

Scaling Behavior Foundation Model for Humanoid Robots

Weishuai Zeng^{1,6,7,*} Kangning Yin^{2,7,*} Xiaojie Niu^{7,*} Shunlin Lu⁷ Weixiang Zhong⁷
Jiahe Chen^{3,7} Feiyu Jia⁷ Xiao Chen⁷ Zirui Wang^{3,7} Furui Xu⁷ Ming Zhou⁷ Kailin Li⁷
Weinan Zhang^{2,7} He Wang^{4,6} Li Yi^{5,6} Dahua Lin^{1,7} Jiangmiao Pang⁷ Jingbo Wang⁷

¹The Chinese University of Hong Kong ²Shanghai Jiao Tong University ³Zhejiang University ⁴Peking University
⁵Tsinghua University ⁶Galbot ⁷Shanghai Artificial Intelligence Laboratory

*Equal Contribution

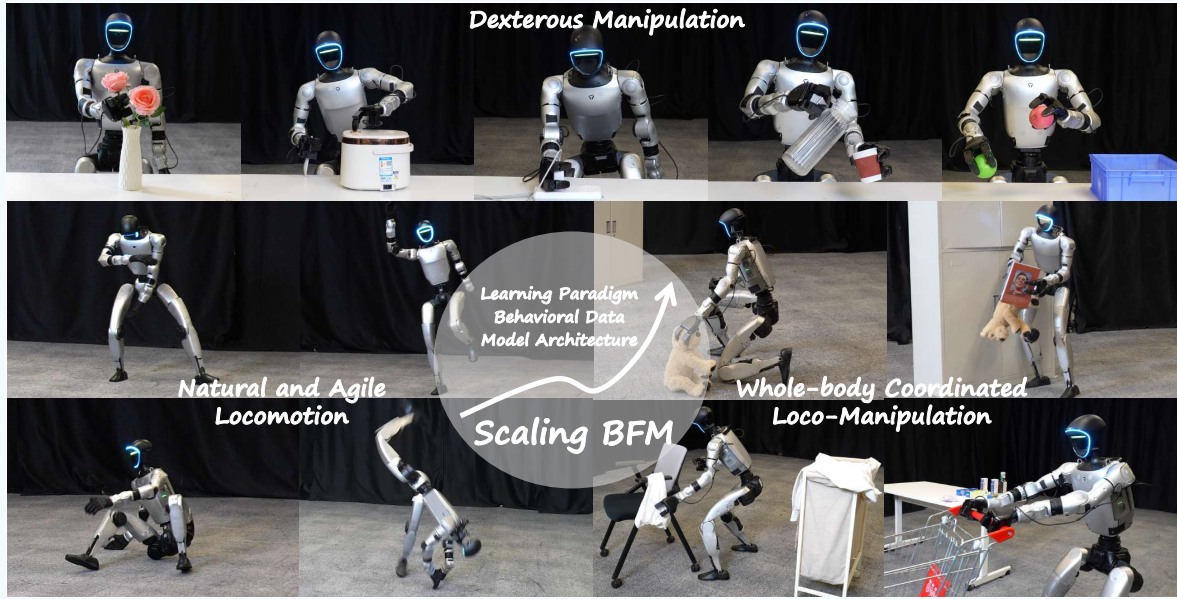


Figure 1: Our **Behavior Foundation Model** enables versatile humanoid behaviors across a broad spectrum of tasks, including dexterous manipulation, natural and agile locomotion and whole-body coordinated loco-manipulation, which highlights its potential as a foundation for general-purpose humanoid control.

Website: scalebfm.github.io

Abstract

Humanoid control requires natural whole-body coordination, precise real-time responses to control signals, and robust generalization across diverse environmental contexts, making it a cornerstone for generalist embodied agents. Behavior Foundation Models (BFMs) have recently emerged as a promising solution to address these challenges by leveraging large-scale behavioral data to achieve superior expressiveness, versatility and generalization. However, despite growing interest in scaling BFMs to further improve their capabilities, it remains unclear how key factors, including the learning paradigm, behavioral data and model architecture should be coordinated to enable effective scaling. In this work, we revisit the scaling recipe for BFMs and demonstrate that substantial performance

gains can be achieved through the coordination of three core components: 1) the learning paradigm of motion tracking that reformulates diverse humanoid control problems as the reproduction of integrated whole-body behaviors in the global frame; 2) the strategic synergy between on-policy rollout quantity and reference motion diversity; and 3) the expressive and scalable model architecture termed Humanoid Transformer that facilitates the natural emergence of structured behavioral representations. Through extensive experiments in both simulation and real-world deployment, we demonstrate that our approach yields significant improvements in control fidelity and task generalization, reducing Mean Per-Keypoint Position Error (MPKPE) on the test set by over 10% in local mode and 82% in global mode compared with existing humanoid controllers. These results establish BFM as a principled and effective foundation for scalable and general-purpose humanoid control.

1 Introduction

Humanoid robots are increasingly viewed as a promising general-purpose embodiment for operating in human-centric environments, owing to their rich actuation, human-like morphology and natural compatibility with daily infrastructures designed for humans. Realizing this potential, however, depends on effective whole-body control that requires the coordinated regulation of numerous degrees of freedom to achieve natural, responsive and robust behaviors across multiple tasks and environmental contexts.

Recent years have witnessed substantial progress in humanoid whole-body control, enabling impressive capabilities across diverse motor skills [1, 2, 3, 4, 5]. Nevertheless, these systems remain largely task-specific, often relying on extensive reward engineering tailored to individual tasks and contextual settings. This specialization limits their versatility and hinders generalization across diverse scenarios.

To address these limitations, recent research has identified behavior as a shared objective across multiple whole-body control tasks [6] and introduced **Behavior Foundation Models (BFMs)**, large-scale pretrained humanoid foundation controllers capable of robust and versatile humanoid control in response to diverse behavior specifications [7, 8, 9, 10, 11]. By decoupling underlying behaviors from task-specific control modes, BFMs represent a paradigm shift from isolated task learning toward holistic behavior learning. In this work, we focus on a special line of BFMs that incorporate an explicit control interface [6, 7, 11]. This design not only enables direct responses to behavioral specifications at deployment, but also promotes the emergence of structured latent representations that organize explicit behavioral intentions within a coherent latent space. Such reusable behavioral representations facilitate strong generalization across diverse specifications and efficient adaptation to downstream applications.

Inspired by the success of foundation models in other domains [12, 13, 14], a natural question arises: *Can behavior foundation models for humanoid robots be effectively scaled to further enhance their capabilities?* Recent studies have taken initial steps toward scaling BFMs [11]. However, these efforts remain fragmented, offering only preliminary conclusions while leaving concrete design principles for effective scaling largely unexplored. To bridge this gap, we revisit current BFM pretraining pipelines and develop a principled scaling recipe that coordinates multiple factors, including a unified and efficient learning paradigm, large-scale and diverse behavioral data, and an expressive yet scalable model architecture.

Our framework identifies motion tracking [15] as a unified and scalable paradigm which reformulates diverse humanoid control problems as the imitation of reference motions, thereby providing dense whole-body guidance for efficient behavior learning. While sharing a similar vision with prior work [9, 11, 16], we adopt a distinct training recipe that requires the robot to reproduce reference motions as *integrated*

whole-body trajectories in the global frame. This design not only mitigates the behavioral ambiguity caused by ignoring root translation, under which behaviors such as walking forward and marching in place may become indistinguishable despite reflecting different behavioral intentions. It also avoids decoupling global root evolution from local pose tracking, allowing deviations in root motion to propagate to the entire body and hence providing stronger guidance for behavior learning. Notably, this formulation preserves flexibility in behavioral specification: the same BFM can support both global control with root localization and local control by re-anchoring the control signals to the robot’s current root state.

Under this learning paradigm, we argue that scaling BFMs is fundamentally driven by the non-trivial interplay between the quantity and diversity of training data. We instantiate motion tracking with Proximal Policy Optimization (PPO) [17], given its mature ecosystem in humanoid control, while noting that alternative algorithms have also shown promising results [18, 19]. Existing studies often equate training data scaling with increasing the number of reference motions. However, under the PPO framework, effective training data are the on-policy rollouts collected through environment interaction. Therefore, data quantity is more directly governed by factors such as the number of parallel environments and the rollout horizon, whereas increasing the number of reference motions primarily shapes the behavioral distribution and enriches its diversity when properly curated. We thus posit that data scaling for BFM pretraining emerges from the synergy between quantity and diversity: quantity is obtained by scaling on-policy data collection, while diversity is achieved by expanding the reference motion corpus.

Finally, we emphasize that an expressive and scalable model architecture is essential for scaling BFMs. While existing works primarily adopt MLPs as the backbone, and recent studies have shown that scaling MLPs can improve performance [11], we investigate whether more expressive architectures exhibit stronger scaling behavior. Beyond architectural design, we further examine whether structured latent representations of behavioral intentions can naturally emerge without relying on auxiliary objectives that may encode human priors. Such representations could facilitate the flexible utilization and efficient transfer of behavioral knowledge learned by BFMs, thereby supporting diverse downstream applications.

Overall, this paper presents a systematic investigation into the scaling behavior of BFMs by revisiting three key ingredients of the scaling recipe: the learning paradigm, behavioral data, and model architecture. Through comprehensive experiments in both simulation and real-world deployment, we validate our findings and demonstrate that our approach significantly improves control fidelity, execution robustness, and generalization capability, reducing Mean Per-Keypoint Position Error (MPKPE) on the test set by over 10% in local mode and 82% in global mode compared with existing humanoid controllers. These results establish our BFM as a principled foundation for scalable, general-purpose humanoid control.

2 Related Work

2.1 Humanoid Whole-body Control

Humanoid control has recently undergone a paradigm shift from task-specific reward engineering toward whole-body reference imitation. Early approaches achieved impressive performance across diverse motor skills through carefully designed reward functions and regularization terms [1, 3, 4]. However, their reliance on task-specific objectives often constrains generalization across heterogeneous control settings. In contrast, mimic-based approaches formulate humanoid control as reference motion track-

ing, providing a unified framework for learning natural, expressive, and coordinated whole-body behaviors [20, 21, 22]. This paradigm has enabled wide applications, including human teleoperation [23, 24, 25], humanoid-scene interaction [26, 27, 28], and humanoid-object interaction [29, 30, 31].

Despite these advances, conventional motion-tracking-based approaches remain constrained by their dependence on explicit whole-body reference motions as the primary control interface. To address this limitation, recent research has explored two complementary directions. The first extends humanoid control to multimodal commands, either preserving whole-body motions as an intermediate representation and then training high-level motion generators for translating multimodal inputs into trackable reference trajectories [32, 33, 34], or developing more flexible control interfaces that can directly respond to diverse command forms without requiring explicit whole-body references during inference [6, 8]. The second direction focuses on learning reusable behavioral representations, where structured latent spaces are acquired through motion tracking to capture diverse motor skills and behavioral priors in a compact and transferable form [35, 36]. Such representations can subsequently be leveraged across downstream tasks, facilitating efficient skill acquisition and rapid adaptation to varying contexts.

2.2 Behavior Foundation Models for Humanoid Robots

Behavior Foundation Models (BFMs) have emerged as a promising paradigm for general-purpose humanoid control, owing to their broad versatility and strong generalization. Recent studies have explored diverse pretraining strategies and control mechanisms for constructing such models. BFM4Humanoid [6] pretrains BFMs within the DAgger framework [37], where actions are labeled by an expert policy trained via motion tracking. It further introduces an explicit control interface, enabling the pretrained model to directly respond to diverse behavior specifications at deployment. Based on a similar distillation framework, BeyondMimic [9] first learns structured behavioral representations as a proxy action space and then performs pretraining through supervised learning on offline behavioral trajectories. At inference time, it applies classifier guidance [38] to the diffusion-based BFM to generate task-conditioned behavior trajectories for humanoid control [39]. BFM-Zero [10] pretrains BFMs via unsupervised reinforcement learning based on forward-backward representations [40, 41]. With the promptable latent space learned through pretraining, it enables zero-shot generalization and few-shot adaptation through sampling-based inference at test time. More recently, SONIC [11] explores the direct usage of motion tracking for BFM pretraining. By learning discrete behavioral representations and enforcing alignment across multiple command forms, it enables versatile control through the resulting shared latent space.

As BFMs continue to demonstrate strong performance across diverse humanoid control settings, scaling has naturally emerged as a key direction for further improving their capabilities. However, their concrete scaling behavior has yet to be systematically identified. While SONIC offers preliminary evidence by scaling the number of reference motions, model capacity, and GPU hours within the motion-tracking framework, it fails to elucidate the coupling among key factors that govern effective scaling, including the learning paradigm, training data, and model architecture. In this work, we revisit the design of motion tracking, clarify the notion of training data in PPO-based pretraining, and introduce a more expressive architecture termed the Humanoid Transformer. These explorations aim to offer a more principled understanding of BFM scaling and lay a stronger foundation for general humanoid control.

3 Method

In this section, we first present our revised formulation and learning objective for BFM pretraining and establish motion tracking as its principled and scalable instantiation. We then detail our motion-tracking recipe, including the state representation, control interface, reward design, and domain randomization strategy. Next, we formalize the notion of training data within our framework and elucidate the corresponding dataset construction pipeline to achieve a synergy between data quantity and diversity. Finally, we introduce a more expressive and scalable architecture termed the *Humanoid Transformer*, which facilitates the natural emergence of structured behavioral representations. Collectively, these components constitute the methodological foundation of our scaling recipe for BFM pretraining.

3.1 Problem Formulation

We formulate humanoid control as a Markov Decision Process (MDP) $M = \langle S, A, T, R, \gamma \rangle$, where S denotes the state space, A the action space, T the transition dynamics, R the reward function, and γ the discount factor. At each timestep t , the state $\mathbf{s}_t \in S$ comprises the agent’s proprioceptive state $\mathbf{s}_t^p$ and a goal state $\mathbf{s}_t^g$ that specifies the control objective. The action $\mathbf{a}_t \in A$ represents desired joint angles, which are then executed by a low-level proportional-derivative (PD) controller to actuate the robot.

Following prior work [6], we identify humanoid behavior $\mathbf{B}$ as the common objective underlying existing WBC systems and formally define it as a *trajectory over the humanoid proprioceptive states $\mathbf{s}_t^p$ and actions $\mathbf{a}_t$* . Notably, goal states $\mathbf{s}_t^g$ are excluded from this definition, as we interpret them as external specifications that guide or constrain behavior rather than the intrinsic components of behavior itself.

$$\mathbf{B} \triangleq (\mathbf{s}_0^p, \mathbf{a}_0, \mathbf{s}_1^p, \mathbf{a}_1, \dots, \mathbf{s}_{T-1}^p, \mathbf{a}_{T-1}, \mathbf{s}_T^p) \quad (3.1)$$

Building upon the formulation above, BFMs intend to decouple behavior learning from task-specific control modes, thereby establishing a unified and scalable objective for learning various behaviors while enabling a versatile control interface that accommodates diverse behavioral specifications of the same underlying behavior. Formally, we structure the pretraining of BFMs as a goal-conditioned reinforcement learning (GCRL) problem, the learning objective of which is to maximize the expected return:

$$\max_{\theta} E_{\tau \sim p_{\theta}(\tau)} \left[\sum_{t=0}^T \gamma^t r_t \right], \quad (3.2)$$

where $p_{\theta}(\tau) = p(\mathbf{s}_0) \prod_{t=0}^T p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t) \pi_{\theta}(\mathbf{a}_t | \mathbf{s}_t)$ is the distribution over trajectories induced by the model π_{θ} . Existing WBC systems typically tailor goal states $\mathbf{s}_t^g$ and reward functions r_t to individual task requirements, failing to establish a unified abstraction across diverse scenarios. In contrast, BFMs aim to 1) identify a proxy task coupled with a general reward formulation that can be shared across a wide range of behavior patterns and control settings; 2) construct a versatile control interface of goal states $\mathbf{s}_t^g$ from diverse forms of behavioral specifications that describe the same underlying behavior.

Intuitively, imitation naturally serves as a general learning paradigm for behavior acquisition, as it aims to reproduce demonstrated trajectories without relying on task-specific requirements or manually designed evaluation protocols. Therefore, we instantiate the proxy task for BFM pretraining as humanoid whole-body motion tracking, where the reward function evaluates how faithfully the humanoid follows the reference motions. This formulation casts behavior learning as a well-defined and measurable optimization problem, thereby establishing a unified and scalable objective across diverse scenarios.

Despite the shared usage of reference motions, the formulation above differs fundamentally from general motion trackers. BFM s employ motion tracking solely as a proxy task for behavior learning rather than as the intended control objective. As a consequence, they are not constrained to a specific form of behavioral specification, whereas conventional motion trackers typically treat whole-body reference states as the only way of applying control and therefore remain specialized for a specific WBC task.

In practice, we optimize the objective of BFM pretraining using the Proximal Policy Optimization (PPO) algorithm [17] for its stability and robustness. Nevertheless, the proposed formulation is algorithm-agnostic and can be readily integrated with a wide range of reinforcement learning algorithms.

3.2 Learning Paradigm

Motion tracking has been widely adopted as an effective learning paradigm for humanoid control. Existing approaches, however, differ substantially in their detailed designs, including state representations, control interfaces, and reward formulations. To enable unified and scalable behavior learning together with versatile humanoid control through diverse forms of behavioral specification, we carefully curate a motion-tracking recipe for BFM pretraining, the key components of which are described below.

Proprioceptive State Design. We adopt the asymmetric actor-critic design in PPO to jointly account for efficient policy learning and practical real-world deployment. The proprioceptive state of actor is defined as $\mathbf{s}_t^p \triangleq (\omega_t^{\text{root}}, g_t, q_t, \dot{q}_t)$ where ω_t is the base angular velocity, g_t the projected gravity, q_t current joint position and $\dot{q}_t$ current joint velocity. In contrast, critic’s proprioception may leverage privileged information in simulators, which is defined as $\hat{\mathbf{s}}_t^p \triangleq (h_t, p_t, \theta_t, v_t, \omega_t, q_t, \dot{q}_t)$ where h_t is the root height, p_t the link positions, θ_t the link orientations, v_t the link linear velocities and ω_t the link angular velocities. All these whole-body states of the critic are expressed in humanoid’s heading frame.

Control Interface Design. To enable versatile humanoid control from diverse forms of behavioral specifications, we design a simple yet expressive control interface based on *masked whole-body target poses in the root-relative Cartesian space*. Specifically, the whole-body target representation is defined as: $\mathbf{x}_t \triangleq (p_{t+1}^{\text{ref}} - p_t^{\text{root}}, p_{t+1}^{\text{ref}} - p_t, \theta_{t+1}^{\text{ref}} \ominus \theta_t^{\text{root}}, \theta_{t+1}^{\text{ref}} \ominus \theta_t)$, which contains the target link poses and their deviations from the current poses, all expressed in the humanoid root frame. We then randomly sample a link-wise mask from a predefined mask set $\mathcal{M} = \{\mathbf{m}_0, \mathbf{m}_1, \dots, \mathbf{m}_7\}$ and construct the actor goal state as $\mathbf{s}_t^g \triangleq (\mathbf{x}_t \odot \mathbf{m}_i, \mathbf{m}_i)$, $\mathbf{m}_i \sim \mathcal{M}$. The masks specify eight control modes, the details of which are provided in the Appendix A.1. In contrast, the critic is provided with complete, unmasked target information to provide more accurate value estimation. Accordingly, we define the critic goal state in the humanoid root frame as $\hat{\mathbf{s}}_t^g \triangleq (p_{t+1}^{\text{ref}} - p_t^{\text{root}}, p_{t+1}^{\text{ref}} - p_t, \theta_{t+1}^{\text{ref}} \ominus \theta_t^{\text{root}}, \theta_{t+1}^{\text{ref}} \ominus \theta_t, v_{t+1}^{\text{ref}} - v_t, \omega_{t+1}^{\text{ref}} - \omega_t)$. This asymmetric design is consistent with BFM objective of decoupling behavior learning from concrete specifications, allowing the critic to focus exclusively on motion tracking as the proxy task for BFM pretraining.

