ADA-BFM: ADAPTING BEHAVIOR FOUNDATION MODELS FOR HUMANOID SCENE INTERACTION

Anonymous authors
Paper under double-blind review

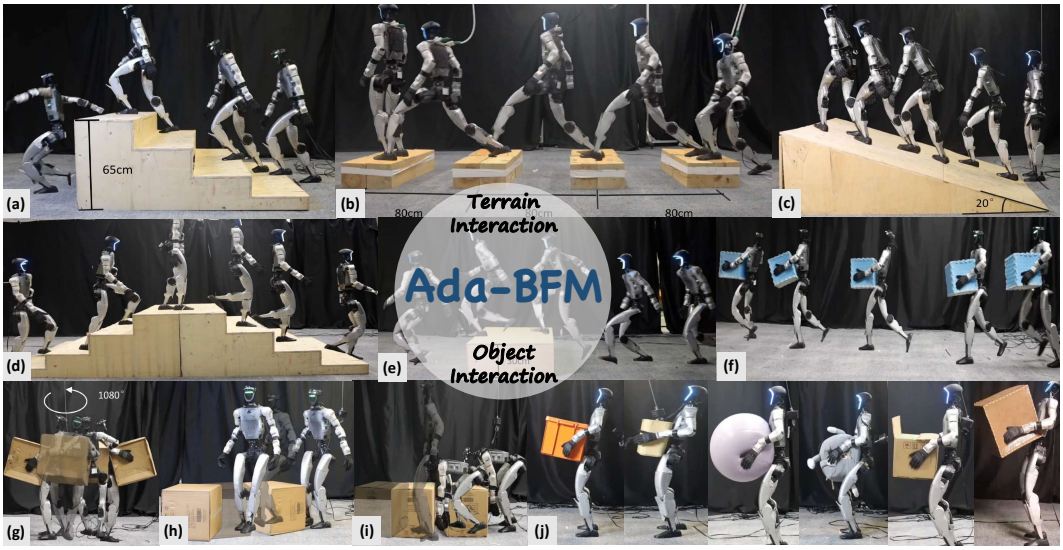


Figure 1: **Ada-BFM** extends the capabilities of pretrained BFMs to humanoid scene interaction, with a single adapted model shared across diverse tasks within each interaction domain. The terrain-interaction model supports (a) climbing stairs and jumping down; (b) repeatedly stepping over curbs; (c) ascending a slope; (d) ascending and descending stairs; and (e) stepping over obstacles. The object-interaction model supports (f) jogging while carrying a box; (g) performing a 1080° turn while holding a box; (h) kicking a box; (i) pushing a box; and (j) lifting objects of varying shapes, materials, and weights.

ABSTRACT

Behavior Foundation Models (BFMs) have emerged as a promising paradigm for general-purpose humanoid control through pretraining on large-scale behavioral data. However, their proficiency in whole-body coordination does not readily extend to scene interaction, which requires adapting motion to scene-dependent geometric constraints and contact dynamics. To bridge this gap, we introduce **Ada-BFM**, a post-training framework that adapts pretrained BFMs for humanoid scene interaction. Ada-BFM augments the original architecture with a lightweight adapter, enabling interaction-specific adjustments while leveraging pretrained control capabilities. Rather than producing a separate controller for each task, the framework yields a shared model for each interaction domain, supporting diverse tasks involving terrain or object interaction. Extensive experiments in simulation and the real world demonstrate that Ada-BFM enables diverse scene interaction behaviors, accelerates learning compared with training from scratch, and largely preserves pretrained behavioral competence. These findings establish Ada-BFM as an effective and efficient approach to extending the capabilities of BFMs beyond their pretraining settings while retaining their strengths in general whole-body control.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Humanoid whole-body control has advanced substantially in recent years, achieving impressive capabilities across diverse motor skills (Ben et al., 2025; Huang et al., 2025; He et al., 2025a). However, existing approaches often rely on extensive reward engineering tailored to specific tasks and contexts, limiting their versatility and hindering generalization across scenarios. To address these limitations, recent research has adopted behavior learning as a shared objective across whole-body control tasks, giving rise to Behavior Foundation Models (BFMs), large-scale pretrained humanoid controllers that enable robust and versatile control in response to diverse behavioral specifications (Zeng et al., 2025; Li et al., 2026). Motion tracking (Peng et al., 2018; Chen et al., 2025; Yin et al., 2026; Liao et al., 2026) is further identified as a unified and scalable task for pretraining these models, allowing BFMs to acquire increasingly rich behavioral knowledge and gain consistent improvements in control capabilities as pretraining scales (Luo et al., 2026; Zeng et al., 2026; Chen et al., 2026a).

