as

a weighted sum of a motion tracking term and an energy penalty term:

$$r_t = w^{\text{track}} r_t^{\text{track}}(s_t, \hat{s}_t) + w^{\text{energy}} r_t^{\text{energy}}, \quad (1)$$

where $r_t^{\text{track}}(s_t, \hat{s}_t)$ measures the discrepancy between the simulated character state s_t and the reference state \hat{s}_t (Da Silva, Abe, and Popović 2008; Lee, Kim, and Lee 2010; Wang et al. 2020). The term r_t^{energy} penalizes excessive joint torques to promote smoother motion. For detailed formulations of each component, refer to Section 5 of Tessler et al. (2024). Our implementation follows the same setting.

After training, the learned policy is evaluated on the HDHM dataset, which contains diverse high-dynamic motions that are not seen during training. Details of the HDHM dataset are provided in Section 4.1 and Appendix A.

3.2 Frame-accelerated augmentation

For each motion clip used in training, a constant acceleration factor is drawn from a uniform distribution, $v \sim \mathcal{U}[1.0, 1.5]$. With the nominal playback step $\Delta t = \frac{1}{30}$ s (30 Hz), the resampled step becomes $\Delta t_{\text{new}} = v \Delta t$. Virtual timestamps are defined as $t_k = k \Delta t_{\text{new}}$ for $k \in \mathbb{N}$. The pose f_k at t_k is obtained by linear interpolation between the two original frames $f_{\lfloor t_k / \Delta t \rfloor}$ and $f_{\lceil t_k / \Delta t \rceil}$. The resulting sequence is fed to the controller at the original 30 Hz rate. Hence, consecutive poses correspond to a physical time interval scaled by v . As illustrated in Figure 3 (a), this augmentation generates motions with enlarged inter-frame displacements, thereby exposing the policy to more dynamic conditions.

3.3 Residual mixture-of-experts

The baseline FC controller tracks low-dynamic motions reliably but fails on rapid pose changes. To extend its dynamic range without disrupting its proven behaviour, we freeze the FC input MLP and Transformer (Vaswani et al. 2017), update only the output MLP, and attach a residual MoE module (Figure 3 (c)). Given token features z_{in} from the frozen input MLP, the frozen Transformer produces $z_{\text{base}} = \text{Transformer}_{\text{FC}}(z_{\text{in}})$. In parallel the MoE computes a residual $\Delta z = \text{MoE}(z_{\text{in}})$. The combined representation $z = z_{\text{base}} + \Delta z$ is fed to the output MLP to yield the action a_t . Thus, the original controller is preserved, while the residual MoE supplies extra corrections for high-dynamic segments. The residual MoE is equipped with a speed-aware router and a dynamic expert-assignment mechanism, described in the following paragraph.

Speed-aware router (SAR) SAR promotes expert specialization by routing inputs according to motion intensity, allowing each expert to focus on a specific velocity regime. To guide this routing behavior, an auxiliary supervision signal is introduced. Specifically, for each training batch, the average joint velocity of the reference frame is computed for each sample. Samples are then sorted and evenly partitioned into three speed groups, labeled as 0, 1, and 2. These labels are used to supervise the router via a cross-entropy loss \mathcal{L}_{CE} . The router is jointly optimized with the main reinforcement learning objective using the combined loss:

$$\mathcal{L}_{\text{router}} = \mathcal{L}_{\text{RL}} + \lambda_{\text{speed}} \mathcal{L}_{\text{CE}}, \quad (2)$$