Reward Design. We formulate the reward as a weighted sum of tracking rewards, penalty terms and survival terms. Unlike prior works that either remove root-position tracking or decouple root following from whole-body pose tracking [9, 11], we require the robot to reproduce reference motions as integrated whole-body trajectories in the global frame. Removing root-position tracking may assign nearly identical learning guidance to behaviors with distinct global semantics, while decoupling root and pose tracking may compromise the coordination of root and pose evolution. Our design instead provides coherent guidance over whole-body behavior, reducing guidance ambiguity and encouraging

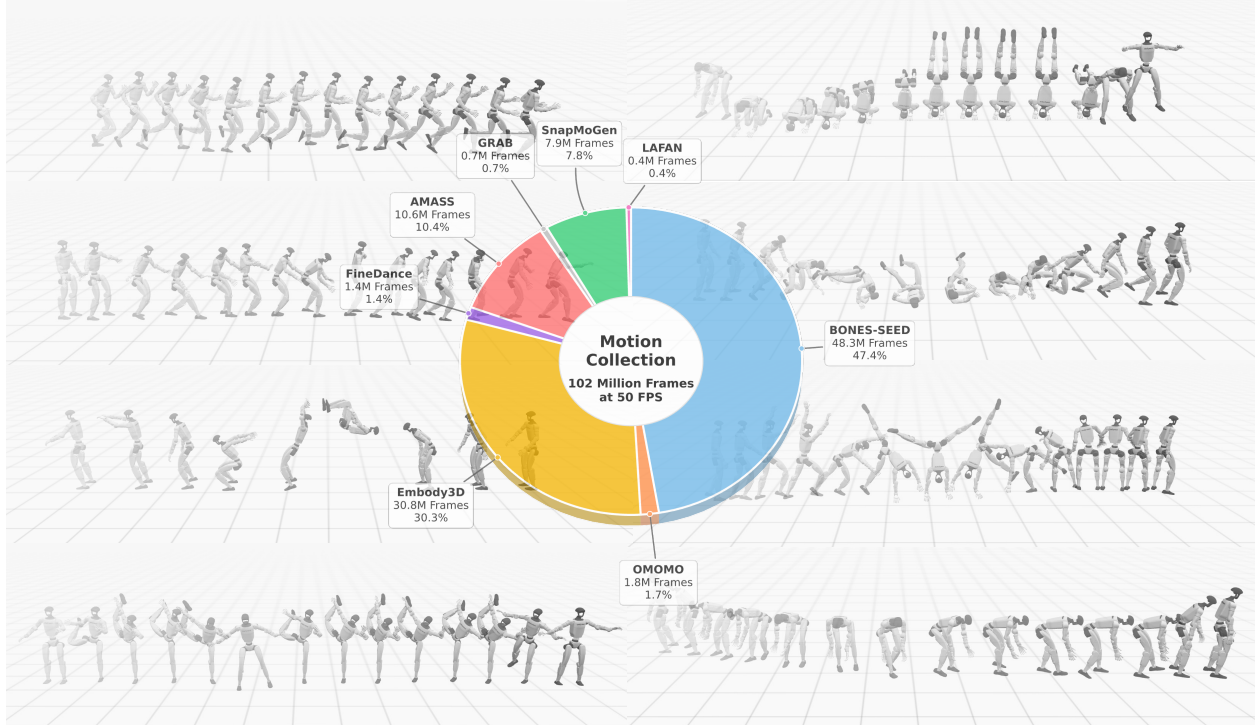


Figure 2: Reference Motion Dataset Composition. We construct a large-scale human motion dataset by aggregating over 102 million frames at 50 FPS from multiple sources, which are subsequently retargeted to the target humanoid. Random samples are visualized to demonstrate its broad coverage of diverse behavior patterns.

more consistent behavior acquisition. Details on reward components are provided in Appendix A.2.

Domain Randomization. To enhance the robustness and generalization of our BFMs, we apply domain randomization during training. We randomize physical parameters including friction coefficient, joint default positions, base center-of-mass positions and mass of hands particularly to accommodate potential variations in end-effector configurations. We also periodically apply random perturbations to robot velocities to simulate external pushes. Detailed components are provided in Appendix A.3.

3.3 Training Data

After establishing motion tracking with versatile control interface as the learning paradigm, we next describe how the training data is collected and organized for scalable BFM pretraining. Existing studies often equate scaling training data with simply increasing the number of reference motions. Under the PPO framework, however, the effective training data are the on-policy rollouts collected through environment interaction, the scale of which is primarily determined by the degree of environment parallelism and the rollout horizon. Increasing the number of reference motions instead influences the distribution of these rollouts by introducing a wider range of behavior patterns for imitation. This benefit, however, depends critically on the diversity of the reference motion corpus rather than its size alone. For instance, a dataset containing an arbitrarily large number of walking forward motions still represents only a single behavior pattern and therefore provides limited behavioral coverage. Consequently, scalable BFM pretraining requires the synergy of quantity and diversity of the training data. In practice,

we increase the number of parallel environments and rollout horizon to collect more on-policy data, while curating reference motions from diverse sources to ensure broad behavioral coverage.

Human Motion Collection. To expand the behavioral coverage of reference motions, we first construct a large-scale human motion collection with fully open-source datasets, including LAFAN [42], AMASS [43], OMOMO [44], GRAB [45], SnapMoGen [46], FineDance [47], BONES-SEED [48], and Embody3D [49]. This collection contains 102M frames at 50 FPS, which is expected to contain rich and coordinated behavior patterns for BFM pretraining. Detailed components may refer to figure 2.

Motion Retargeting. We adopt a two-stage retargeting pipeline to bridge the embodiment gap between human and humanoid. In the first stage, we align the human and humanoid skeleton by optimizing the human shape coefficient β to minimize the difference of corresponding links at the same rest pose. Notably, our reference motion corpus comprises heterogeneous human motion representations, including both SMPL and BVH formats. For SMPL motions, we optimize the built-in shape coefficients, whereas for BVH motions, we optimize the local skeletal offsets. Human skeletons from diverse sources are thereby aligned with the target humanoid morphology. In the second stage, we follow prior work and retarget human motions to the humanoid in a sequential frame-wise manner [50]. For each frame t , the humanoid root pose and joint configurations are initialized from the solution of the previous frame and subsequently optimized by solving an inverse kinematics problem that minimizes the weighted pose discrepancy between selected links. Visualization of retargeted motions is also presented in figure 2.

Early Termination and Reference State Initialization. Early termination and reference state initialization (RSI) jointly shape the on-policy data distribution for BFM pretraining. We terminate an episode whenever any humanoid link deviates by more than 0.5 m from the reference motion in the global frame, thereby preventing the collection of low-quality rollout data. After termination, we adopt the RSI strategy [15], where the initial states are directly derived from the reference motions. Rather than randomly selecting an arbitrary reference frame, we reinitialize the humanoid with the reference state next to the termination point, bringing it back to the desired trajectory manifold from which it departs.

Adaptive Sampling. Adaptive sampling intends to bias the data distribution toward challenging motion sequences to improve sample efficiency while maintaining sufficient behavioral coverage to avoid catastrophic forgetting. Inspired by the implementation of ProtoMotions [51], we assign a uniform sampling weight $w_0 = 1$ to each motion sequence at the beginning of training and then periodically evaluate the model on the full training set every T_{eval} epochs. A trajectory is marked as failed if any humanoid link at any timestep deviates by more than 0.5 m from the reference in the global frame. Given a discount factor β where $0 < \beta < 1$, the sampling weight of each trajectory after the t -th evaluation is updated as

$$w_t = \begin{cases} w_{t-1} / \beta^{T_{\text{eval}}}, & \text{if the trajectory fails,} \\ w_{t-1} \cdot \beta^{T_{\text{eval}}}, & \text{if the trajectory succeeds.} \end{cases} \quad (3.3)$$

To avoid excessive concentration of the sampling distribution, we clamp the updated sampling weights:

$$w_t \leftarrow \text{clip}(w_t, w_{\min}, w_{\max}). \quad (3.4)$$

In practice, we set $T_{\text{eval}} = 200$, $\beta = 0.999$, $w_{\min} = 0.03$, and $w_{\max} = 1.0$ and normalize the weights to obtain the sampling distribution. This forms a closed loop of evaluation, feedback, and resampling, allowing training to focus more on difficult motions while retaining coverage of the reference corpus.

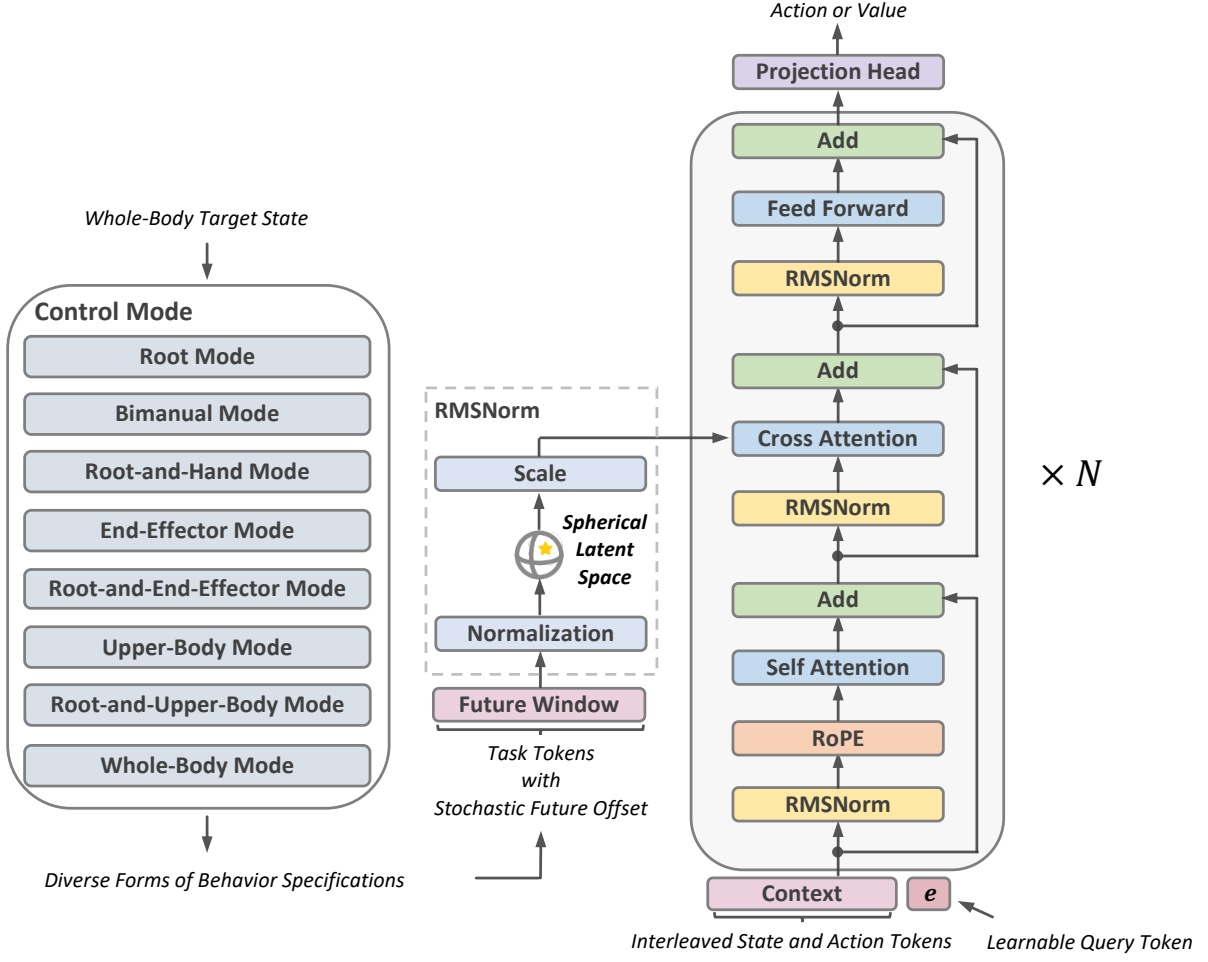


Figure 3: Overview of Model Architecture. We introduce an expressive and scalable architecture termed the *Humanoid Transformer*, which facilitates the natural emergence of structured behavioral representations.

3.4 Model Architecture

We finally introduce the *Humanoid Transformer*, an expressive and scalable architecture tailored for BFM pretraining. To provide richer temporal information for behavior learning, we first extend the proprioceptive states, goal states, and actions into finite temporal windows. These inputs are then encoded by modality-specific tokenizers, the parameters of which are shared across all inputs of the same modality. The resulting proprioception and action tokens are interleaved to form the context sequence, followed by a learnable query token for action or value prediction. During self-attention, the query token is prevented from being attended to by the context tokens, while retaining full attention over the entire context sequence to aggregate historical information. Goal states, together with their temporal offsets relative to the current timestep, are concatenated along the temporal dimension, tokenized, and injected into the backbone through cross-attention, providing a flexible mechanism for conditioning the model on behavioral specifications. An overview of the architecture is presented in figure 3.

Notably, we employ RMSNorm [52] to normalize the goal embeddings onto a continuous and bounded hypersphere, thereby naturally inducing a structured latent representation space for behavioral intentions. Rather than explicitly regularizing this latent space with auxiliary objectives, it is shaped solely through the objective of motion tracking during pretraining. Interestingly, a similar structural design of the latent space has also been explored in generative frameworks for image generation [53].

We employ the same backbone architecture for both the actor and the critic, while adopting different temporal offset schemes. Let index 0 denote the immediate next target frame. The future window of actor contains the next five consecutive frames with indices $\{0, 1, 2, 3, 4\}$, together with an additional frame randomly sampled from the range $[5, 32]$. This *stochastic future offset* provides a practical mechanism for mitigating system delays during real-time deployment, as the last future frame can be dynamically adjusted according to the timestamp of the received control signals. In contrast, the critic adopts fixed exponentially spaced indices $\{0, 1, 2, 4, 8, 16, 32\}$, providing access to both fine-grained short-term dynamics and long-horizon outcomes. This design improves value estimation while maintaining a comparable token budget. More details on the model architecture are provided in Appendix A.4.

4 Experiments

In this section, we first investigate the scaling behavior of BFM from the perspectives of training data and model architecture. Building on these findings, we then analyze the learned latent space to show that structured representations of behavioral intentions emerge naturally without auxiliary objectives and are further refined through scaling. Finally, we benchmark the resulting BFMs against existing humanoid controllers to evaluate their effectiveness as humanoid foundation controllers.

4.1 Experiment Setup

Environments. We conduct experiments on Unitree G1 [54], which stands 1.3 meters tall and has 29 degrees of freedom. All models are trained in IsaacLab [55] and evaluated in MuJoCo [56] to assess robustness under dynamic transfer. Details on real-world deployment are provided in Appendix B.1.

Evaluation Protocol. BFM pretraining is instantiated as motion tracking, which naturally provides direct evaluation protocols by measuring how accurately the humanoid tracks the control signals under diverse control modes. Accordingly, we report the success rate (Succ) together with global and local tracking errors. These metrics are computed only over the activated links specified by each control mode, as the remaining links are left unspecified and are therefore naturally inpainted. A motion sequence is deemed unsuccessful if any selected link deviates from its desired state in the global frame by more than a predefined threshold. G-MPKPE and G-MPKRE denote the mean per-keypoint position and rotation errors of the selected links in the global frame, whereas L-MPKPE and L-MPKRE measure the corresponding root-relative errors after removing the horizontal translation offsets and heading differences [9]. Formal definitions of all the evaluation metrics are provided in Appendix B.2.

Evaluation Dataset. We evaluate all the models on two test sets. The first (BONES Test Set) consists of 10,000 held-out motion sequences sampled from the BONES dataset, forming a source-aligned benchmark of approximately 3.6 million frames at 50 FPS. The second (Ours Test Set) comprises high-quality,

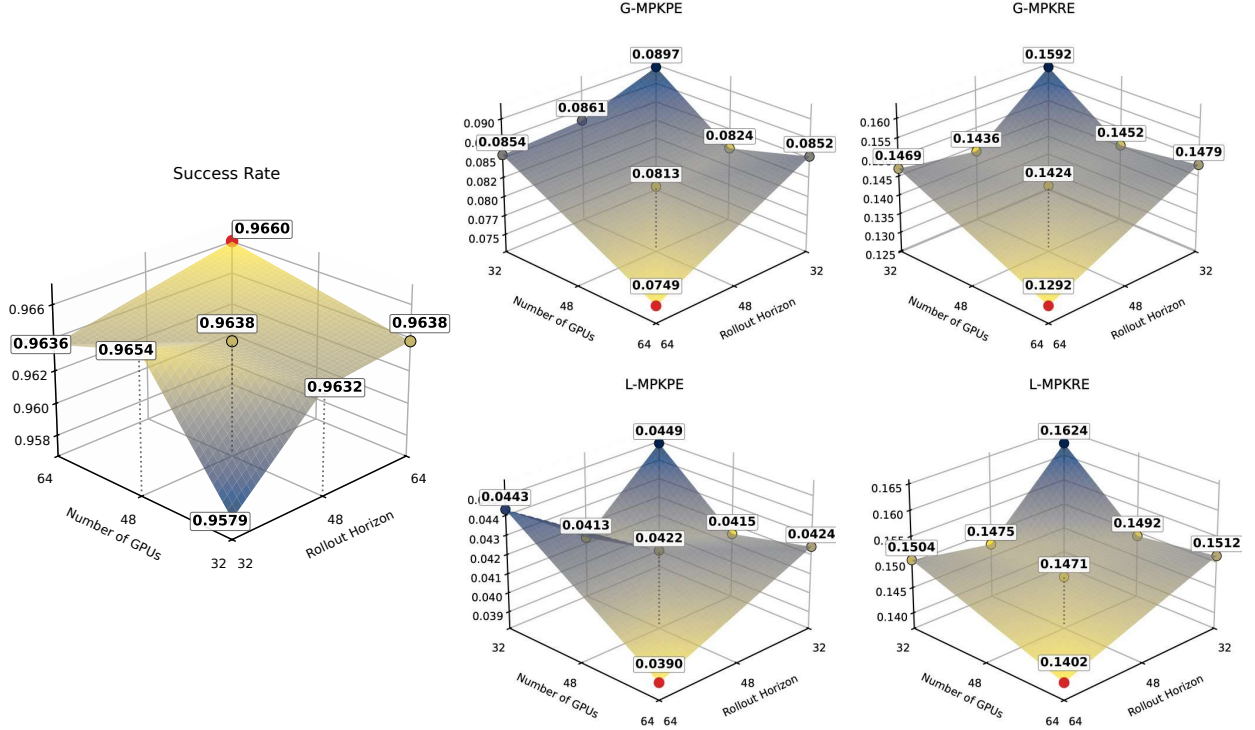


Figure 4: Results of scaling on-policy data collection on the BONES Test Set under whole-body control mode. Best outcomes with the highest success rate and lowest tracking errors are highlighted with red circles.

previously unseen motions, including 839 sequences captured using Xsens [57] and 810 sequences from 100Style [58]. Together, these cross-source sequences comprise around 4.5 million frames at 50 FPS.