Despite this growth in behavioral knowledge, existing BFM pretraining remains largely *scene-agnostic*, focusing on the realization of behavioral specifications without explicitly accounting for the surrounding environments. This limits BFMs' ability to interact with terrain and objects, where successful execution requires not only following control commands but adapting the execution to scene-dependent geometric constraints and contact dynamics (Yang et al., 2026; Wu et al., 2026b; Weng et al., 2025b; Yu et al., 2026b). Although a broad repertoire of behaviors does not, by itself, ensure effective scene interaction, the underlying behavioral knowledge provides a potentially valuable foundation for efficient transfer. This motivates scene-aware adaptation of BFMs that builds upon pretrained behavioral capabilities to meet the demands of physical interaction.

To this end, we propose **Ada-BFM**, a post-training framework that adapts the pretrained BFM for humanoid scene interaction. The framework is designed around three core adaptation objectives: 1) *Effectiveness*, enabling a single adapted model within each interaction domain to handle diverse tasks and outperform baseline approaches; 2) *Efficiency*, leveraging the behavioral knowledge acquired through pretraining to accelerate the acquisition of scene-interaction capabilities; and 3) *Retention*, preserving the broad behavioral repertoire of the pretrained model while minimizing degradation of its original control capabilities. Together, these objectives extend the capabilities of BFMs beyond their pretraining while preserving their strengths as general-purpose humanoid controllers.

As a concrete instantiation, Ada-BFM is built upon a pretrained BFM from ScaleBFM (Zeng et al., 2026). Its *Humanoid Transformer* backbone employs self-attention to capture temporal dependencies within behavioral trajectories and cross-attention to incorporate behavioral specifications that guide action generation. A learnable query token is appended to the end of the context sequence to aggregate control-relevant information and its output representation is used to predict actions. To enable scene-aware adaptation, Ada-BFM introduces a lightweight adapter into each Transformer layer to incorporate scene-relevant information. Each adapter comprises a gated cross-attention module and a gated feed-forward layer. This design mirrors the backbone's existing conditioning mechanism, treating scene inputs as another form of behavioral specifications. Rather than concatenating scene inputs with the original specifications, the adapters incorporate them through a separate conditioning pathway designed to minimize interference with pretrained control capabilities.

To evaluate the performance of the proposed framework in adapting pretrained BFMs for humanoid scene interaction, we consider two representative domains: terrain interaction and object interaction. Within each domain, we train a single adapted model to handle diverse tasks rather than a separate model for each task. Both domains use the same adaptation framework and training pipeline, differing in their reward configurations. Depth images provide a common representation of scene information across both domains. Throughout post-training, the pretrained backbone remains frozen and continues to process control commands, while the trainable adaptation pathway learns to leverage depth perception to modulate the resulting behavior for scene-aware physical interaction.

Comprehensive experiments in simulation and the real world demonstrate that Ada-BFM outperforms baseline approaches on terrain-interaction and object-interaction tasks, accelerates learning of scene-interaction behaviors compared with training from scratch and largely preserves pretrained control capabilities. These findings demonstrate the effectiveness, efficiency and retention of the proposed framework, establishing it as a promising approach for extending BFMs beyond their pretraining settings while retaining their strengths in general whole-body control.

2 Related Works

2.1 General Humanoid Motion Tracking and Behavior Foundation Models

Humanoid motion tracking obtains whole-body motor competence by imitating human motion data: physics-based imitation established that reinforcement learning reproduces complex reference motions (Peng et al., 2018; 2022), and later work scaled one controller to perpetual control, whole-body teleoperation, and universal tracking (Luo et al., 2023b; He et al., 2025b; Yin et al., 2026; Chen et al., 2025). Behavior foundation models turn the same recipe into pretraining: masked online distillation supports flexible control modes (Zeng et al., 2025), SONIC exposes a unified token interface (Luo et al., 2026), and ScaleBFM coordinates the tracking formulation and reference-motion diversity with a Humanoid Transformer (Zeng et al., 2026). Perceptive BFM instead synthesizes terrain-conformal references to supervise a depth-conditioned policy (Wang et al., 2026).

2.2 Humanoid Scene Interaction

Humanoid scene interaction splits into the two settings we evaluate: terrain interaction and object interaction. Perceptive Humanoid Parkour distills retargeted human skills into one depth-conditioned policy (Wu et al., 2026a). Hiking in the Wild maps depth and proprioception to joint actions in one reinforcement-learning stage, adding a terrain-edge foothold-safety mechanism and flat-patch sampling (Zhu et al., 2026). HDMI retargets human-object trajectories from monocular RGB video and co-tracks both states with a unified object representation and a residual action space (Weng et al., 2025a). OmniContact has a low-level tracker follow a contact flow of body trajectories and binary contact signals, extended by a rule-based planner (Yu et al., 2026a). PhysHSI pairs adversarial-motion-prior policy learning with a coarse-to-fine LiDAR-and-camera object localizer and validates carrying, sitting, and standing in the real world (Wang et al., 2025), and SceneBot conditions one tracker on per-link contact labels, with scene-interaction graphs reconstructed in hindsight from retargeted human motion (Chen et al., 2026b). None of them starts from a pretrained tracker, so the whole-body behavior and its scene conditioning are learned together.