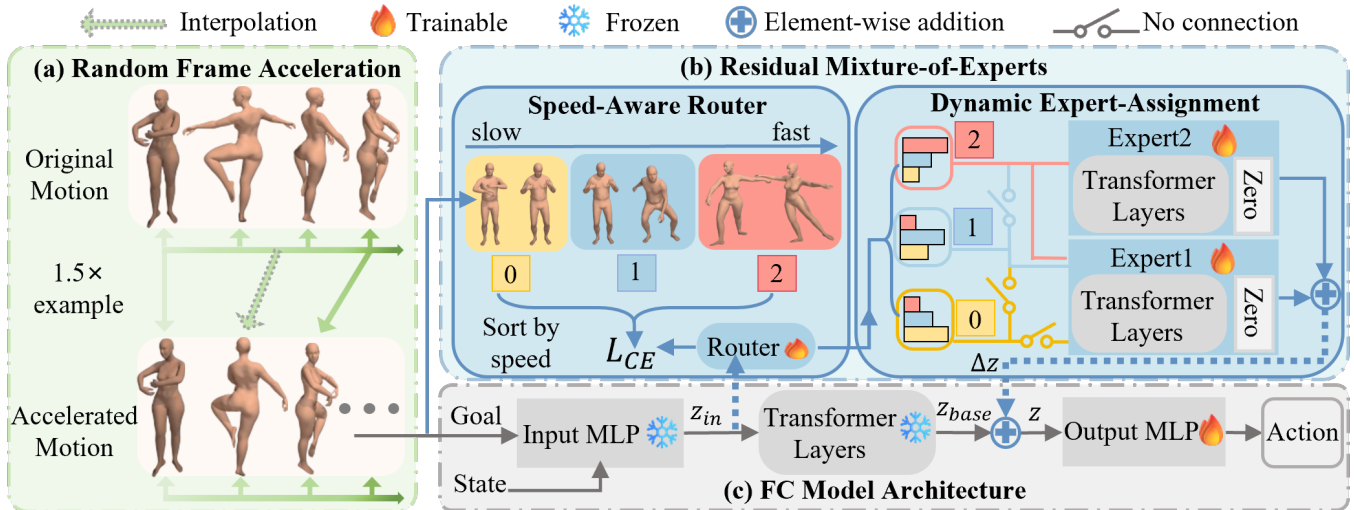


Figure 3: Overview of the FARM framework. Frame-accelerated augmentation increases frame intervals by uniformly downsampling motion sequences with random speed to simulate high-dynamic motions. The FARM model consists of a base FC model for general motion tracking and a residual MoE module that enhances high-dynamic motion control through the speed-aware router and dynamic expert-assignment.

where λ_{speed} balances the auxiliary supervision. This formulation ensures that each expert receives data in a specific speed band, encouraging specialization and avoiding the common issue (Mu and Lin 2025) in unconstrained MoE setups where a single expert dominates and others are rarely selected.

Dynamic expert-assignment (DEA) The DEA mechanism is inspired by human attention allocation: routine actions require minimal focus, whereas high-dynamic activities demand greater cognitive effort. Similarly, our method activates a variable number of experts based on motion intensity. Let $p \in \mathbb{R}^E$ be the router’s softmax output, where E is the total number of experts. We define $k = \arg \max(p)$ and activate experts indexed from 1 to k . The residual is computed as

$$\Delta z = \sum_{i=1}^k w_i \cdot \text{Expert}_i(z_{\text{in}}), \quad w_i = \sum_{j=i}^{E-1} p_j. \quad (3)$$

When $k = 0$, no expert is used and the model reduces to the base FC controller. Followed by ControlNet (Zhang, Rao, and Agrawala 2023), each expert copies the transformer architecture and weights of FC and appends a zero-initialized linear projection layer, ensuring the residual path does not affect the initial behavior. This design enables adaptive capacity scaling with motion difficulty while preserving the reliability of the original controller.

4 Experiments

In this section, we conduct a comprehensive evaluation of the proposed FARM framework. Section 4.1 describes the datasets used in our experiments. Section 4.2 presents both quantitative and qualitative comparisons. Section 4.3 analyzes the expert selection behavior within the residual MoE module. Section 4.4 provides ablation studies to investigate the contribution of each component in the FARM framework. Additional experiment details are provided in Appendix B.