4.2 Scaling with Training data

We first study how training data affects the scaling behavior of BFM from two complementary aspects: scaling on-policy data collection and scaling reference motions under the PPO-based motion-tracking framework. Specifically, scaling on-policy data collection primarily increases the *quantity* of training data per update, whereas scaling reference motions is expected to enrich the *diversity* by introducing a broader range of behavioral patterns, thereby shaping the data distribution encountered during training.

4.2.1 Scaling On-policy Data Collection

We scale on-policy data collection by increasing both the *width* and *depth* of environmental interaction. With the number of environments per GPU held constant, the *width* is increased by scaling the number of parallel GPUs, whereas the *depth* is increased by extending the rollout horizon of each environment.

To evaluate the effect of scaling on-policy data collection, we keep all other training configurations fixed across experiments while varying the rollout horizon and the number of GPUs across three levels. We report results on two test sets under four representative control modes. The whole-body control results are presented in figure 4 and 5, whereas the results for the other three control modes, together with the

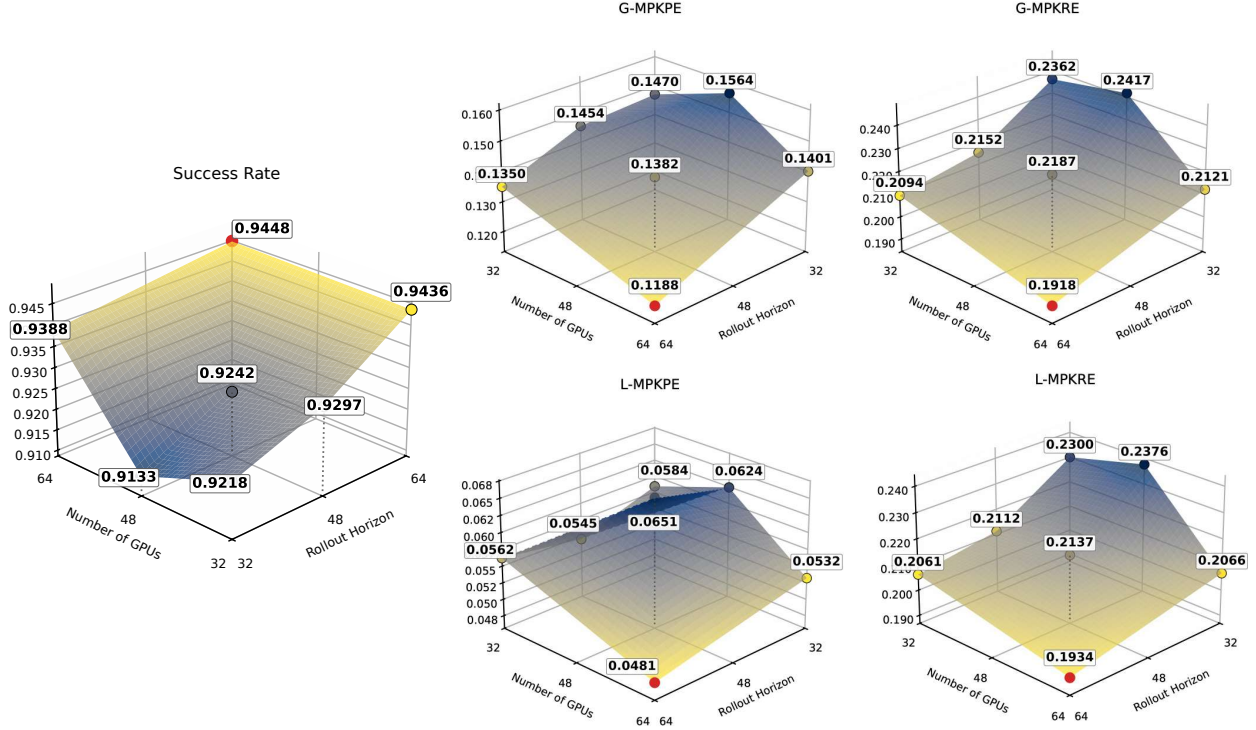


Figure 5: Results of scaling on-policy data collection on the Ours Test Set under whole-body control mode. Best outcomes with the highest success rate and lowest tracking errors are highlighted with red circles.

experimental designs adopted to ensure a fair comparison, are further provided in Appendix B.4.

As shown by the results, jointly scaling the *width* and *depth* dimensions substantially improves performance across diverse control modes, with the largest configuration (64 GPUs and a rollout horizon of 64) achieving the best performance in nearly all experiments. We attribute these gains primarily to the increased amount of on-policy experience collected for each policy update, which provides more reliable estimates of quantities involved in PPO optimization, thereby leading to more effective policy updates.

However, scaling either dimension alone does not always produce a consistent trend of improvement. This observation suggests that the benefits of scaling on-policy data collection depend not only on the total quantity of collected experience but also on how it is accumulated. In other words, effectively scaling on-policy data collection may require an appropriate balance between the *width* and *depth* dimensions, rather than increasing either one in isolation. Understanding the intricate interplay between these two complementary factors remains an interesting direction for future investigation.

4.2.2 Scaling Reference Motions

We evaluate the effect of scaling reference motions by partitioning the full dataset into five nested subsets of increasing size, such that each larger subset fully contains all smaller ones. These subsets are denoted as XXS, XS, S, M, and L. Specifically, XXS, XS, and S correspond to one-third, two-thirds, and the complete BONES-SEED dataset excluding the held-out test split, respectively. The M subset comprises all reference motion datasets except Embody3D, whereas L contains the complete collection.

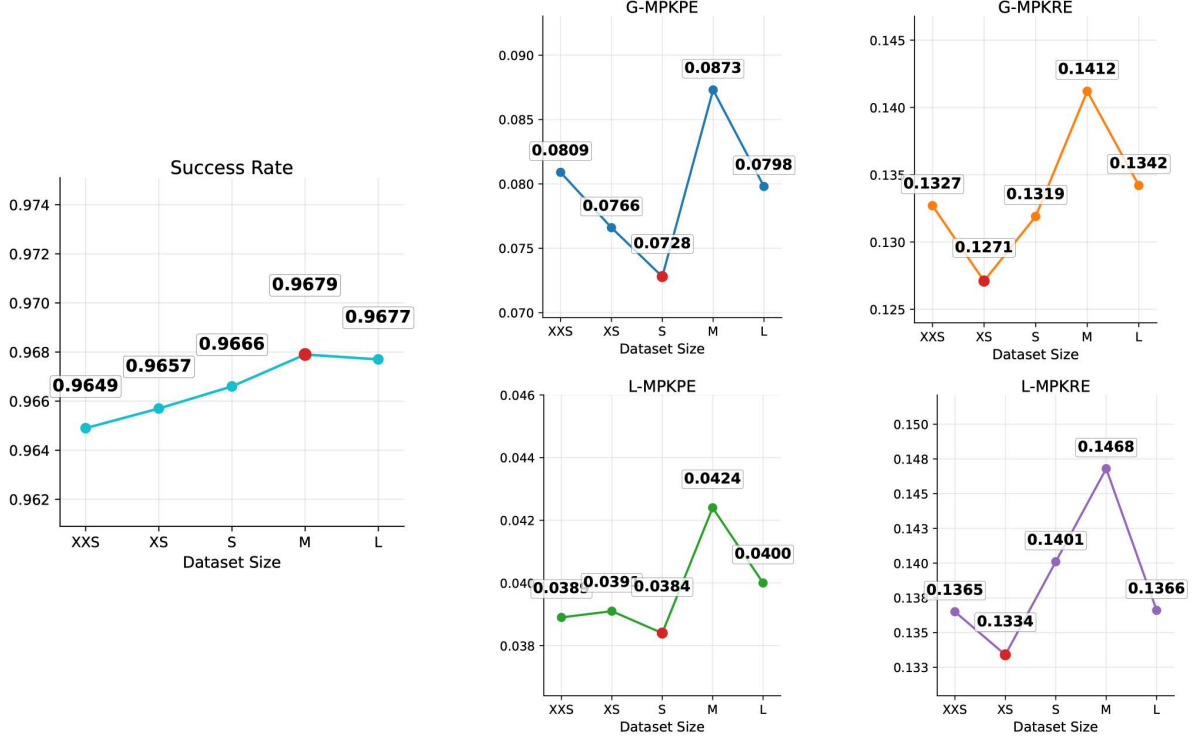


Figure 6: Results of scaling reference motions on the BONES Test Set under whole-body mode. Best outcomes corresponding to the highest success rate and lowest tracking errors are highlighted with red circles.

All other training configurations are kept fixed across experiments, and the optimal on-policy data configuration is adopted throughout. We still report results of two test sets under four representative control modes. The whole-body control results are presented in figure 6 and 7, whereas those for the remaining three control modes, together with the detailed experimental designs are provided in Appendix B.5.

Before interpreting the detailed evaluation results, we first quantitatively analyze the behavioral characteristics of the training partitions. Specifically, we extract behavioral features from all motion frames and fit a clustering model using features randomly sampled from the complete L partition. An equal number of features are subsequently sampled from each partition and assigned to the nearest cluster defined by the shared clustering model. The **occupancy rate** is therefore defined as the ratio of occupied clusters to the total number of clusters.

Partition	Occupancy Rate
XXS	0.9365
XS	0.9350
S	0.9365
M	0.9825
L	0.9995

Table 1: Occupancy Rates of Training Partitions

As shown in table 1, the occupancy rate remains nearly constant from XXS to S, but increases substantially from S to L. This observation indicates that scaling from XXS to S may primarily increase the amount of reference motions while yielding little measurable expansion in the behavioral coverage. In contrast, scaling from S to L not only increases the amount of reference motions but also broadens the range of represented behaviors. Additional details of this analysis are provided in the Appendix B.5.

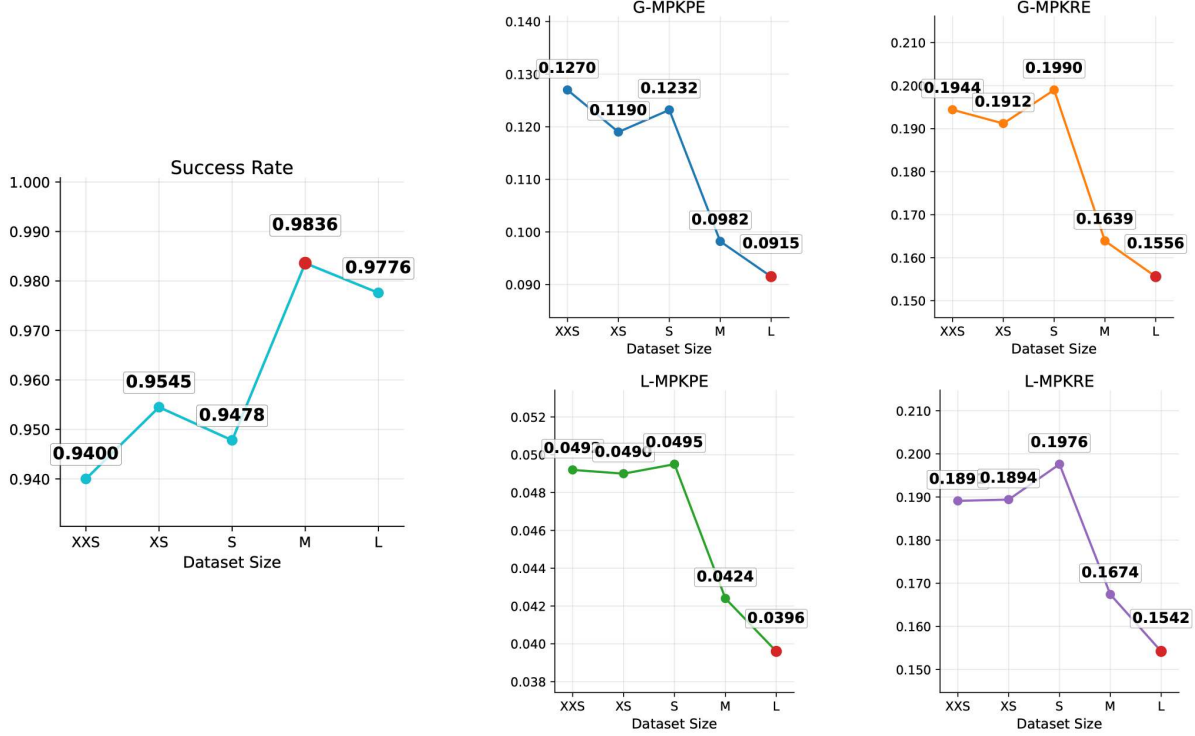


Figure 7: Results of scaling **reference motions** on the **Ours Test Set** under **whole-body mode**. Best outcomes corresponding to the highest success rate and lowest tracking errors are highlighted with red circles.

The behavioral coverage analysis naturally divides reference-motion scaling into two distinct regimes: *homogeneous scaling* increases the amount of reference motions while leaving the measured behavioral coverage nearly unchanged (XXS→S), whereas *heterogeneous scaling* increases both the amount of reference motions and the measured behavioral coverage (S→L). As shown in the experiments, homogeneous scaling yields only marginal improvements on both the BONES and Ours Test Set, despite increasing the amount of reference motions from the same domain. In contrast, heterogeneous scaling produces little additional benefit on the BONES Test Set but leads to substantial performance gains on the Ours Test Set. Notably, significant improvements are observed only when scaling markedly expands measured behavioral coverage, and only on the benchmark containing motions from novel data sources.

These observations suggest that the effectiveness of scaling reference motions depends not only on the amount of reference motions, but also on whether scaling expands the behavioral coverage and whether the newly introduced behaviors are relevant to the target evaluation benchmark. Intuitively, a BFM trained exclusively on walking motions is unlikely to perform a cartwheel without prior exposure to similar behaviors. Expanding the training set to include cartwheel-like motions can potentially improve performance on such behaviors. Conversely, if the current objective is to improve walking performance, introducing cartwheel motions into the training set is unlikely to provide significant additional benefit.

The findings above provide several practical insights for future research: 1) If the objective is to improve performance within a specific domain, simply increasing the amount of in-domain reference motions is likely to yield diminishing returns. Once the dominant behavioral patterns of the target domain are well represented, further gains are more likely to be achieved through a more carefully curated train-

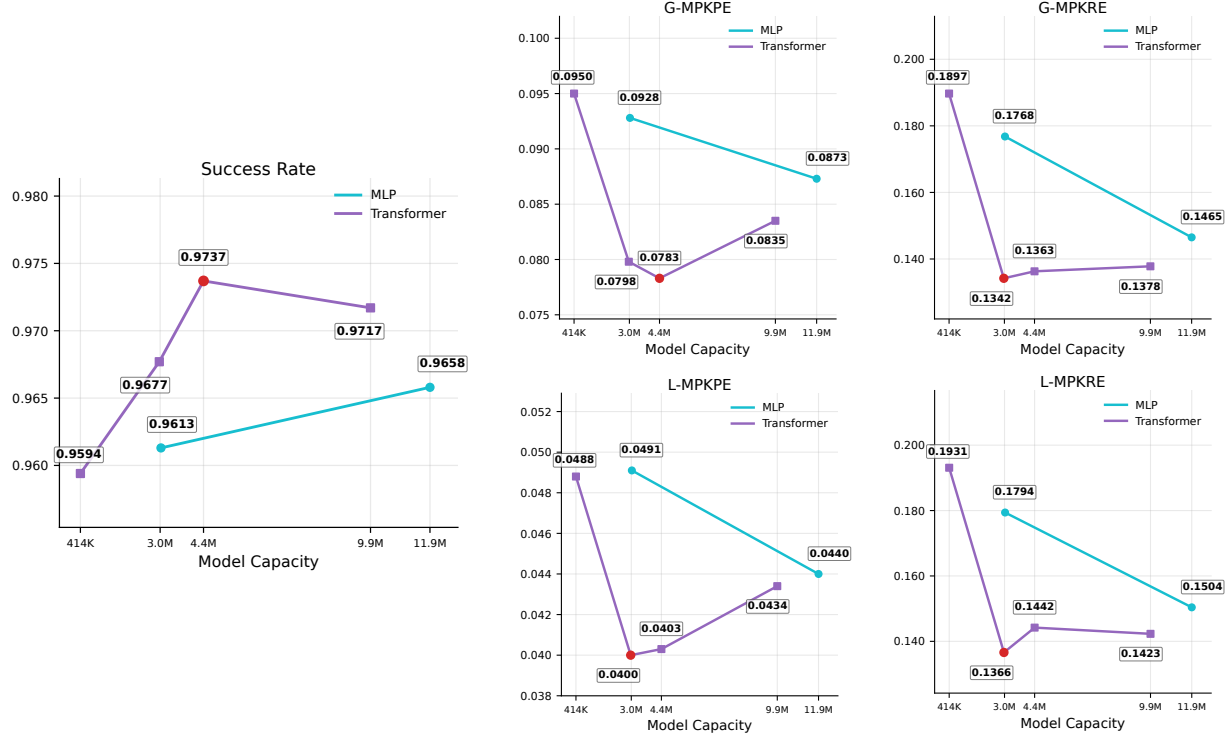


Figure 8: Results of scaling **model architectures** on the **BONES Test Set** under **whole-body control mode**. Best outcomes with the highest success rate and lowest tracking errors are highlighted with red circles.

ing set with fine-grained coverage of domain-specific behaviors. 2) Since no finite dataset can exhaust the space of possible behaviors, rare and long-tail behavioral patterns will inevitably remain absent. A more meaningful question is therefore not how to achieve a fine-grained behavioral coverage but how to curate a finite training set that achieves an effective balance between data efficiency and behavioral coverage. 3) If the objective is to improve the general capabilities of BFMs, incorporating diverse data sources appears to be a practical strategy. Although additional data sources do not inherently guarantee broader behavioral coverage, they often contribute complementary behavioral patterns—for example, LAFAN primarily provides locomotion behaviors, whereas OMOMO introduces human-object interaction behaviors. Combining such complementary sources can therefore provide a stronger behavioral foundation for BFM pretraining by exposing the model to a broader range of behavioral patterns.

4.3 Scaling with Model Architecture

We next investigate the effect of model architectures on scaling BFMs. Specifically, we intend to examine whether the proposed Humanoid Transformer achieves superior performance compared with previously adopted MLP architecture, and whether increasing its capacity leads to further performance gains.