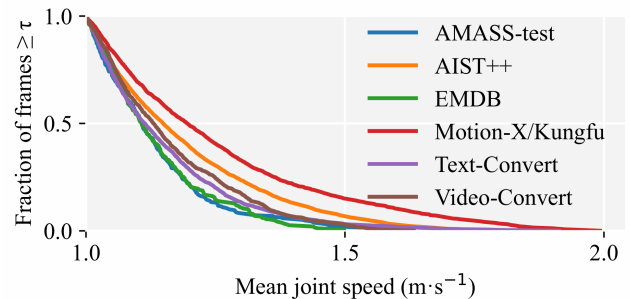


Figure 4: Survival functions of per-frame mean joint speeds for six datasets. τ represents the speed threshold. For visual clarity, only speeds between $1.0\text{--}2.0\text{m}\cdot\text{s}^{-1}$ are shown. Almost all five HDHM-derived datasets lie above the AMASS-test curve, indicating greater dynamic activity.

4.1 Datasets

Following Luo et al. (2024), we filter the AMASS dataset to remove physically implausible motions, and split it into two subsets: AMASS-train for training and AMASS-test for evaluation. To further assess high-dynamic motion tracking, we construct the HDHM dataset, composed of curated clips from five sources: AIST++ (Li et al. 2021; Tsuchida et al. 2019), EMDB (Kaufmann et al. 2023), Motion-X Kungfu subset (Lin et al. 2023), Text-Convert, and Video-Convert. All clips are manually filtered to exclude those involving explicit environment interaction or exhibiting non-physical artifacts such as body self-intersections, floating, or ground penetration. All sequences are temporally resampled to 30 Hz. AIST++ contains diverse dance sequences, EMDB covers both indoor daily activities and outdoor sports, and the Motion-X Kungfu subset features complex kungfu movements. The

	AIST++			EMDB			Motion-X/Kungfu		
	Success \uparrow	MPJPE _g \downarrow	MPJPE _l \downarrow	Success \uparrow	MPJPE _g \downarrow	MPJPE _l \downarrow	Success \uparrow	MPJPE _g \downarrow	MPJPE _l \downarrow
PHC+	81.5%	210.7	102.0	71.0%	302.8	89.5	70.2%	265.5	122.4
FC	93.1%	108.5	76.6	95.5%	98.3	57.4	84.1%	155.3	86.5
Ours	95.6%	94.2	69.6	97.7%	91.2	57.0	90.7%	124.2	72.0
	Text-Convert			Video-Convert			Total(HDHM)		
	Success \uparrow	MPJPE _g \downarrow	MPJPE _l \downarrow	Success \uparrow	MPJPE _g \downarrow	MPJPE _l \downarrow	Success \uparrow	MPJPE _g \downarrow	MPJPE _l \downarrow
PHC+	87.3%	166.0	77.3	57.8%	303.4	144.5	80.3%	217.6	100.6
FC	96.6%	76.9	50.7	83.2%	173.0	91.4	92.3%	111.3	71.7
Ours	98.6%	71.2	48.5	90.7%	116.8	71.8	95.6%	95.0	64.4

Table 1: Comparison of our method with PHC+ and FC on the five subsets of the HDHM dataset and the combined overall set, with all results averaged over four random seeds. \uparrow indicates that higher values are better, while \downarrow indicates that lower values are better. Bold numbers highlight the best performance in each column. Our method consistently outperforms baselines across all subsets and the overall average. The results for PHC+ and FC are reproduced using the official released weights evaluated locally.

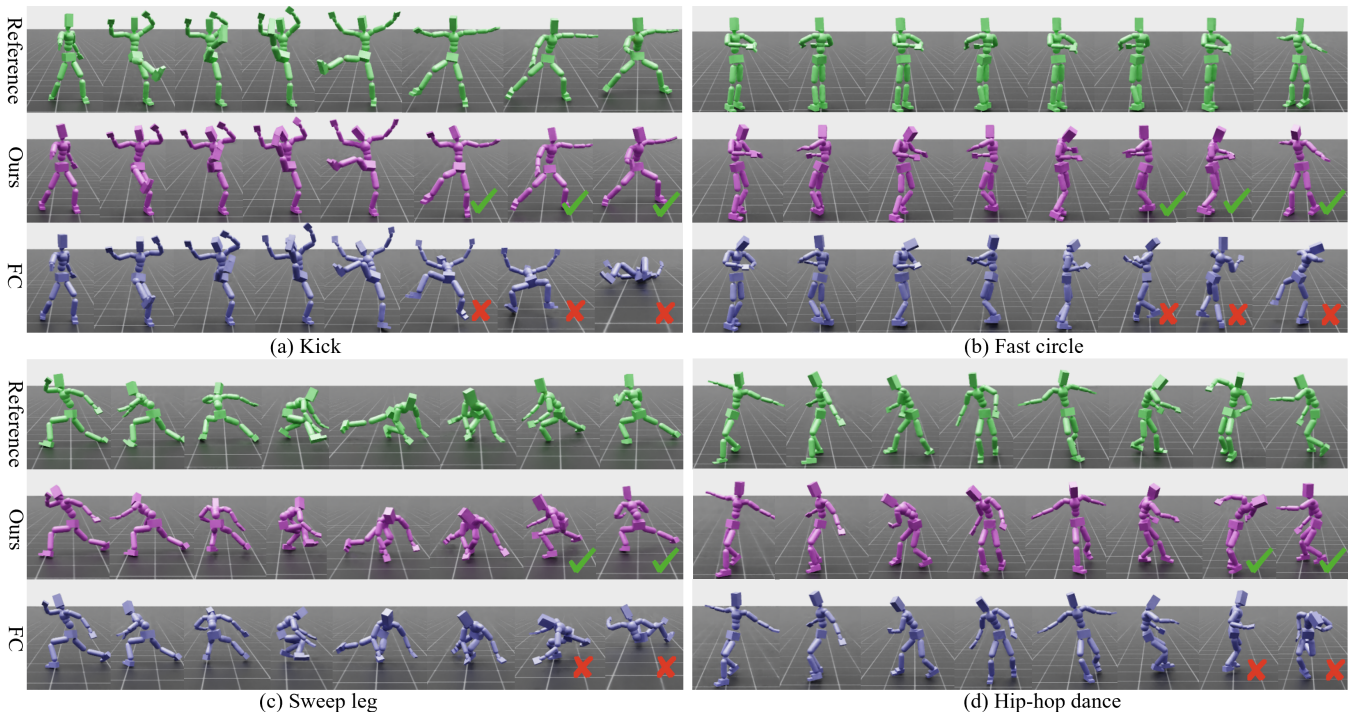


Figure 5: Qualitative comparison on four high-dynamic motion clips. Motion frames are presented in playback order from left to right. The reference motion is shown in green, our FARM results in purple, and the FC baseline in blue. FARM consistently maintains accurate motion tracking (green check), while FC exhibits noticeable failures under rapid pose changes (red cross).

Text-Convert set is generated using motion diffusion models (Tevet et al. 2022, 2024) conditioned on high-dynamic prompts, while the Video-Convert set is created by converting YouTube martial arts videos via the GVHMR model (Shen et al. 2024). Figure 4 illustrates that the HDHM dataset exhibits higher motion dynamics compared to AMASS. Among these, AIST++, EMDB, and the Motion-X Kungfu subset are publicly available datasets, while Text-Convert and Video-Convert are constructed by us. In total, the HDHM dataset contains 3593 clips with an average duration of 9.4 seconds, with AIST++ (1320, 12.4s), EMDB (45, 36.3s), Motion-X

Kungfu (663, 10.1s), Text-Convert (1392, 5.9s), and Video-Convert (173, 5.5s). For more details, refer to Appendix A.

4.2 Performance Evaluation

Quantitative comparison. We compare FARM with two state-of-the-art baselines, FC (Tessler et al. 2024) and PHC+ (Luo et al. 2024), on both the HDHM and AMASS datasets. Evaluation metrics include the motion tracking success rate, the global mean per-joint position error (MPJPE_g), and the root-relative MPJPE (MPJPE_l), following the definitions used in Luo et al. (2021). The success rate measures whether



Figure 6: Visualization of expert activation throughout a cartwheel motion. Each frame is colored according to the number of active experts (yellow for 0, blue for 1, and red for 2). We observe that expert activation aligns with motion dynamics.