All other training configurations are kept fixed across experiments, including the optimal on-policy data and reference motion configurations. We report results on both test sets under four representative control modes. The whole-body control results are presented in figures 8 and 9, while those for the remaining three modes, together with the detailed experimental setup are provided in Appendix B.6.

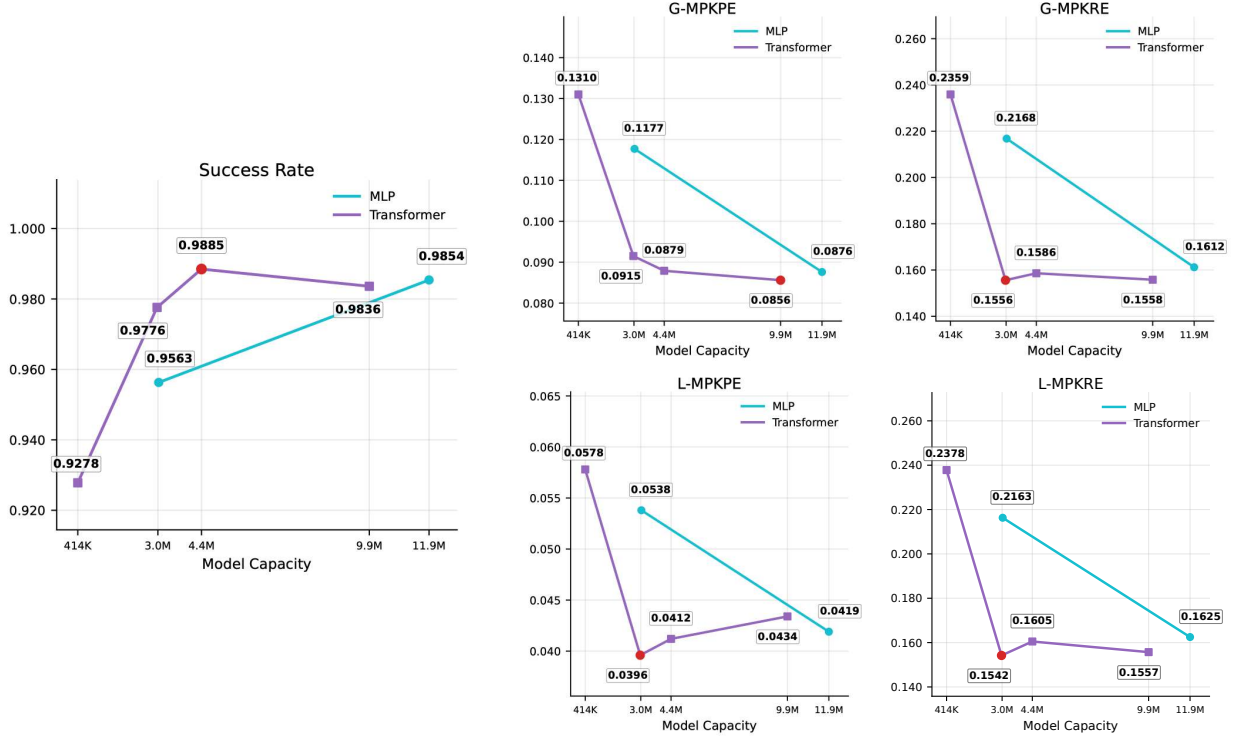


Figure 9: Results of scaling **model architectures** on the **Ours Test Set** under **whole-body control mode**. Best outcomes with the highest success rate and lowest tracking errors are highlighted with red circles.

As shown in the results, the Humanoid Transformer generally achieves higher success rates and lower tracking errors than the MLP across diverse experiments. Notably, the medium-sized Humanoid Transformer already achieves performance comparable to, or even better than, the substantially larger MLP while further increasing the Transformer size yields only diminishing returns. These results suggest that the proposed architecture serves as a stronger backbone for BFM pretraining, achieving superior performance through more effective utilization of model capacity under the same optimization budget.

However, we also observe that scaling Humanoid Transformer does not consistently improve performance across all the control modes. While some modes continue to benefit from increased model capacities, others exhibit early saturation at relatively modest scales. We hypothesize that this behavior arises from the imbalance in learning difficulty among different control modes. For example, the whole-body mode may be easier to learn than the root-and-end-effector mode, as the latter requires the model to infer natural whole-body behaviors from only sparse constraints. As training progresses, optimization trade-offs between these modes may become increasingly pronounced, allowing the model to further improve certain modes at the expense of slight performance degradation in others. To better understand this phenomenon, we provide a clearer analysis of the underlying learning dynamics in Section 4.4.

4.4 Latent Analysis

Beyond evaluating the performance of the proposed Humanoid Transformer, we further investigate whether structured latent representations can emerge without introducing auxiliary objectives. In this work, a latent space is considered structured if it satisfies: 1) Locality. Temporally close commands

Test Set	Noise Level	Success Rate $\uparrow$	G-MPKPE $\downarrow$	G-MPKRE $\downarrow$	L-MPKPE $\downarrow$	L-MPKRE $\downarrow$
BONES	0	0.9717	0.0835	0.1378	0.0434	0.1423
	5	0.9720	0.0846	0.1389	0.0437	0.1438
	10	0.9710	0.0883	0.1450	0.0447	0.1501
	15	0.9701	0.0956	0.1565	0.0468	0.1617
	20	0.9646	0.1067	0.1736	0.0505	0.1784
Ours	0	0.9836	0.0856	0.1558	0.0434	0.1557
	5	0.9830	0.0877	0.1567	0.0435	0.1565
	10	0.9776	0.0970	0.1640	0.0450	0.1631
	15	0.9715	0.1100	0.1744	0.0470	0.1724
	20	0.9400	0.1387	0.2033	0.0543	0.1991

Table 2: Evaluation of Latent Space Robustness. Latent representations on the hypersphere are perturbed by rotating each latent vector along a randomly sampled direction through an angle specified by the noise level.

should be mapped to nearby latent representations, preserving local neighborhood relationships and forming a locally continuous latent manifold. 2) Global Organization. Beyond preserving local neighborhood, the latent space should exhibit a coherent global structure, in which distinct behavioral intentions are organized into distinguishable regions rather than being arbitrarily intermixed. 3) Robustness. The latent representations should be resilient to perturbations, such that moderate noise on the latent representations does not substantially affect the model’s interpretation of the intended behavior.

As introduced in Section 3.4, the latent space of the Humanoid Transformer is designed as a continuous manifold on the hypersphere. Throughout this paper, we focus on the unit hypersphere, which is obtained by removing the multiplicative scaling factor in the RMSNorm module and therefore remains consistent across different layers. To examine the locality and global organization of this latent space, we extract the normalized latent representation corresponding to the first future target at each timestep of a motion sequence. The resulting latent representations are then projected into a three-dimensional space using a random Gaussian projection matrix and subsequently normalized to ensure unit length for visualization.

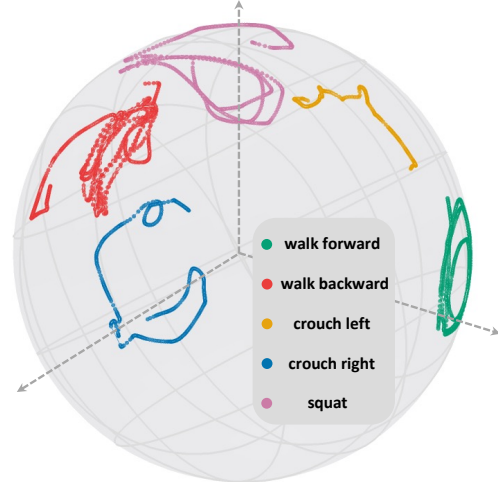


Figure 10: Visualization of the Latent Space.

According to the visualization results in figure 10, the latent space exhibits both locality and global organization across five selected motion sequences from the complete motion collection. From a local perspective, each motion sequence forms a continuous and smooth trajectory on the latent manifold, without abrupt transitions or high-frequency fluctuations. From a global perspective, trajectories corresponding to behavioral intentions in different directions occupy distinguishable regions of the latent space rather than being arbitrarily intermixed. Furthermore, the latent representations preserve the underlying directional relationships among behaviors: walking forward and backward are distributed on



Figure 11: Comparison of Latent-space Visualizations across Multiple Control Modes for Models of Different Sizes. Each row corresponds to the same motion sequence, and each column corresponds to a different model checkpoint. The left and right columns show the relatively small and relatively large models, respectively. Colors indicate different control modes, with the same color representing the same mode across all visualizations.

opposite sides of one latent dimension, while crouching left and right are separated along another. The visualization above provides qualitative evidence of the locality and global organization of the learned latent space, while the quantitative results in table 2 show that the model maintains strong performance even when relatively large perturbations are introduced into the latent representations. This observation provides empirical evidence for the robustness of the learned latent space, suggesting that the model can preserve its interpretation of the underlying behavioral intentions despite moderate latent noise.

In addition to the structural analysis, we also examine how diverse behavioral specifications are organized within this shared representation space and how this organization evolves as BFMs scale with stronger backbones. Specifically, we compare latent-space visualizations of multiple control modes over randomly sampled motion sequences between a relatively small model and a relatively large model. As shown in figure 11, increasing model capacity leads to convergence of the latent representations across different control modes. This observation may help explain the scaling behavior reported in Section 4.3, where improved performance on certain control modes is accompanied by slight degradation on others, potentially reflecting a strong coupling among control modes in the shared latent space.

Test Set	Method	Success Rate $\uparrow$	G-MPKPE $\downarrow$	G-MPKRE $\downarrow$	L-MPKPE $\downarrow$	L-MPKRE $\downarrow$
BONES	GMT	0.4407	0.5674	0.4789	0.0761	0.3177
	TWIST	0.4226	0.6212	0.5487	0.0704	0.3337
	SONIC [†]	0.9239	0.1740	0.1810	0.0436	0.1762
	BFM-Bym	0.9644	0.1005	0.1406	0.0406	0.1360
	BFM-Local	0.7286	0.2281	0.1279	0.0398	0.1312
	BFM-Global	0.9677	0.0798	0.1342	0.0400	0.1366
Ours	GMT	0.0855	1.8693	0.9829	0.0896	0.3158
	TWIST	0.0831	2.2744	1.0705	0.0596	0.3002
	SONIC	0.5937	0.5035	0.2098	0.0430	0.1974
	BFM-Bym	0.9709	0.1224	0.1648	0.0403	0.1592
	BFM-Local	0.2553	0.4731	0.1473	0.0388	0.1474
	BFM-Global	0.9776	0.0915	0.1556	0.0396	0.1542

Table 3: Performance Benchmark on Whole-body Control. We compare the 3M-parameter variant of the proposed BFM with off-the-shelf whole-body controllers and a self-curated ablation baseline (BFM-Bym), trained using the reward design of BeyondMimic and evaluated under the global control mode. The best result for each metric is highlighted in bold. [†]: The randomly sampled BONES Test Set may overlap with SONIC’s training set.

4.5 Model Performance Benchmarking

To benchmark the performance of the proposed BFM, we select whole-body control as a common evaluation task and compare our model with existing whole-body controllers. The baselines include several off-the-shelf models, namely GMT [20], TWIST [24, 25], and SONIC [11]. We further include an ablation baseline trained using the same pipeline but with the reward design of BeyondMimic [9], allowing us to evaluate the contribution of the proposed reward formulation. As shown in table 3, the proposed BFM consistently achieves higher success rates and lower tracking errors across both local and global control modes, outperforming all competing methods by a clear margin. Notably, the proposed reward formulation consistently yields lower global tracking errors than the ablation baseline, suggesting that tracking integrated whole-body trajectories in the global frame may provide more coherent behavioral guidance without compromising the coordination between root motion and whole-body pose evolution. Detailed descriptions of the local and global control modes are provided in Appendix B.1.

5 Discussion

In this paper, we present a systematic investigation of the scaling behavior of Behavior Foundation Models for humanoid robots by revisiting three fundamental aspects: the learning paradigm, behavioral data, and model architecture. Through extensive empirical studies, we derive an effective scaling recipe that yields substantial improvements in control fidelity, execution robustness, and generalization capability. Nevertheless, the proposed framework has its limitations. Although we curate a unified control interface with diverse control modes, it remains unclear whether this interface is the most appropriate abstraction and how these modes should be integrated with future high-level policies. Consequently, the design of BFMs should continue to evolve alongside advances in high-level policy learning. Moreover, the current infrastructure for BFM pretraining remains relatively preliminary and constrained, limiting its scalability to broader settings. We hope that this work provides a foundation for systematic development of BFMs, ultimately contributing to the realization of general-purpose humanoid intelligence.

References

- [1] Tao Huang, Junli Ren, Huayi Wang, Zirui Wang, Qingwei Ben, Muning Wen, Xiao Chen, Jianan Li, and Jiangmiao Pang. Learning humanoid standing-up control across diverse postures. *arXiv preprint arXiv:2502.08378*, 2025.
- [2] Xialin He, Runpei Dong, Zixuan Chen, and Saurabh Gupta. Learning getting-up policies for real-world humanoid robots. *arXiv preprint arXiv:2502.12152*, 2025.
- [3] Yufei Xue, Wentao Dong, Minghuan Liu, Weinan Zhang, and Jiangmiao Pang. A unified and general humanoid whole-body controller for versatile locomotion. *arXiv preprint arXiv:2502.03206*, 2025.
- [4] Qingwei Ben, Feiyu Jia, Jia Zeng, Juntong Dong, Dahua Lin, and Jiangmiao Pang. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. *arXiv preprint arXiv:2502.13013*, 2025.
- [5] Tairan He, Jiawei Gao, Wenli Xiao, Yuanhang Zhang, Zi Wang, Jiashun Wang, Zhengyi Luo, Guanqi He, Nikhil Sobanbab, Chaoyi Pan, et al. Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. *arXiv preprint arXiv:2502.01143*, 2025.
- [6] Weishuai Zeng, Shunlin Lu, Kangning Yin, Xiaojie Niu, Minyue Dai, Jingbo Wang, and Jiangmiao Pang. Behavior foundation model for humanoid robots. *arXiv preprint arXiv:2509.13780*, 2025.
- [7] Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions On Graphics (TOG)*, 43(6):1–21, 2024.
- [8] Tairan He, Wenli Xiao, Toru Lin, Zhengyi Luo, Zhenjia Xu, Zhenyu Jiang, Jan Kautz, Changliu Liu, Guanya Shi, Xiaolong Wang, et al. Hover: Versatile neural whole-body controller for humanoid robots. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9989–9996. IEEE, 2025.
- [9] Qiayuan Liao, Takara E Truong, Xiaoyu Huang, Yuman Gao, Guy Tevet, Koushil Sreenath, and C Karen Liu. Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. *arXiv preprint arXiv:2508.08241*, 2025.
- [10] Yitang Li, Zhengyi Luo, Tonghe Zhang, Cunxi Dai, Anssi Kanervisto, Andrea Tirinzoni, Haoyang Weng, Kris Kitani, Mateusz Guzek, Ahmed Touati, et al. Bfm-zero: A promptable behavioral foundation model for humanoid control using unsupervised reinforcement learning. *arXiv preprint arXiv:2511.04131*, 2025.
- [11] Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Sirui Chen, Fernando Castaneda, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *arXiv preprint arXiv:2511.07820*, 2025.
- [12] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

- [13] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.
- [14] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- [15] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018.
- [16] Maiyue Chen, Kaihui Wang, Bo Zhang, Xihan Ma, Zhiyuan Yang, Yi Ren, Qijun Huang, Zihao Zhu, Yucheng Wang, and Zhizhong Su. Holomotion-1 technical report. *arXiv preprint arXiv:2605.15336*, 2026.
- [17] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [18] Younggyo Seo, Carmelo Sferrazza, Haoran Geng, Michal Nauman, Zhao-Heng Yin, and Pieter Abbeel. Fasttd3: Simple, fast, and capable reinforcement learning for humanoid control. *arXiv preprint arXiv:2505.22642*, 2025.
- [19] Brent Yi, Hongsuk Choi, Himanshu Gaurav Singh, Xiaoyu Huang, Takara E Truong, Carmelo Sferrazza, Yi Ma, Rocky Duan, Pieter Abbeel, Guanya Shi, et al. Flow policy gradients for robot control. *arXiv preprint arXiv:2602.02481*, 2026.
- [20] Zixuan Chen, Mazeyu Ji, Xuxin Cheng, Xuanbin Peng, Xue Bin Peng, and Xiaolong Wang. Gmt: General motion tracking for humanoid whole-body control. *arXiv preprint arXiv:2506.14770*, 2025.
- [21] Kangning Yin, Weishuai Zeng, Ke Fan, Minyue Dai, Zirui Wang, Qiang Zhang, Zheng Tian, Jingbo Wang, Jiangmiao Pang, and Weinan Zhang. Unitracker: Learning universal whole-body motion tracker for humanoid robots. *IEEE Robotics and Automation Letters*, 2026.
- [22] Yunshen Wang, Shaohang Zhu, Peiyuan Zhi, Yuhan Li, Jiaxin Li, Yong-Lu Li, Yuchen Xiao, Xingxing Wang, Baoxiong Jia, and Siyuan Huang. Omnixtreme: Breaking the generality barrier in high-dynamic humanoid control. *arXiv preprint arXiv:2602.23843*, 2026.
- [23] Yixuan Li, Yutang Lin, Jieming Cui, Tengyu Liu, Wei Liang, Yixin Zhu, and Siyuan Huang. Clone: Closed-loop whole-body humanoid teleoperation for long-horizon tasks. In *9th Annual Conference on Robot Learning*, 2025.
- [24] Yanjie Ze, Zixuan Chen, Joao Pedro Araújo, Zi-ang Cao, Xue Bin Peng, Jiajun Wu, and C Karen Liu. Twist: Teleoperated whole-body imitation system. *arXiv preprint arXiv:2505.02833*, 2025.
- [25] Yanjie Ze, Siheng Zhao, Weizhuo Wang, Angjoo Kanazawa, Rocky Duan, Pieter Abbeel, Guanya Shi, Jiajun Wu, and C Karen Liu. Twist2: Scalable, portable, and holistic humanoid data collection system. *arXiv preprint arXiv:2511.02832*, 2025.