	AMASS-train		AMASS-test	
	Success \uparrow	MPJPE _g \downarrow	Success \uparrow	MPJPE _g \downarrow
PHC+	99.9%	28.8	98.5%	38.1
FC	99.7%	33.3	100%	40.8
Ours	99.9%	35.2	100%	42.1

Table 2: Performance on AMASS-train and AMASS-test datasets. The AMASS dataset is primarily composed of low-dynamic motions. Our method achieves high tracking success rates comparable to the baselines, and the increase in MPJPE_g compared to FC remains below 2 mm.

the humanoid can track the reference motion without falling or significantly lagging behind. MPJPE_l and MPJPE_g quantify the imitation accuracy in root-relative and global coordinates, respectively. Table 1 summarizes the performance on the five subsets of the HDHM dataset and the overall aggregate. Our method achieves the highest success rate and lowest MPJPE scores on all subsets, demonstrating improved tracking robustness and accuracy, especially under high-dynamic motions. Table 2 shows that our method retains high success rates comparable to the baselines, and introduces less than a 2mm increase in MPJPE_g relative to FC. This shows that FARM not only improves tracking on challenging motions but also preserves performance on standard motions.

Qualitative comparison. To further assess high-dynamic tracking quality, we visually compare FARM with the FC baseline on four challenging clips in Figure 5. While FC often drifts or collapses during fast transitions, FARM successfully maintains balance and closely follows the reference, demonstrating improved robustness under rapid pose changes.

4.3 Expert behaviour analysis

Case study. We present a case study on a cartwheel motion to illustrate how the residual MoE dynamically adjusts expert usage. As shown in Figure 6, each humanoid is colored based on the number of activated experts. We observe that expert activations correlate well with motion dynamics. Fewer experts are used during low-motion phases, while high-dynamic segments trigger more experts. This indicates that the SAR successfully learns to distinguish motion dynamics, providing a foundation for effective dynamic expert-assignment.

Statistical and spatial analysis. To further understand the behavior of our residual MoE, we analyze both the distribution of activated expert counts and the spatial organization of expert outputs. As shown in Figure 7 (left), the AMASS-test

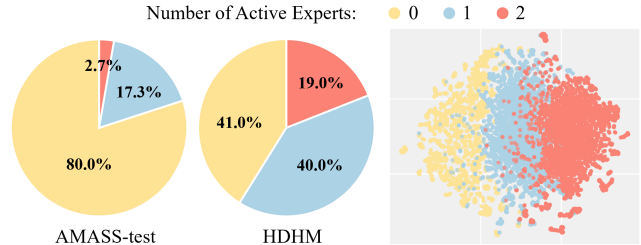


Figure 7: **Left:** Distribution of expert usage indicates that our controller engages more experts on HDHM in response to higher motion dynamics. **Right:** t-SNE visualization of Video-Convert embeddings demonstrates that the residual MoE encodes distinct and structured features.

set, which primarily contains low-dynamic motions, rarely activates experts, with 80% of frames using only the base controller. In contrast, the HDHM dataset exhibits more frequent activation of experts, with over half of the frames involving one or two experts. This confirms that our model allocates more expert capacity for high-dynamic motions.

To investigate whether these expert activations lead to distinguishable control behaviors, we extract the output embeddings of the residual MoE under different expert counts, and visualize them using t-SNE, as shown in Figure 7 (right). Each point represents the final embedding z from a motion frame, colored by the number of activated experts. The resulting projection reveals a clear spatial separation, indicating that the residual MoE learns structured and distinct representations corresponding to different motion complexities. This further supports the effectiveness of our SAR and DEA mechanism.

4.4 Ablation study

Component ablation. We group the ablation results in Table 3 by major components. Row 5 shows that applying frame acceleration (FA) alone yields minimal improvement, indicating that the base model cannot fully exploit high-dynamic data. In contrast, combining FA with MoE (Rows 3 and 8) leads to clear gains, suggesting that additional expert capacity is essential for leveraging harder samples. Among MoE designs, the residual variant (Res-MoE, Rows 2 and 8) consistently outperforms both the No-MoE baseline (Row 5) and the Full-MoE variant (Row 3), which jointly updates the base and expert modules. Full-MoE degrades performance on AMASS-test, highlighting that modifying the pretrained base harms generalization. In contrast, Res-MoE preserves the base and adapts only the residual, improving

Row	Data	MoE			Router				HDHM		AMASS-test	
	FA	Res-MoE	Full-MoE	No-MoE	SAR	DEA	Top1	Top2	Success \uparrow	MPJPE _g \downarrow	Success \uparrow	MPJPE _g \downarrow
1	\times	\times	\times	\times	\times	\times	\times	\times	92.3%	111.3	100%	40.8
2	\times	\checkmark	\times	\times	\checkmark	\checkmark	\times	\times	95.1% (+3.0%)	97.0 (-12.9%)	100%	41.6 (+2.0%)
3	\checkmark	\times	\checkmark	\times	\checkmark	\checkmark	\times	\times	94.3% (+2.2%)	104.9 (-5.8%)	100%	47.3 (+15.9%)
4	\checkmark	\checkmark	\times	\times	\times	\checkmark	\times	\times	94.8% (+2.7%)	97.5 (-12.4%)	100%	44.6 (+9.3%)
5	\checkmark	\times	\times	\checkmark	\times	\times	\times	\times	92.8% (+0.5%)	111.4 (+0.1%)	100%	46.2 (+13.2%)
6	\checkmark	\checkmark	\times	\times	\checkmark	\times	\checkmark	\times	95.5% (+3.5%)	97.5 (-12.4%)	100%	44.2 (+8.3%)
7	\checkmark	\checkmark	\times	\times	\checkmark	\times	\times	\checkmark	95.3% (+3.3%)	98.1 (-11.9%)	100%	44.8 (+9.8%)
8	\checkmark	\checkmark	\times	\times	\checkmark	\checkmark	\times	\times	95.6% (+3.6%)	95.0 (-14.6%)	100%	42.1 (+3.2%)

Table 3: Ablation results on the HDHM and AMASS-test datasets. Row 1 shows the performance of the baseline FC model. Values in parentheses indicate relative change with respect to this baseline. We ablate the effects of frame acceleration (FA), different MoE designs—including No MoE (no expert module), Res-MoE (our residual formulation), Full-MoE (jointly trained base and expert networks), and router configurations including speed-aware router (SAR), dynamic expert-assignment (DEA), and hard expert selection strategies (Top1 and Top2).

high-dynamic performance without sacrificing low-dynamic accuracy. Router-wise, SAR proves crucial, as its absence (Row 4) leads to notable drops. DEA (Row 8) surpasses both Top1 (Row 6), which overloads a single expert, and Top2 (Row 7), which involves unnecessary experts. DEA better balances expert capacity based on motion intensity.

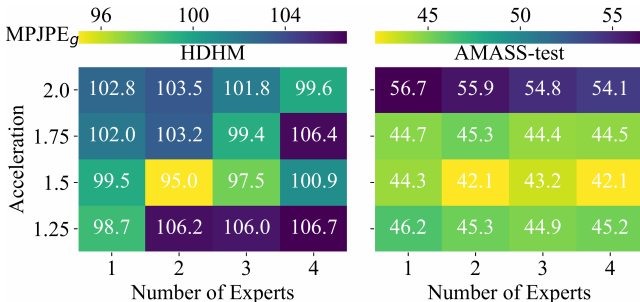


Figure 8: Heatmaps of MPJPE_g across total number of experts and maximum acceleration factor for the HDHM and AMASS-test datasets.

Impact of frame acceleration and number of experts.