- [26] Huayi Wang, Wentao Zhang, Runyi Yu, Tao Huang, Junli Ren, Feiyu Jia, Zirui Wang, Xiaojie Niu, Xiao Chen, Jiahe Chen, et al. Physshi: Towards a real-world generalizable and natural humanoid-scene interaction system. *arXiv preprint arXiv:2510.11072*, 2025.
- [27] Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C Karen Liu, Rocky Duan, and Guanya Shi. Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. *arXiv preprint arXiv:2509.26633*, 2025.
- [28] Ziwen Zhuang, Shaoting Zhu, Mengjie Zhao, and Hang Zhao. Deep whole-body parkour. *arXiv preprint arXiv:2601.07701*, 2026.
- [29] Haoyang Weng, Yitang Li, Nikhil Sobanbabu, Zihan Wang, Zhengyi Luo, Tairan He, Deva Ramanan, and Guanya Shi. Hdmi: Learning interactive humanoid whole-body control from human videos. *arXiv preprint arXiv:2509.16757*, 2025.
- [30] Yinhuai Wang, Qihan Zhao, Yuen Fui Lau, Runyi Yu, Hok Wai Tsui, Qifeng Chen, Jingbo Wang, Jiangmiao Pang, and Ping Tan. Humanx: Toward agile and generalizable humanoid interaction skills from human videos. *arXiv preprint arXiv:2602.02473*, 2026.
- [31] Xialin He, Sirui Xu, Xinyao Li, Runpei Dong, Liuyu Bian, Yu-Xiong Wang, and Liang-Yan Gui. Ultra: Unified multimodal control for autonomous humanoid whole-body loco-manipulation. *arXiv preprint arXiv:2603.03279*, 2026.
- [32] Nan Jiang, Zimo He, Wanhe Yu, Lexi Pang, Yunhao Li, Hongjie Li, Jieming Cui, Yuhan Li, Yizhou Wang, Yixin Zhu, et al. Uniact: Unified motion generation and action streaming for humanoid robots. *arXiv preprint arXiv:2512.24321*, 2025.
- [33] Weiwei Xie, Jiakun Zheng, Jinrui Han, Jiyuan Shi, Weinan Zhang, Chenjia Bai, and Xuelong Li. Textop: Real-time interactive text-driven humanoid robot motion generation and control. *arXiv preprint arXiv:2602.07439*, 2026.
- [34] Zelin Tao, Zeran Su, Peiran Liu, Jingkai Sun, Wenqiang Que, Jiahao Ma, Jialin Yu, Jiahang Cao, Pihai Sun, Hao Liang, et al. Heracles: Bridging precise tracking and generative synthesis for general humanoid control. *arXiv preprint arXiv:2603.27756*, 2026.
- [35] Kangning Yin, Zhe Cao, Wentao Dong, Weishuai Zeng, Tianyi Zhang, Qiang Zhang, Jingbo Wang, Jiangmiao Pang, Ming Zhou, and Weinan Zhang. Robostriker: Hierarchical decision-making for autonomous humanoid boxing. *arXiv preprint arXiv:2601.22517*, 2026.
- [36] Zhikai Zhang, Haofei Lu, Yunrui Lian, Ziqing Chen, Yun Liu, Chenghuai Lin, Han Xue, Zicheng Zeng, Zekun Qi, Shaolin Zheng, et al. Learning athletic humanoid tennis skills from imperfect human motion data. *arXiv preprint arXiv:2603.12686*, 2026.
- [37] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

- [38] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- [39] Xiaoyu Huang, Takara Truong, Yunbo Zhang, Fangzhou Yu, Jean Pierre Sleiman, Jessica Hodgins, Koushil Sreenath, and Farbod Farshidian. Diffuse-cloc: Guided diffusion for physics-based character look-ahead control. *ACM Transactions on Graphics (TOG)*, 44(4):1–12, 2025.
- [40] Ahmed Touati and Yann Ollivier. Learning one representation to optimize all rewards. *Advances in Neural Information Processing Systems*, 34:13–23, 2021.
- [41] Andrea Tirinzoni, Ahmed Touati, Jesse Farebrother, Mateusz Guzek, Anssi Kanervisto, Yingchen Xu, Alessandro Lazaric, and Matteo Pirota. Zero-shot whole-body humanoid control via behavioral foundation models. *arXiv preprint arXiv:2504.11054*, 2025.
- [42] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020.
- [43] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019.
- [44] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023.
- [45] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020.
- [46] Chuan Guo, Inwoo Hwang, Jian Wang, and Bing Zhou. Snapmogen: Human motion generation from expressive texts. *arXiv preprint arXiv:2507.09122*, 2025.
- [47] Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li. Finedance: A fine-grained choreography dataset for 3d full body dance generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10234–10243, 2023.
- [48] Bones Studio. Bones-seed: Skeletal everyday embodiment dataset. <https://huggingface.co/datasets/bones-studio/seed>, 2026.
- [49] Claire McLean, Makenzie Meendering, Tristan Swartz, Orri Gabbay, Alexandra Olsen, Rachel Jacobs, Nicholas Rosen, Philippe de Bree, Tony Garcia, Gadsden Merrill, et al. Embody 3d: A large-scale multimodal motion and behavior dataset. *arXiv preprint arXiv:2510.16258*, 2025.
- [50] Joao Pedro Araujo, Yanjie Ze, Pei Xu, Jiajun Wu, and C Karen Liu. Retargeting matters: General motion retargeting for humanoid motion tracking. *arXiv preprint arXiv:2510.02252*, 2025.
- [51] Chen Tessler*, Yifeng Jiang*, Xue Bin Peng, Erwin Coumans, Yi Shi, Haotian Zhang, Davis Rempe, Gal Chechik†, and Sanja Fidler†. Protomotions3: An open-source framework for humanoid simulation and control. <https://github.com/NVlabs/ProtoMotions/>, 2025.

- [52] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.
- [53] Kaiyu Yue, Menglin Jia, Ji Hou, and Tom Goldstein. Image generation with a sphere encoder. *arXiv preprint arXiv:2602.15030*, 2026.
- [54] Unitree. Unitree g1 humanoid agent ai avatar, 2024.
- [55] Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Munoz, Xinjie Yao, René Zurbügg, Nikita Rudin, et al. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025.
- [56] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.
- [57] Martin Schepers, Matteo Giuberti, Giovanni Bellusci, et al. Xsens mvn: Consistent tracking of human motion using inertial sensing. *Xsens Technol*, 1(8):1–8, 2018.
- [58] Ian Mason, Sebastian Starke, and Taku Komura. Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 5(1), may 2022.
- [59] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.
- [60] Albert S. Huang, Edwin Olson, and David C. Moore. Lcm: Lightweight communications and marshallng. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4057–4062, 2010.
- [61] HTC VIVE. Vive ultimate tracker – full-body tracking for standalone vr, 2024. Accessed: 2026-01-18.

A Method Details

A.1 Control Modes

Control Mode	# Links	Activated Links
Root Mode	1	Pelvis
Bimanual Mode	2	Left Hand, Right Hand
Root-and-Hand Mode	3	Pelvis, Left Hand, Right Hand
End-Effector Mode	4	Left Hand, Right Hand, Left Foot, Right Foot
Root-and-End-Effector Mode	5	Pelvis, Left Hand, Right Hand, Left Foot, Right Foot
Upper-Body Mode	6	Left Shoulder, Right Shoulder, Left Elbow Right Elbow, Left Hand, Right Hand
Root-and-Upper-Body Mode	7	Pelvis, Left Shoulder, Right Shoulder Left Elbow, Right Elbow, Left Hand, Right Hand
Whole-Body Mode	14	Pelvis, Left Shoulder, Right Shoulder, Left Elbow Right Elbow, Left Hand, Right Hand, Torso, Left Hip Right Hip, Left Knee, Right Knee, Left Foot, Right Foot

Table 4: Curated control modes featuring different activated target links. Instead of constructing arbitrary link-wise masks by independently activating each link according to a Bernoulli distribution with probability 0.5 [6], we curate a set of eight representative control modes spanning different levels of specification granularity. By randomly sampling a control mode at each environment reset, the same underlying behavior can be represented by diverse forms of behavioral specifications. Such design enables versatile humanoid control across multiple control modes while naturally supporting behavior inpainting under sparse control signals.

A.2 Reward Design

Term	Expression	Weight	Notes
Tracking Reward			
Global Root Height	$\exp(-\ h_t - h_t^{\text{ref}}\ _2^2 / \sigma_{\text{height}}^2)$	0.5	$\sigma_{\text{height}} = 0.3$
Global Body Position	$\exp(-\frac{1}{ B } \sum_{b \in B} \ p_{t,b} - p_{t,b}^{\text{ref}}\ _2^2 / \sigma_{\text{pos}}^2)$	1.0	$\sigma_{\text{pos}} = 0.3$
Global Body Rotation	$\exp(-\frac{1}{ B } \sum_{b \in B} \ \theta_{t,b} - \theta_{t,b}^{\text{ref}}\ _2^2 / \sigma_{\text{rot}}^2)$	1.0	$\sigma_{\text{rot}} = 0.4$
Global Body Linear Velocity	$\exp(-\frac{1}{ B } \sum_{b \in B} \ v_{t,b} - v_{t,b}^{\text{ref}}\ _2^2 / \sigma_{\text{lin}}^2)$	1.0	$\sigma_{\text{lin}} = 1.0$
Global Body Angular Velocity	$\exp(-\frac{1}{ B } \sum_{b \in B} \ \omega_{t,b} - \omega_{t,b}^{\text{ref}}\ _2^2 / \sigma_{\text{ang}}^2)$	1.0	$\sigma_{\text{ang}} = 3.14$
Penalty Term			
Action Rate	$\ \mathbf{a}_t - \mathbf{a}_{t-1}\ _2^2$	-0.1	
Joint Position Limits	$\sum_{j \in J} \mathbf{1}[q_{t,j} \notin [q_j^{\min}, q_j^{\max}]]$	-10.0	
Survival Reward			
Survival	$\mathbf{1}[\text{is_alive}]$	1.0	

Table 5: Reward components and weights. Variables annotated with the superscript ref are derived from the reference motion, whereas those without this annotation correspond to the humanoid’s current state. Reward weights are automatically scaled by the environment according to the simulation timestep dt .

A.3 Domain Randomization

Domain Randomization	Sampling Distribution
<i>Physical parameters</i>	
Static friction coefficients	$\mu_s \sim U[0.3, 1.6]$
Dynamic friction coefficients	$\mu_d \sim U[0.3, 1.2]$
Default joint positions	$q_0 \sim q_0 + U[-0.01, 0.01]$
Base COM offset (x, y, z)	$\Delta x \sim U[-0.025, 0.025], \Delta y \sim U[-0.05, 0.05], \Delta z \sim U[-0.05, 0.05]$
Hand mass	$m_0 \sim m_0 + U[0.0, 1.0]$
<i>Root velocity perturbations</i>	
Root linear vel (x, y, z)	$v_x \sim U[-0.5, 0.5], v_y \sim U[-0.5, 0.5], v_z \sim U[-0.2, 0.2]$
Root angular vel	$\omega_{\text{roll}} \sim U[-0.52, 0.52], \omega_{\text{pitch}} \sim U[-0.52, 0.52], \omega_{\text{yaw}} \sim U[-0.78, 0.78]$
Push duration	$\Delta t \sim U[1, 3]\text{s}$

Table 6: Domain randomization parameters. Randomization of physical parameters is applied at the start of training. Root velocity perturbations are introduced periodically during training. $U[\cdot]$: uniform distribution.

A.4 Model Architecture

We provide the implementation details of the proposed *Humanoid Transformer* for BFM pretraining.

Let L denotes the history length. Proprioceptions and actions are first extended to temporal windows:

$$H_t \triangleq (\mathbf{s}_{t-L+1}^p, \mathbf{a}_{t-L+1}, \dots, \mathbf{s}_{t-1}^p, \mathbf{a}_{t-1}, \mathbf{s}_t^p). \quad (\text{A.1})$$

Given the historical information, we encode proprioceptive states and actions with separate tokenizers:

$$z_i^{s^p} = E_{s^p}(\mathbf{s}_i^p) \quad z_i^a = E_a(\mathbf{a}_i). \quad (\text{A.2})$$

The historical context is then represented as a sequence of interleaved proprioceptive and action tokens:

$$(z_{t-L+1}^{s^p}, z_{t-L+1}^a, \dots, z_{t-1}^{s^p}, z_{t-1}^a, z_t^{s^p}). \quad (\text{A.3})$$

We additionally append a learnable query token e at the end and obtain the complete context:

$$Z_t^p \triangleq (z_{t-L+1}^{s^p}, z_{t-L+1}^a, \dots, z_{t-1}^{s^p}, z_{t-1}^a, z_t^{s^p}, e). \quad (\text{A.4})$$

Let N denote the number of future frames. The goal states are also extended to a temporal window:

$$G_t \triangleq (\mathbf{s}_{t+\delta_1}^g, \mathbf{s}_{t+\delta_2}^g, \dots, \mathbf{s}_{t+\delta_N}^g), \quad (\text{A.5})$$

where δ_i is the temporal offset of future frame i relative to current frame and $\mathbf{s}_{t+\delta_i}^g$ is defined as:

$$\begin{aligned} \mathbf{x}_{t+\delta_i} &\triangleq (p_{t+\delta_i}^{\text{ref}} - p_t^{\text{root}}, p_{t+\delta_i}^{\text{ref}} - p_t, \theta_{t+\delta_i}^{\text{ref}} \ominus \theta_t^{\text{root}}, \theta_{t+\delta_i}^{\text{ref}} \ominus \theta_t) \\ \mathbf{s}_{t+\delta_i}^g &\triangleq (\mathbf{x}_{t+\delta_i} \odot \mathbf{m}_i, \mathbf{m}_i), \mathbf{m}_i \sim M \end{aligned} \quad (\text{A.6})$$

We concatenate goal states with the temporal offsets and encode them with a shared tokenizer:

$$z_i^{s^g} = E_{s^g}(\mathbf{s}_{t+\delta_i}^g). \quad (\text{A.7})$$

The future window then becomes a sequence of task tokens, each corresponding to a future frame:

$$Z_t^g \triangleq (z_1^{s^g}, z_2^{s^g}, \dots, z_N^{s^g}) \quad (\text{A.8})$$

Given the historical context Z_t^p and future window Z_t^g , we employ self-attention to capture the temporal dependencies and dynamic evolution in historical context and cross-attention to condition the model on diverse behavioral specifications. This design decouples the modeling of the behavioral manifold from the injection of control conditions, naturally aligning with our objective of building BFM.

Each layer of the transformer consists of the self-attention, the cross-attention, the feed-forward and the RMSNorm module. Let $Z_{t,(i-1)}^p$ denote the context after layer $i - 1$. The updating rule in layer i is:

$$\begin{aligned} Z_{t,(i-1)}^p &= Z_{t,(i-1)}^p + \text{SelfAttn}(\text{RoPE}(\text{RMSNorm}(Z_{t,(i-1)}^p))) \\ Z_{t,(i-1)}^p &= Z_{t,(i-1)}^p + \text{CrossAttn}(\text{RMSNorm}(Z_{t,(i-1)}^p), \text{RMSNorm}(Z_t^g)) \\ Z_{t,(i)}^p &= Z_{t,(i-1)}^p + \text{FFN}(\text{RMSNorm}(Z_{t,(i-1)}^p)) \end{aligned} \quad (\text{A.9})$$

where the learnable query token is masked from being attended to by the other context tokens while it can attend to all context tokens to aggregate historical information in the self-attention module.

The RMSNorm module projects goal embeddings onto a continuous and bounded hypersphere via:

$$\text{RMSNorm}(\mathbf{x}) = \gamma \cdot \frac{\mathbf{x}}{\sqrt{\frac{1}{d} \sum_{i=1}^d x_i^2}}, \quad (\text{A.10})$$

where γ is a learnable scalable parameter. Throughout this paper, we mainly focus on the unit hypersphere, regardless of any multiplicative scaling coefficient. We do not employ auxiliary objectives to regularize the latent space; instead, it is shaped solely through the motion-tracking objective.

Given the output of the backbone, we denote the latent embedding corresponding to the learnable query token as $\hat{e}$ and use a projection head f to predict the action or value, which is expressed as $f(\hat{e})$.

In practice, we implement the proprioception tokenizer E_{sp} , action tokenizer E_a , goal tokenizer E_{sg} , projection head f all as single linear layers, and the feed-forward module as SwiGLU [59].

B Experiment Details

B.1 Real-World Deployment

All real-world experiments are conducted on the Unitree G1 humanoid robot with 29 degrees of freedom. We decouple high-level BFM inference from the low-level joint-space PD controller and employ LCM [60] for inter-process communication. This architecture provides direct access to the low-level control loop, thereby facilitating the implementation of safety mechanisms, such as binding emergency termination to a dedicated button on the remote controller for direct human interference.

The low-level controller is implemented upon Unitree SDK2 and operates at 200 Hz, whereas high-level BFM inference is accelerated with TensorRT and runs at 50 Hz. Both processes can be executed either

onboard with the assistance of the Jetson Orin GPU, or on a local PC connected to the robot via a network cable. For scenarios involving online streaming of control signals, we transmit data over Wi-Fi using TCP and buffer the received signals with their timestamps. During inference, the index of the final future reference frame is dynamically adjusted according to the recorded timestamps, providing an effective mechanism for compensating communication and system latency during real-time deployment.

Although BFM s are pretrained with objectives to reproduce reference motions as integrated whole-body trajectories in the global frame, they naturally support both global and local control during real-world deployment. In the global control mode, we adopt HTC VIVE Ultimate Trackers [61] for root localization. The root position on the xy -plane and the heading direction of the control signals are aligned with those of the robot through first-frame calibration. In the local control mode, the current root position specified by the control signal is treated as the robot’s actual root position. While the reference heading could likewise be regarded as the robot’s current heading, we instead preserve the heading derived from the IMU and align the direction of control signals through the same first-frame calibration procedure. This design avoids modifying the robot’s raw sensory inputs while still enabling accurate tracking of the desired root orientation. The overall procedure can be interpreted as a reference-forcing mechanism: it assumes that the robot has already reached the desired root position and therefore only needs to track the remaining root orientation and whole-body pose, a strategy commonly adopted in prior work.

Since BFM s support eight curated control modes, any of them can be activated and switched online during real-world deployment. Notably, some control modes do not provide an explicit specification of the root target state; consequently, local control is not directly applicable under these modes.

B.2 Evaluation Protocol

We report the success rate together with global and local tracking errors across diverse control modes. A motion sequence is considered unsuccessful if any selected link deviates from its desired position by more than 0.5m in the global frame. G-MPKPE and G-MPKRE denote the mean per-keypoint position and rotation errors of the selected links in the global frame, whereas L-MPKPE and L-MPKRE measure the corresponding root-relative errors after removing horizontal translation offsets and heading differences. Position errors (G-MPKPE and L-MPKPE) are reported in meters, whereas rotation errors (G-MPKRE and L-MPKRE) are reported in radians. The metrics are defined below, where T is the number of frames and B the selected link set.

$$\text{G-MPKPE} = \frac{1}{T \cdot |B|} \sum_{t=1}^T \sum_{b \in B} \|p_{t,b} - p_{t,b}^{\text{ref}}\|_2 \quad (\text{B.1})$$

$$\text{G-MPKRE} = \frac{1}{T \cdot |B|} \sum_{t=1}^T \sum_{b \in B} \|\theta_{t,b} \ominus \theta_{t,b}^{\text{ref}}\|_2 \quad (\text{B.2})$$

$$\text{L-MPKPE} = \frac{1}{T \cdot |B|} \sum_{t=1}^T \sum_{b \in B} \|p_{t,b}^{\text{root-relative}} - p_{t,b}^{\text{root-relative,ref}}\|_2 \quad (\text{B.3})$$

$$\text{L-MPKRE} = \frac{1}{T \cdot |B|} \sum_{t=1}^T \sum_{b \in B} \|\theta_{t,b}^{\text{root-relative}} \ominus \theta_{t,b}^{\text{root-relative,ref}}\|_2 \quad (\text{B.4})$$

B.3 Training Hyperparameters

Term	Value
Optimizer	Adam
Value coefficient	1.0
PPO Clipping	0.2
Entropy coefficient	5.0e-3
Learning epochs	2
Actor learning rate	2.0e-5
Critic learning rate	1.0e-3
Discount factor γ	0.99
GAE λ	0.95
Desired KL value	0.01
Maximum gradient norm	1.0
Training environments per GPU	8192
Initial noise std	0.8
GPU Type	NVIDIA RTX 4090

Table 7: Hyperparameter Details. These terms remain fixed across experiments to ensure a fair comparison.