Figure 8 presents heatmaps analyzing the impact of two key design choices: the maximum random frame acceleration factor and the total number of experts in the residual MoE. The results show a clear trend that both insufficient or excessive acceleration, as well as too few or too many experts, can lead to suboptimal performance. On the one hand, a maximum acceleration factor that is too low fails to introduce sufficient high-dynamic diversity, while excessively large factors may generate physically implausible motions that hinder learning. On the other hand, using only one expert overloads it with a wide range of dynamics, while having too many experts results in over-fragmented expert assignments, making it harder for each expert to learn meaningful patterns.

Fine-tuning data. Table 4 compares full-data fine-tuning with fine-tuning only on failure cases. Fine-tuning on failure data achieves slightly better performance on both the

Data	HDHM		AMASS-test		AMASS-train
	Success \uparrow	MPJPE _g \downarrow	Success \uparrow	MPJPE _g \downarrow	Training time
Full	95.1%	99.2	100%	44.4	35 h
Failure	95.6%	95.0	100%	42.1	6 h

Table 4: Comparison of fine-tuning on full data versus failure data. Fine-tuning on failure data achieves slightly higher success rates, lower MPJPE_g, and dramatically faster convergence, demonstrating the efficiency of targeted fine-tuning on challenging samples. All experiments are conducted on a single RTX 4090 GPU.

HDHM and AMASS-test datasets, with higher success rates and lower MPJPE_g. Notably, this targeted approach also reduces the training time from 35 hours to just 6 hours, representing a substantial gain in efficiency. These results suggest that difficult samples provide more informative gradients and that focusing training on such challenging segments can accelerate convergence while improving final performance.

5 Conclusion

We present FARM, a high-dynamic humanoid control framework that combines frame-accelerated augmentation with a residual MoE architecture. To evaluate high-dynamic performance, we curate the HDHM dataset, which contains diverse, high-velocity motion clips collected from multiple sources. Experimental results demonstrate that FARM significantly improves tracking accuracy on HDHM while maintaining strong generalization on AMASS. Despite these advances, our framework still exhibits failure cases on HDHM, particularly on motions with subtle artifacts such as ground penetration, floating, or joint jitter. While these artifacts may appear visually acceptable, they pose non-trivial challenges for controllers. Future work will investigate how to eliminate such artifacts at the data generation stage and explore architectural improvements to enhance robustness against noise. Another important direction is to deploy the FARM framework on real-world humanoid robots.

Acknowledgements

This work was supported by the 1+1+1: HKUST–HKUST(GZ) “1+1+1” Joint Funding Program (G042), the National Natural Science Foundation of China (Grant No. 62405255), and BrainCo Inc.