B.4 Scaling On-Policy Data Collection

We scale on-policy data collection along two complementary dimensions: the *width* and the *depth* of environmental interaction. With the number of environments per GPU held constant, the *width* is increased by scaling the number of GPUs, while the *depth* is increased by extending the rollout horizon.

To ensure a fair comparison, we keep all the training hyperparameters fixed across experiments, as detailed in table 7. In particular, the number of mini-batches is adjusted according to the rollout horizon so that only the total amount of collected on-policy data varies across experiments, while the mini-batch size remains constant. We further adopt the reference motion dataset of medium size, which consists solely of the BONES-SEED dataset excluding the held-out test split. This removes potential benefits arising from heterogeneous data sources in the full reference motion dataset, allowing us to isolate the impact of scaling on-policy data collection. Finally, all experiments employ the same Humanoid Transformer model architecture of medium size, which provides a favorable balance between control performance and real-time onboard inference efficiency. Collectively, these design choices ensure that the observed differences can be primarily attributed to the scale of on-policy data collection.

To evaluate the effect of scaling on-policy data collection, we report results of two test sets under four representative control modes: root-and-hand mode (figure 12 and 13), root-and-end-effector mode (figure 14 and 15), root-and-upper-body mode (figure 16 and 17)), and whole-body mode (figure 4 and 5). Both the rollout horizon and the number of GPUs are varied across three levels (32, 48, and 64). For all configurations, we evaluate the checkpoint obtained after 20,000 training epochs.

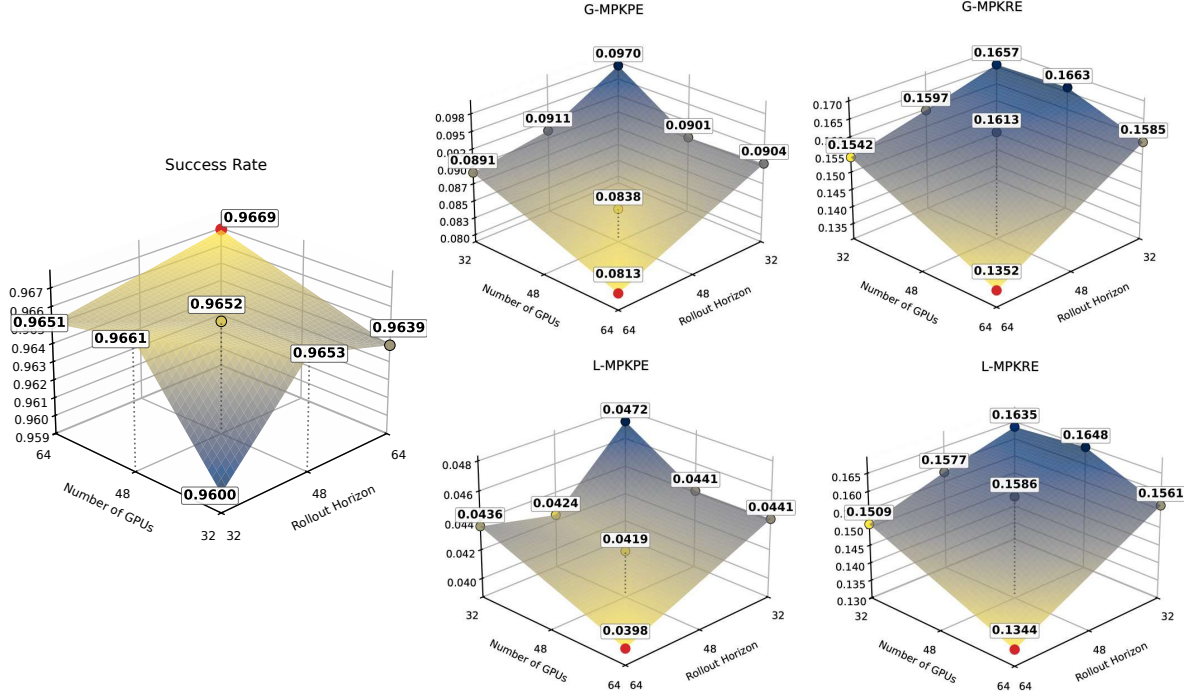


Figure 12: Results of scaling on-policy data collection on the BONES Test Set under root-and-hand control mode. Best results with the highest success rate and lowest tracking errors are highlighted with red circles.

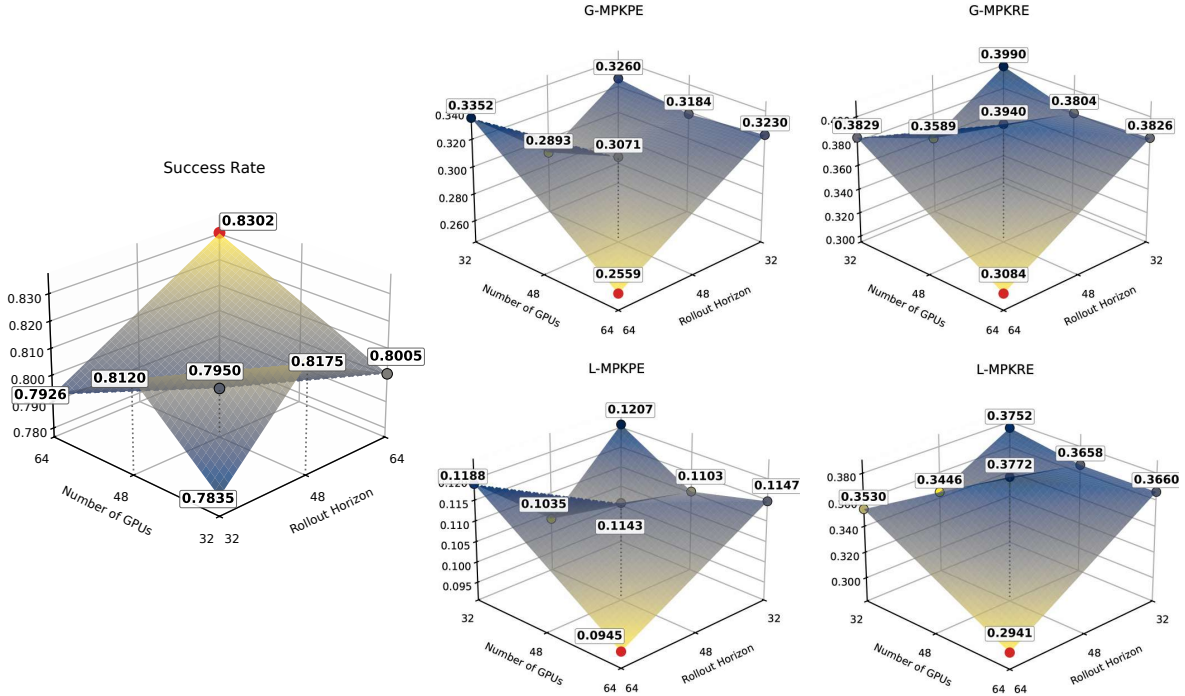


Figure 13: Results of scaling on-policy data collection on the Ours Test Set under root-and-hand control mode. Best results with the highest success rate and lowest tracking errors are highlighted with red circles.

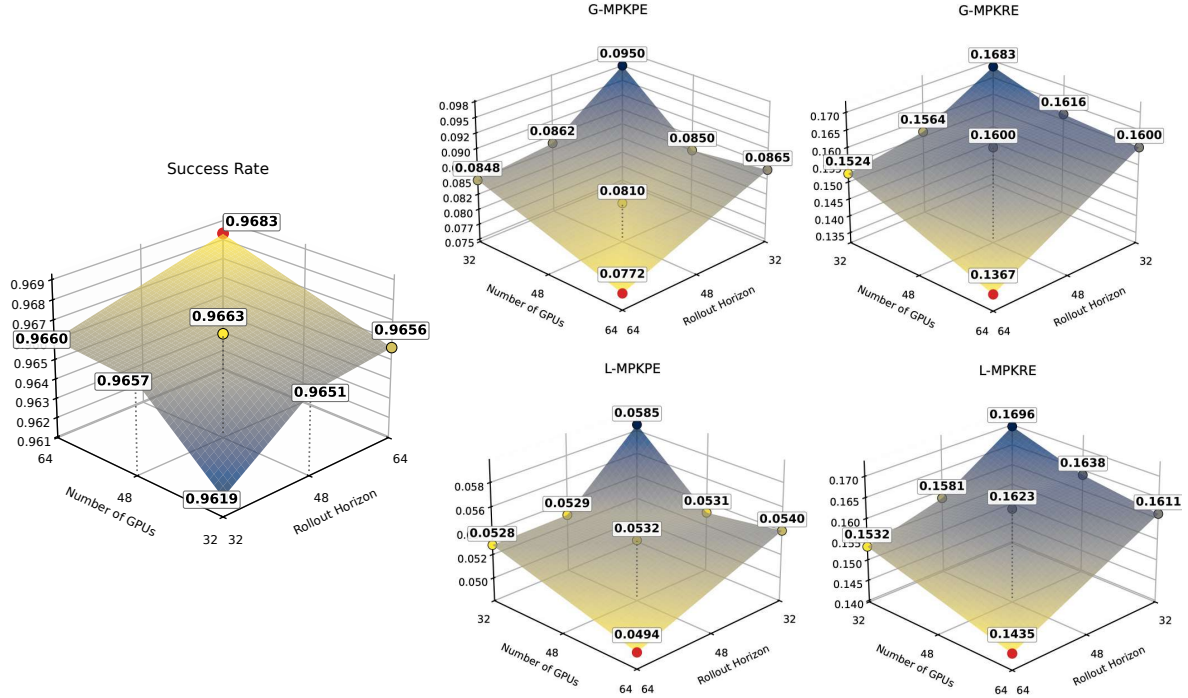


Figure 14: Results of scaling on-policy data collection on the BONES Test Set under root-and-end-effector mode. Best results with the highest success rate and lowest tracking errors are highlighted with red circles.

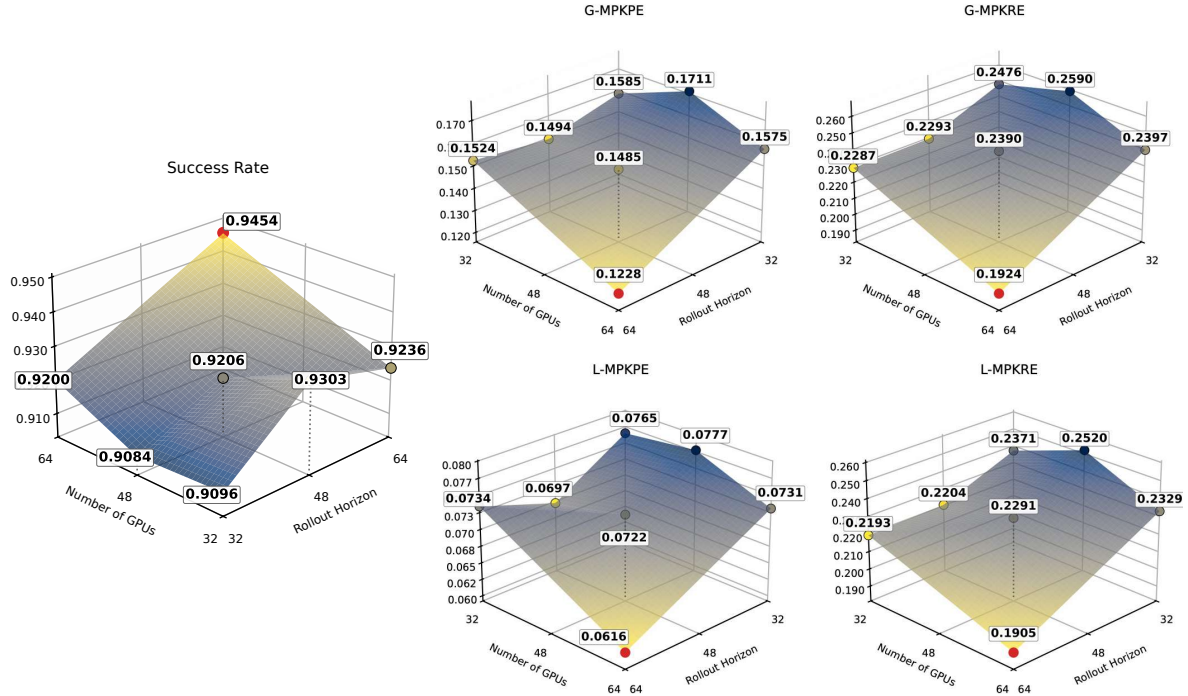


Figure 15: Results of scaling on-policy data collection on the Ours Test Set under root-and-end-effector mode. Best results with the highest success rate and lowest tracking errors are highlighted with red circles.

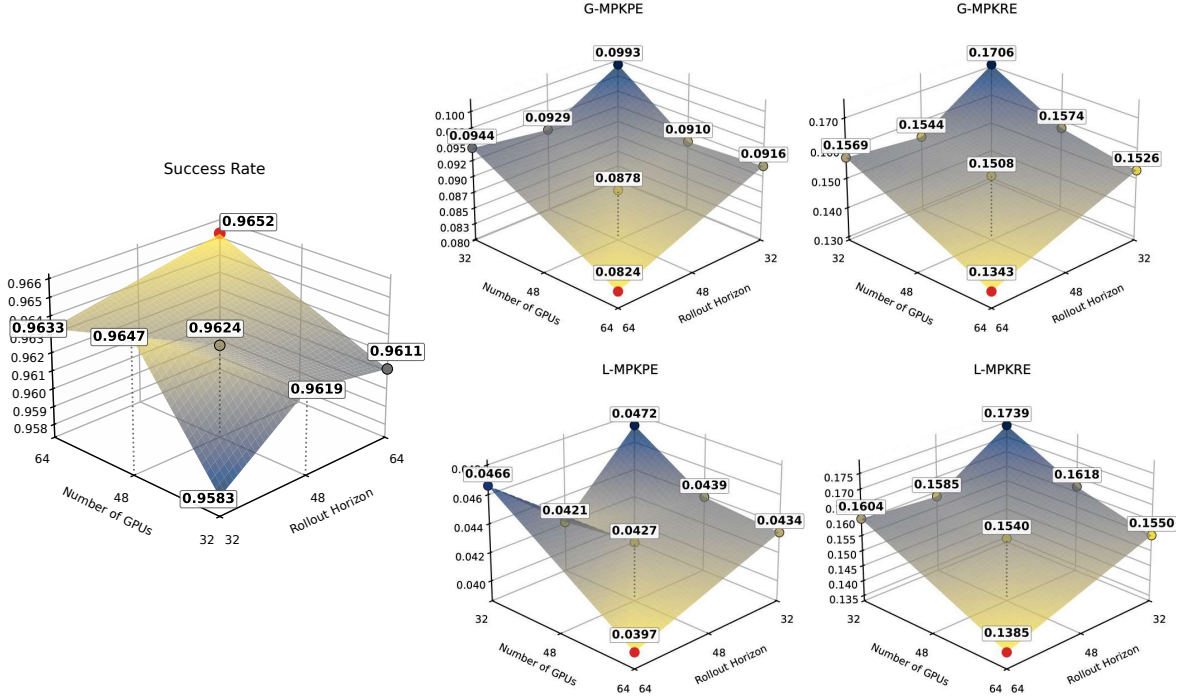


Figure 16: Results of scaling on-policy data collection on the BONES Test Set under root-and-upper-body mode. Best results with the highest success rate and lowest tracking errors are highlighted with red circles.

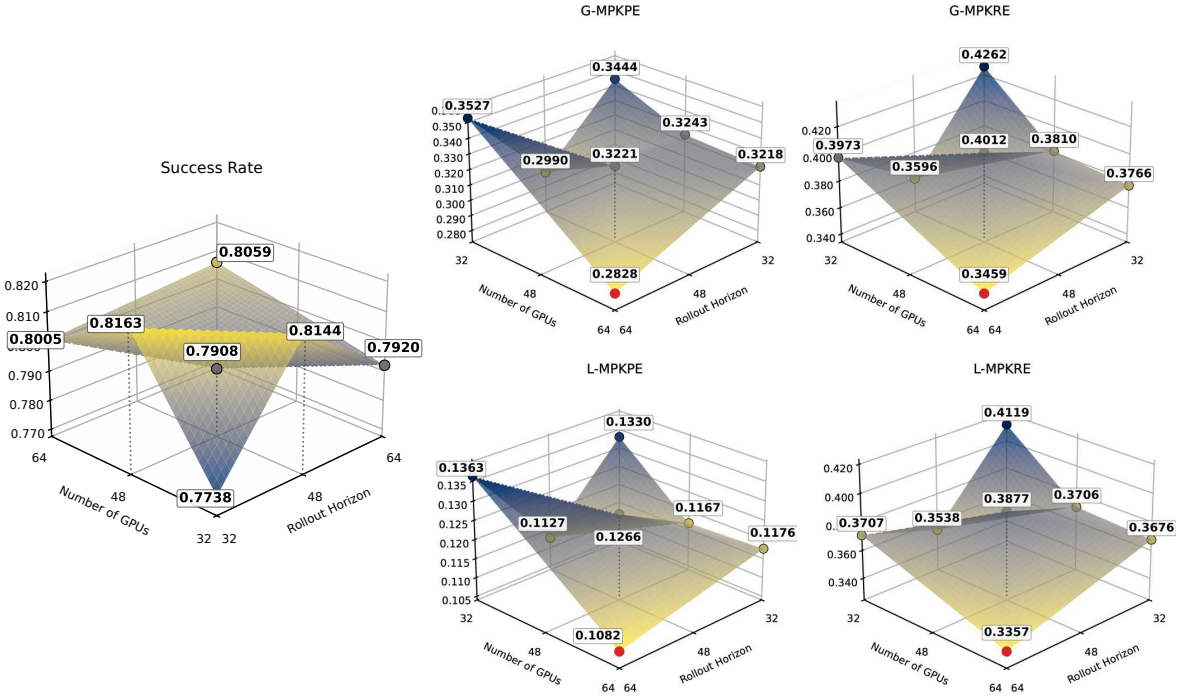


Figure 17: Results of scaling on-policy data collection on the Ours Test Set under root-and-upper-body mode. Best results with the highest success rate and lowest tracking errors are highlighted with red circles.

B.5 Scaling Reference Motions

We evaluate the effect of scaling reference motions by partitioning the full dataset into five nested subsets of increasing size, such that each larger subset fully contains all smaller ones. These subsets are denoted as XXS, XS, S, M, and L. Specifically, XXS, XS, and S correspond to one-third, two-thirds, and the complete BONES-SEED dataset excluding the held-out test split, respectively. The M subset comprises all reference motion datasets except Embody3D, whereas L contains the complete collection. Details on the statistics of these subsets are presented in table 8.