References

- Chen, Z.; Ji, M.; Cheng, X.; Peng, X.; Peng, X. B.; and Wang, X. 2025. GMT: General Motion Tracking for Humanoid Whole-Body Control. *arXiv preprint arXiv:2506.14770*.
- Da Silva, M.; Abe, Y.; and Popović, J. 2008. Simulation of human motion data using short-horizon model-predictive control. In *Computer Graphics Forum*, volume 27, 371–380. Wiley Online Library.
- Huang, R.; Zhu, S.; Du, Y.; and Zhao, H. 2025. MoE-LoCo: Mixture of Experts for Multitask Locomotion. *arXiv preprint arXiv:2503.08564*.
- Kaufmann, M.; Song, J.; Guo, C.; Shen, K.; Jiang, T.; Tang, C.; Zárate, J. J.; and Hilliges, O. 2023. Emdb: The electromagnetic database of global 3d human pose and shape in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14632–14643.
- Lee, Y.; Kim, S.; and Lee, J. 2010. Data-driven biped control. In *ACM SIGGRAPH 2010 Papers*, SIGGRAPH '10. New York, NY, USA: Association for Computing Machinery. ISBN 9781450302104.
- Li, R.; Yang, S.; Ross, D. A.; and Kanazawa, A. 2021. Learn to dance with aist++: Music conditioned 3d dance generation. *arXiv preprint arXiv:2101.08779*, 2(3).
- Lin, J.; Zeng, A.; Lu, S.; Cai, Y.; Zhang, R.; Wang, H.; and Zhang, L. 2023. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36: 25268–25280.
- Luo, Z.; Cao, J.; Kitani, K.; Xu, W.; et al. 2023. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10895–10904.
- Luo, Z.; Cao, J.; Merel, J.; Winkler, A.; Huang, J.; Kitani, K. M.; and Xu, W. 2024. Universal Humanoid Motion Representations for Physics-Based Control. In *The Twelfth International Conference on Learning Representations*.
- Luo, Z.; Hachiuma, R.; Yuan, Y.; and Kitani, K. 2021. Dynamics-regulated kinematic policy for egocentric pose estimation. *Advances in Neural Information Processing Systems*, 34: 25019–25032.
- Luo, Z.; Yuan, Y.; and Kitani, K. M. 2022. From universal humanoid control to automatic physically valid character creation. *arXiv preprint arXiv:2206.09286*.
- Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *International Conference on Computer Vision, ICCV*, 5442–5451.
- Mittal, M.; Yu, C.; Yu, Q.; Liu, J.; Rudin, N.; Hoeller, D.; Yuan, J. L.; Singh, R.; Guo, Y.; Mazhar, H.; Mandlekar, A.; Babich, B.; State, G.; Hutter, M.; and Garg, A. 2023. Orbit: A Unified Simulation Framework for Interactive Robot Learning Environments. *IEEE Robotics and Automation Letters*, 8(6): 3740–3747.
- Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, 1928–1937. PmLR.
- Mu, S.; and Lin, S. 2025. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *arXiv preprint arXiv:2503.07137*.
- Peng, X. B.; Abbeel, P.; Levine, S.; and Van de Panne, M. 2018. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4): 1–14.
- Peng, X. B.; Ma, Z.; Abbeel, P.; Levine, S.; and Kanazawa, A. 2021. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4): 1–20.
- Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Shen, Z.; Pi, H.; Xia, Y.; Cen, Z.; Peng, S.; Hu, Z.; Bao, H.; Hu, R.; and Zhou, X. 2024. World-Grounded Human Motion Recovery via Gravity-View Coordinates. In *SIGGRAPH Asia Conference Proceedings*.
- Tessler, C.; Guo, Y.; Nabati, O.; Chechik, G.; and Peng, X. B. 2024. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics (TOG)*, 43(6): 1–21.
- Tevet, G.; Raab, S.; Cohan, S.; Reda, D.; Luo, Z.; Peng, X. B.; Bermano, A. H.; and van de Panne, M. 2024. CLOSD: Closing the Loop between Simulation and Diffusion for multi-task character control. *arXiv:2410.03441*.
- Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; and Bermano, A. H. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*.
- Tsuchida, S.; Fukayama, S.; Hamasaki, M.; and Goto, M. 2019. AIST Dance Video Database: Multi-Genre, Multi-Dancer, and Multi-Camera Database for Dance Information Processing. In *ISMIR*, volume 1, 6.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, D.; Wang, X.; Liu, X.; Shi, J.; Zhao, Y.; Bai, C.; and Li, X. 2025. MoRE: Mixture of Residual Experts for Humanoid Lifelike Gaits Learning on Complex Terrains. *arXiv preprint arXiv:2506.08840*.
- Wang, T.; Guo, Y.; Shugrina, M.; and Fidler, S. 2020. Unicorn: Universal neural controller for physics-based character motion. *arXiv preprint arXiv:2011.15119*.
- Yuan, Y.; and Kitani, K. 2020. Residual force control for agile human behavior imitation and extended motion synthesis. *Advances in Neural Information Processing Systems*, 33: 21763–21774.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhu, Q.; Zhang, H.; Lan, M.; and Han, L. 2023. Neural categorical priors for physics-based character control. *ACM Transactions on Graphics (TOG)*, 42(6): 1–16.