Subset	Motion Sequences	Motion Frames
XXS	44,073	16,159,416
XS	88,146	32,230,197
S	132,220	48,321,234
M	165,028	71,130,505
L	184,206	101,979,598

Table 8: Reference Motion Subset Statistics. The frames are calculated based on 50 FPS.

All training configurations and the model size are kept fixed across experiments, and the optimal on-policy data configuration (64 GPUs and a rollout horizon of 64) is adopted throughout. We still report results of two test sets under four representative control modes: root-and-hand mode (figure 18 and 19), root-and-end-effector mode (figure 20 and 21), root-and-upper-body mode (figure 22 and 23), and whole-body mode (figure 6 and 7). We evaluate the checkpoint obtained after 22,200 training epochs.

To provide context for interpreting the observed scaling behavior, we first quantitatively analyze the behavioral characteristics of the training partitions. Specifically, we define a behavioral feature as:

$$(h, r, p, v_x, v_y, \omega_z, j_p, j_v)$$

where h denotes the root height, r and p are the root roll and pitch angles, v_x and v_y are the root linear velocities on the XY-plane, ω_z is the root angular velocity along the z-axis, j_p is the joint positions and j_v is the joint velocities. The root velocities are all expressed in the humanoid heading frame. Together, these components form an expressive feature vector of 64 dimensions given Unitree G1 (1+2+2+1+29+29).

We extract such behavioral features from all motion frames, and fit a K-Means clustering model using 200,000 randomly sampled features from the complete L partition, yielding a shared clustering space consisting of 2,000 clusters. Subsequently, an equal number of behavioral features are randomly sampled from each partition and assigned to the nearest cluster in the shared clustering space. A cluster is considered occupied if at least one sampled feature is assigned to it. The **occupancy rate** is defined as the ratio of occupied clusters to the total number of clusters. A higher occupancy rate indicates that the partition spans a broader range of behavioral regions in the shared feature space and therefore exhibits greater behavioral coverage. It is worth noting that this quantitative measure is inherently dependent on the choice of reference motions used to construct the shared clustering space. In this paper, we use the complete partition and implicitly assume that it provides sufficiently broad behavioral support for meaningful comparisons across partitions. Nevertheless, no finite dataset can exhaust the space of possible behaviors and rare or long-tail behavioral patterns may still be absent. Consequently, determining an appropriate support for behavioral coverage analysis remains an important direction for future work.

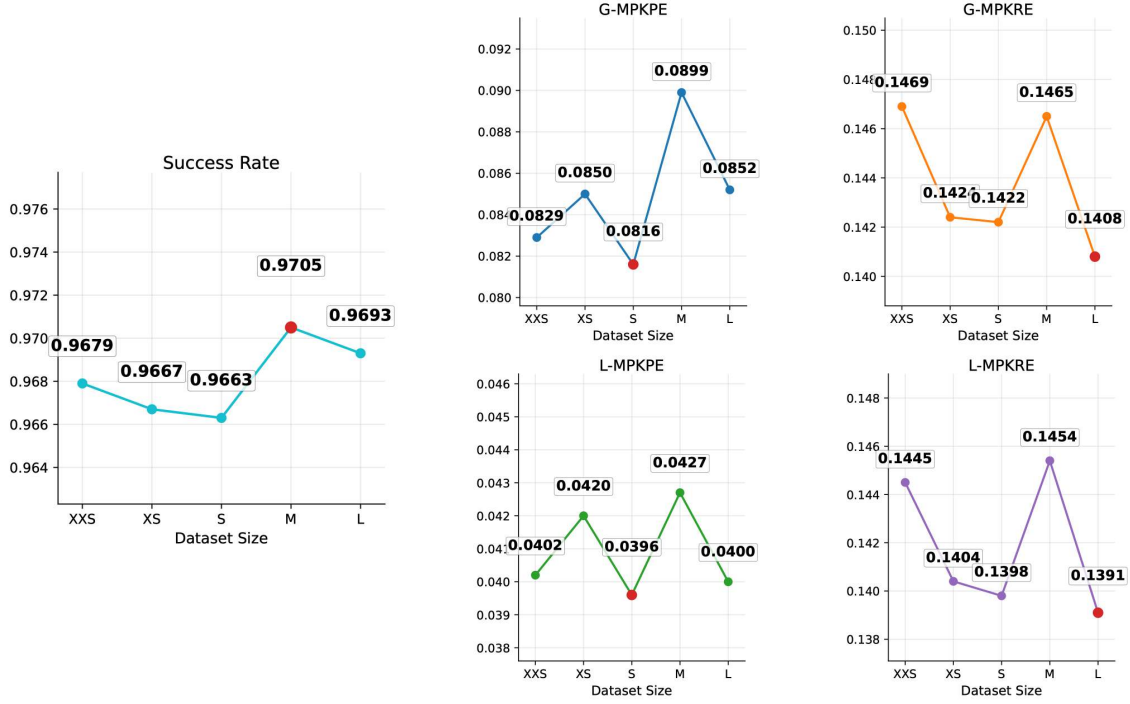


Figure 18: Results of scaling reference motions on BONES Test Set under root-and-hand mode. Best results corresponding to the highest success rate and lowest tracking errors are highlighted with red circles.

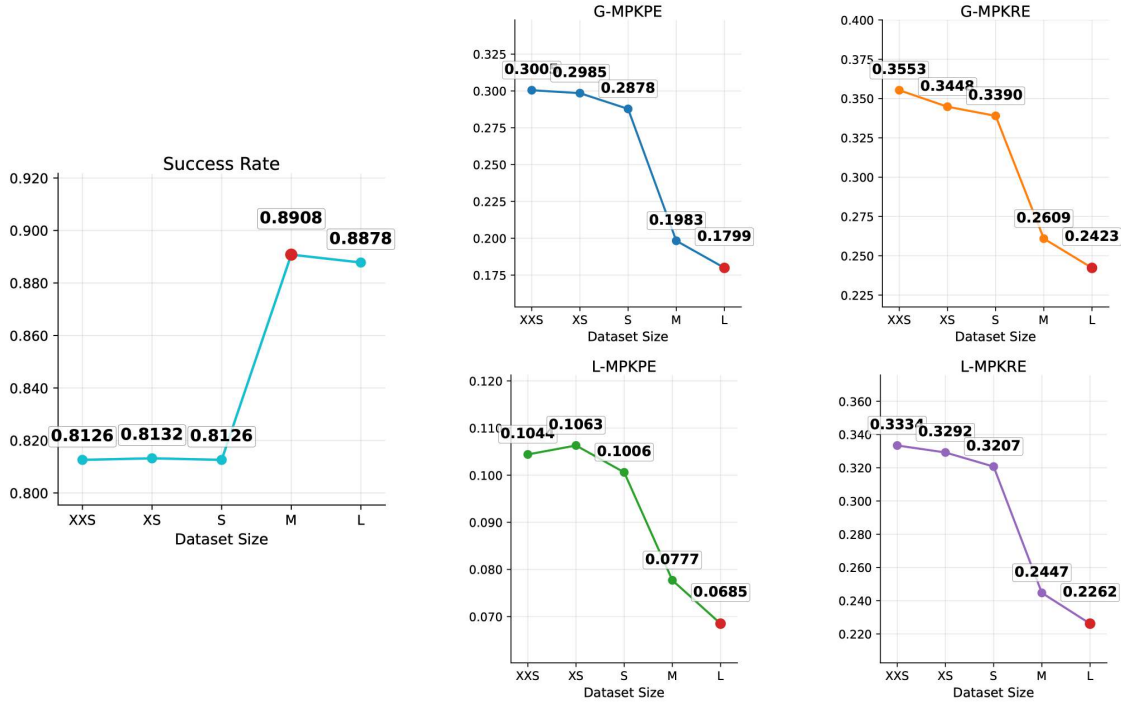


Figure 19: Results of scaling reference motions on Ours Test Set under root-and-hand mode. Best results corresponding to the highest success rate and lowest tracking errors are highlighted with red circles.

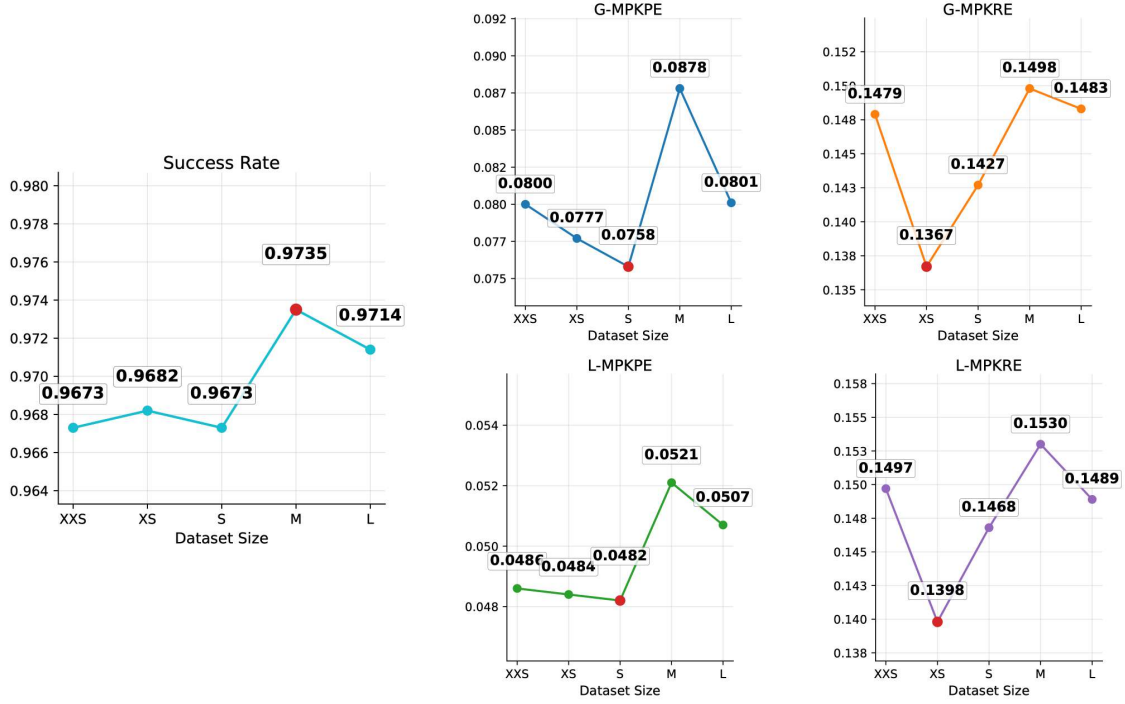


Figure 20: Results of scaling reference motions on BONES Test Set under root-and-end-effector mode. Best results corresponding to the highest success rate and lowest tracking errors are highlighted with red circles.

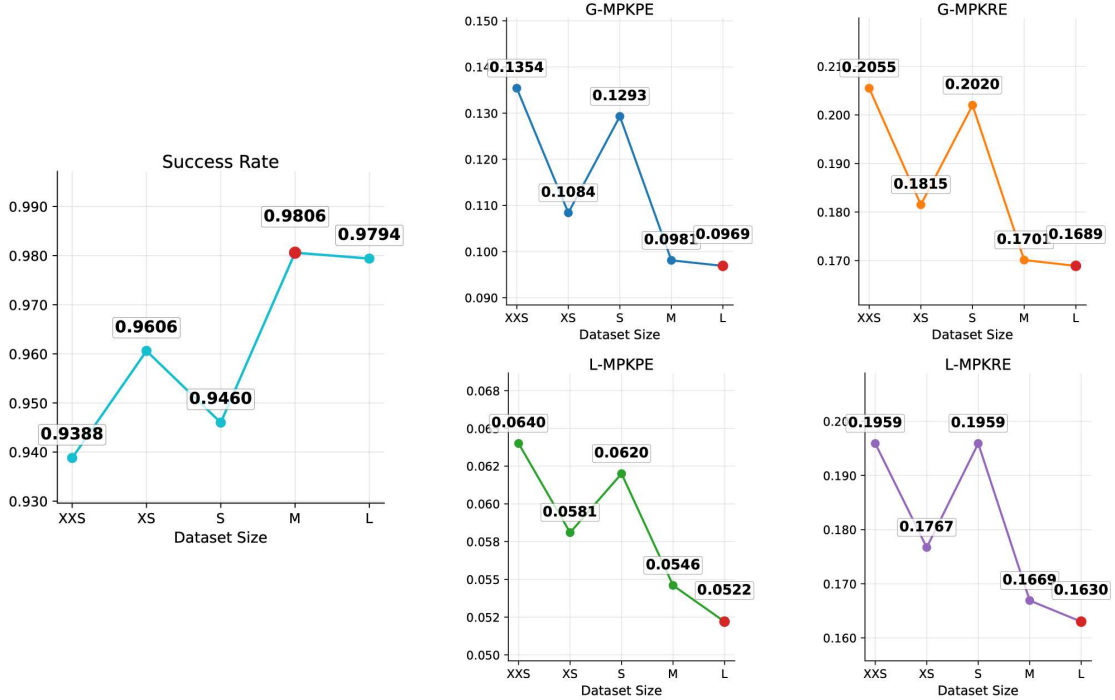


Figure 21: Results of scaling reference motions on Ours Test Set under root-and-end-effector mode. Best results corresponding to the highest success rate and lowest tracking errors are highlighted with red circles.

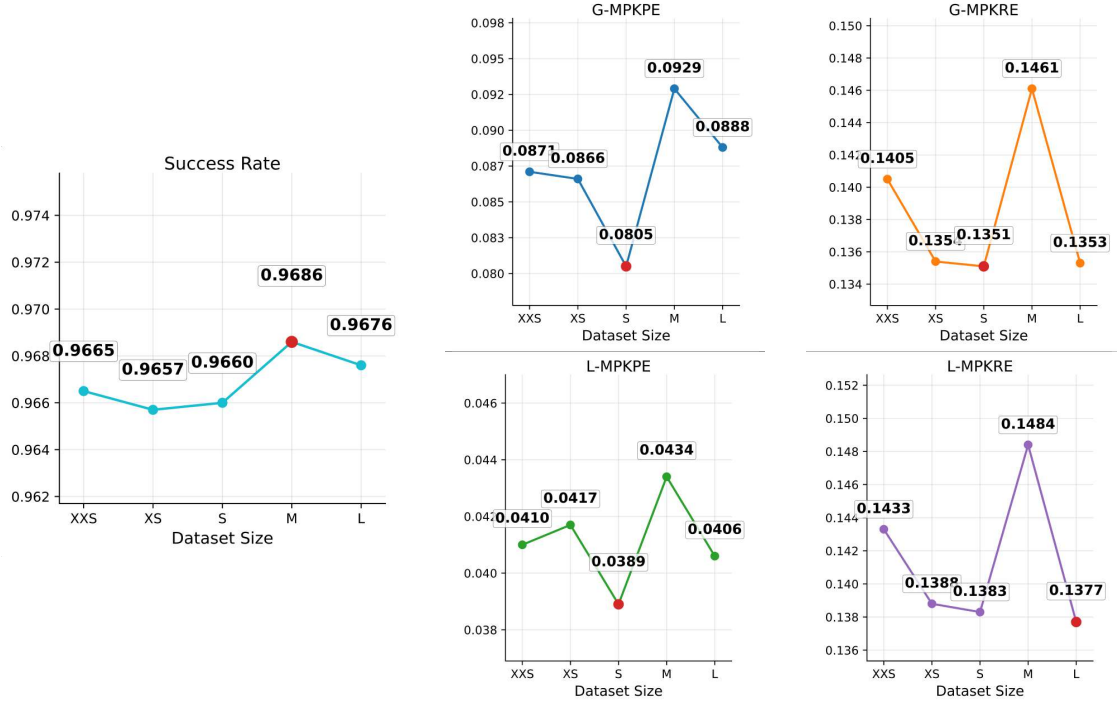


Figure 22: Results of scaling reference motions on BONES Test Set under root-and-upper-body mode. Best results corresponding to the highest success rate and lowest tracking errors are highlighted with red circles.

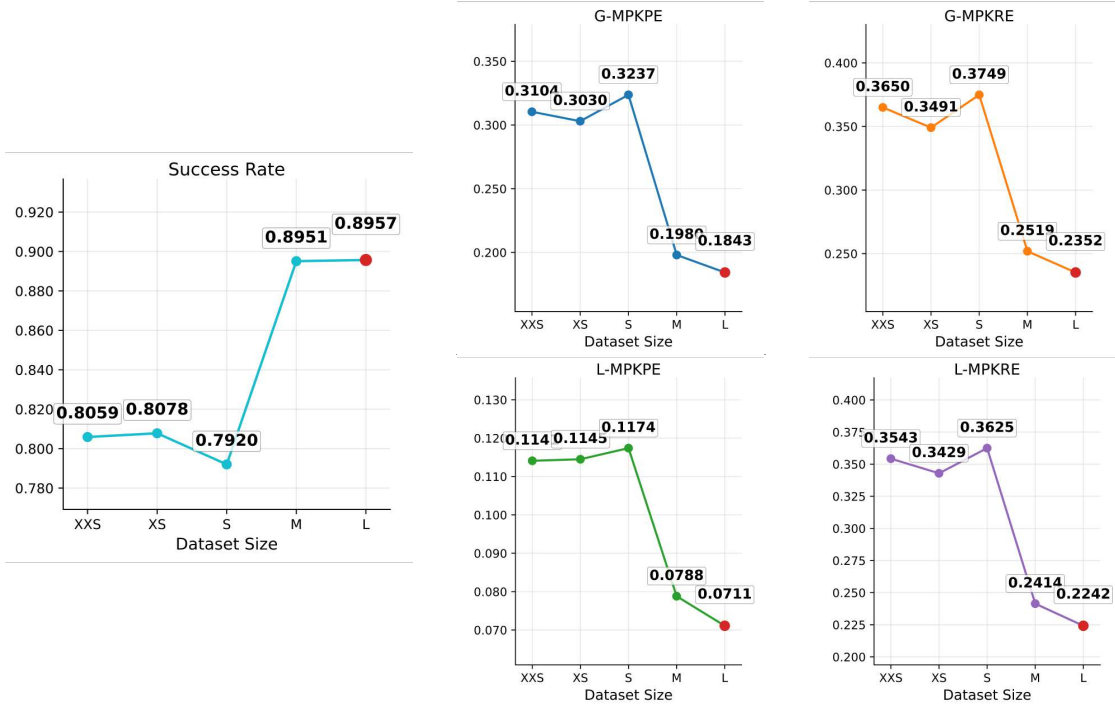


Figure 23: Results of scaling reference motions on Ours Test Set under root-and-upper-body mode. Best results corresponding to the highest success rate and lowest tracking errors are highlighted with red circles.

B.6 Scaling with Model Architectures

We next explore the effect of model architectures on the scaling behavior of BFM. Specifically, we examine whether the proposed Humanoid Transformer outperforms the MLP architecture and whether further increasing its capacity leads to additional performance gains. Throughout this paper, the actor and critic are instantiated with identical backbone architectures and model capacities. They differ only in their input and output dimensions, reflecting the asymmetric observations and prediction objectives.

All algorithmic and environmental configurations are kept fixed across experiments, including the PPO hyperparameters listed in table 7, the on-policy data collection configuration of 64 GPUs with a rollout horizon of 64, and the L partition of the reference motion dataset. We still report results of two test sets under four representative control modes: root-and-hand mode (figure 24 and 25), root-and-end-effector mode (figure 26 and 27), root-and-upper-body mode (figure 28 and 29), and whole-body mode (figure 8 and 9). We evaluate the checkpoint obtained after 22,200 training epochs for each experiment.

For the MLP backbone, we adopt two representative configurations following prior work. The small configuration contains 3.05M parameters, with hidden dimensions of [1024, 768, 512], while the large configuration contains 11.86M parameters, with hidden dimensions of [2048, 2048, 1024, 1024, 512, 512]. Both MLP variants employ ELU activations. No normalization is applied to the observations.

For the proposed Humanoid Transformer, we evaluate four model configurations by varying four critical components: the embedding dimension, feed-forward dimension, number of attention heads, and network depth. The S model contains 0.41M parameters, with an embedding dimension of 128, 2 attention heads, a feed-forward dimension of 128, and 2 Transformer layers. The M model contains 3.00M parameters, with an embedding dimension of 256, 4 attention heads, a feed-forward dimension of 256, and 4 layers. The L model contains 4.44M parameters and shares the same width as the M model but increases the depth to 6 layers, thereby isolating the effect of scaling network depth. Finally, the XL model contains 9.91M parameters, with an embedding dimension of 384, 6 attention heads, a feed-forward dimension of 384, and 6 layers, representing joint scaling of both network width and depth.

It is worth noting that our scaling study remains relatively limited compared with those conducted for Large Language Models. This limitation stems from two primary factors. First, humanoid robot control still lacks the mature infrastructure required to support truly large-scale pretraining. Distributed training remains fragile due to multi-GPU communication, simulator instability, and the limited exploration of efficient training systems tailored to humanoid foundation models, not to mention the community has yet to reach a consensus on the appropriate foundation model paradigm for humanoid robots. Second, practical deployment imposes strict computational constraints. To sustain an inference frequency above 50 Hz while reserving computational budget for future integration with high-level models, the parameter count of Behavior Foundation Models cannot be increased indefinitely under current hardware constraints. Nevertheless, as onboard computing capabilities continue to improve, establishing a scalable training pipeline is of great value for future humanoid foundation models. We are actively working to address these limitations with the constraints of current hardware and software systems. Nevertheless, we believe that substantial progress will ultimately require coordinated efforts from the broader community. To facilitate such progress, we will open-source all the resources, with the hope of providing a scalable foundation upon which the community can collectively build.

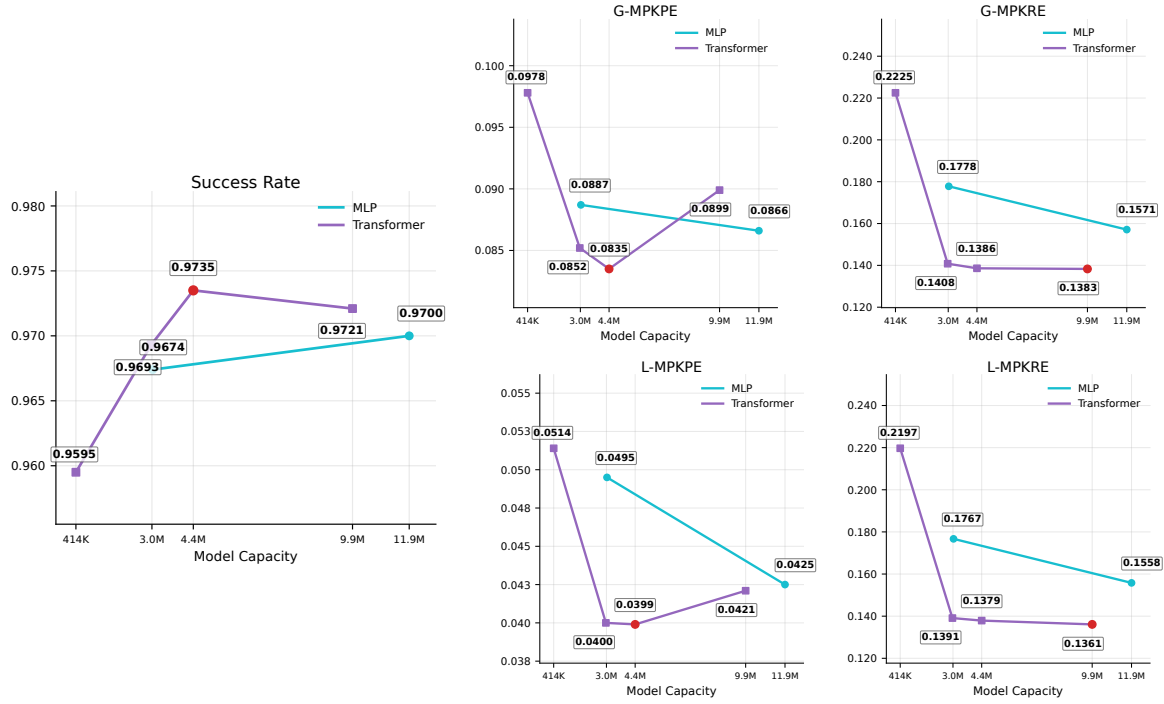


Figure 24: Results of scaling model architectures on **BONES Test Set** under **root-and-hand mode**. Best results corresponding to the highest success rate and lowest tracking errors are highlighted with red circles.

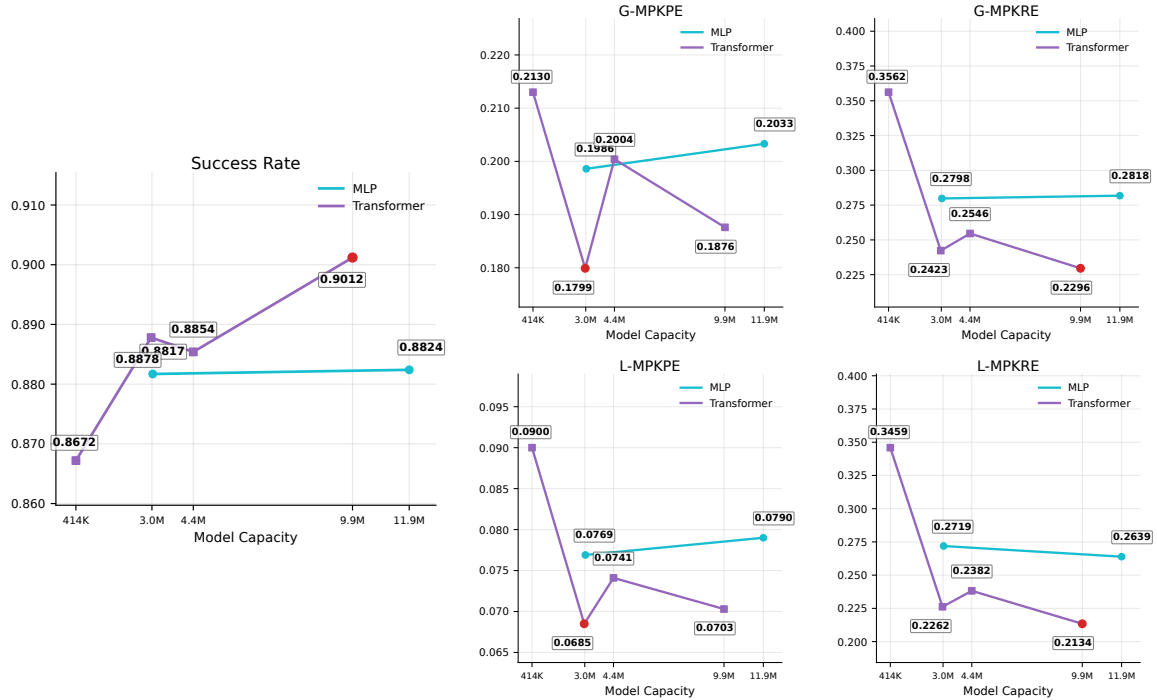


Figure 25: Results of scaling model architectures on **Ours Test Set** under **root-and-hand mode**. Best results corresponding to the highest success rate and lowest tracking errors are highlighted with red circles.

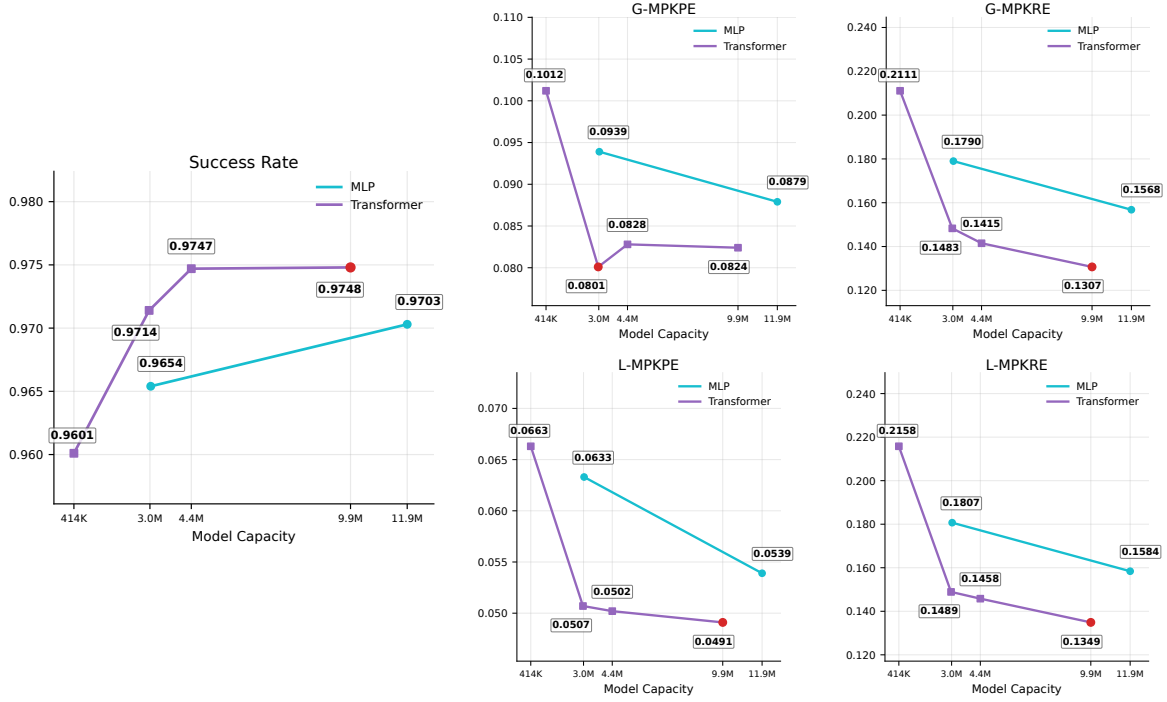


Figure 26: Results of scaling model architectures on **BONES Test Set** under **root-and-end-effector mode**. Best results corresponding to the highest success rate and lowest tracking errors are highlighted with red circles.

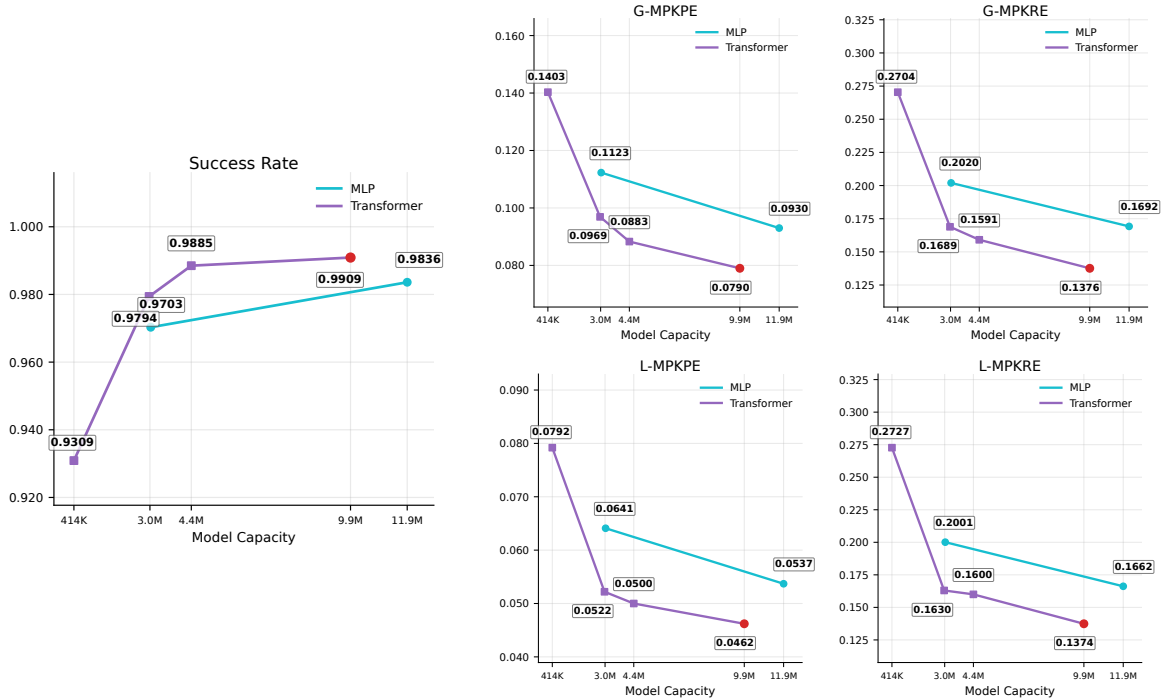


Figure 27: Results of scaling model architectures on **Ours Test Set** under **root-and-end-effector mode**. Best results corresponding to the highest success rate and lowest tracking errors are highlighted with red circles.

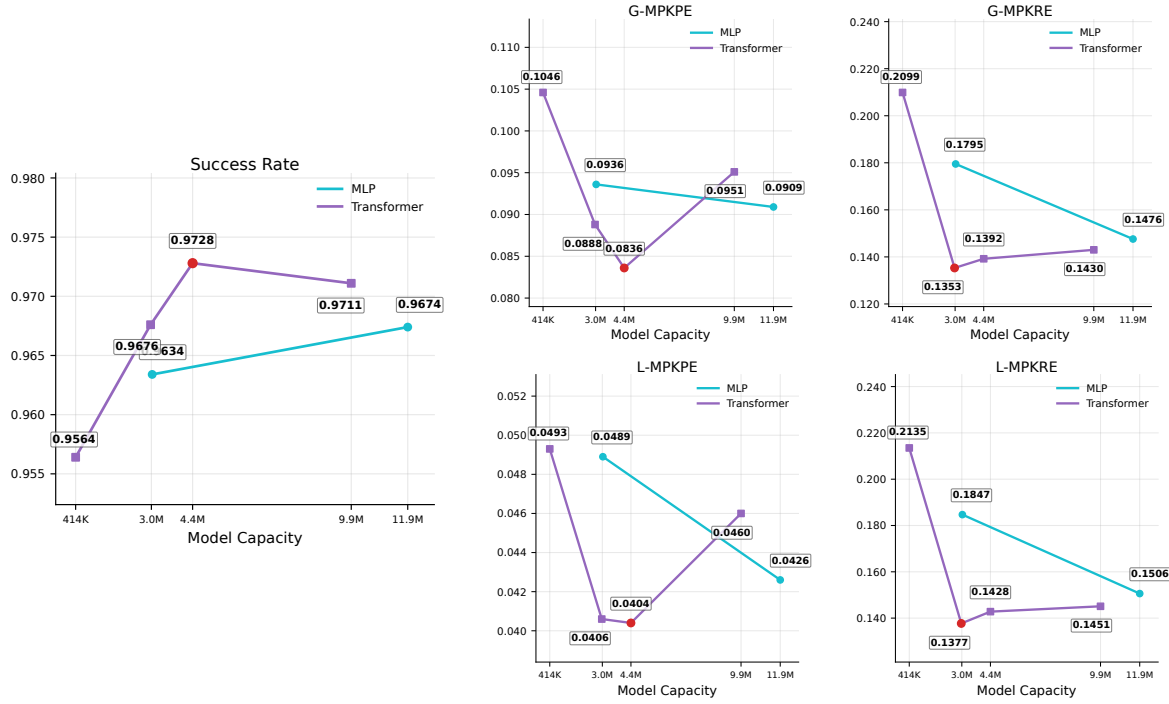


Figure 28: Results of scaling model architectures on BONES Test Set under root-and-upper-body mode. Best results corresponding to the highest success rate and lowest tracking errors are highlighted with red circles.

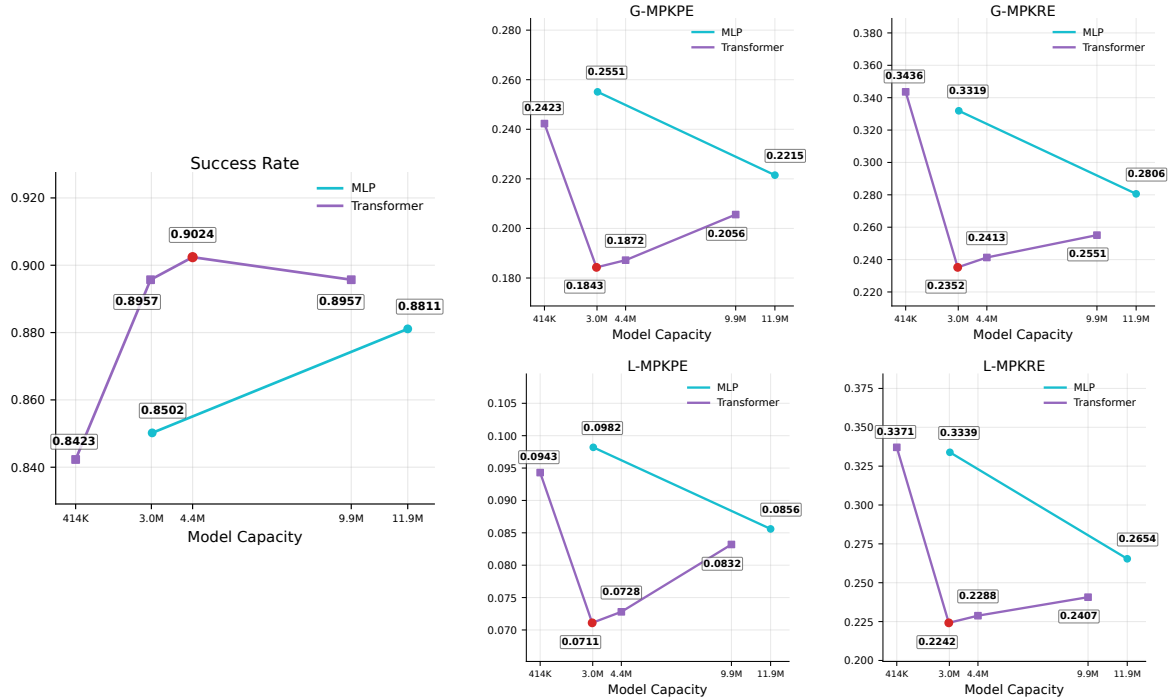


Figure 29: Results of scaling model architectures on Ours Test Set under root-and-upper-body mode. Best results corresponding to the highest success rate and lowest tracking errors are highlighted with red circles.