2.3 Adapting Transformer Policies

Ada-BFM borrows from parameter-efficient adaptation—residual adapters (Houlsby et al., 2019; Rebuffi et al., 2017) and LoRA (Hu et al., 2022)—and from gated injection into frozen backbones: zero-initialized gated cross-attention (Alayrac et al., 2022), fixed-length visual token interfaces (Li et al., 2023), and zero-convolution branches (Zhang et al., 2023). Post-training on a narrow objective costs part of what a pretrained model already knows: preference alignment degrades the pretraining tasks the objective does not cover (Lin et al., 2024), instruction tuning forgets domain knowledge and reasoning (Luo et al., 2023a), and fine-tuning on benign data shifts behavior beyond the fine-tuned distribution (Qi et al., 2024). Parameter-efficient adaptation such as LoRA updates fewer parameters but still modifies the pretrained weights, so a policy adapted to one interaction task can lose the general tracking behavior it started from—the stability–plasticity trade-off that adaptation in this setting must resolve. Robotic cases are scarce: CLARE adds adapters to VLA models (Römer et al., 2026), AdaptNet modifies a controller's own layers (Xu et al., 2023). The closest relatives change the deployed network instead: residual policy learning adds a correction to a pretrained controller (Johannink et al., 2019; Silver et al., 2018), absorbing sim-to-real mismatch (He et al., 2025a), adding loco-manipulation to a general tracker (Zhao et al., 2025), or fitting one to teleoperation interfaces by distillation (Sun et al., 2026). Task Tokens (Vainshtein et al., 2025) also freezes the BFM but appends one learned goal token to its input sequence, adding no environment evidence.

3 Method

Our approach freezes the pretrained BFM and adds adapters for scene interaction. We instantiate the pretrained BFM as ScaleBFM (Zeng et al., 2026); the formulation applies to Transformer-based BFMs with compatible motion-conditioning interfaces. We first formalize the adaptation problem (Section 3.1), then describe the training setup (Section 3.2) and model architecture (Section 3.3).

3.1 Problem Formulation

We model each downstream interaction domain k as a Markov Decision Process (MDP) $\mathcal{M}^k = \langle \mathcal{S}^k, \mathcal{A}, \mathcal{T}^k, \mathcal{R}^k, \gamma \rangle$, where $\mathcal{S}^k$ and $\mathcal{A}$ are the state and action spaces, $\mathcal{T}^k$ is the transition dynamics, $\mathcal{R}^k$ is the reward, and γ is the discount factor. At timestep t, the controller observes the humanoid proprioception $\boldsymbol{s}_t^p$ and a motion goal $\mathcal{G}_t$ that specifies reference whole-body poses over a finite future window in the robot's root-relative frame. The action $\boldsymbol{a}_t \in \mathcal{A}$ represents desired joint positions, which are converted to joint torques by a low-level proportional–derivative controller.

Following the BFM formulation (Zeng et al., 2025; 2026), we define a humanoid behavior as the trajectory of proprioceptive states and actions over T control steps:

$$\mathcal{B} \equiv \left(\boldsymbol{s}_0^p, \boldsymbol{a}_0, \boldsymbol{s}_1^p, \boldsymbol{a}_1, \ldots, \boldsymbol{s}_{T-1}^p, \boldsymbol{a}_{T-1}, \boldsymbol{s}_T^p\right). \tag{1}$$

The goal $\mathcal{G}_t$ is not part of the behavior; it is an external specification of the behavior to execute. A pretrained BFM therefore implements a goal-conditioned policy $\pi_{\boldsymbol{\theta}_0}(\boldsymbol{a}_t \mid \mathcal{H}_t, \mathcal{G}_t)$, conditioned on the proprioception–action history $\mathcal{H}_t$ and the motion goal $\mathcal{G}_t$. $\boldsymbol{\theta}_0$ denotes parameters pretrained by motion tracking. This policy captures reusable balance and whole-body coordination but lacks the perceptual inputs needed for terrain and object interaction.

A downstream interaction domain adds observation $\boldsymbol{o}_t^k$ to this interface. Augmenting the frozen BFM with task-pathway parameters $\boldsymbol{\phi}_k$ yields the task-conditioned policy $\pi_{\boldsymbol{\theta}_0,\boldsymbol{\phi}_k}(\boldsymbol{a}_t \mid \mathcal{H}_t, \mathcal{G}_t, \boldsymbol{o}_t^k)$. The task pathway is trained once per domain and shared by all its tasks. With $\boldsymbol{\theta}_0$ fixed, we optimize $\boldsymbol{\phi}_k$ by on-policy reinforcement learning to maximize the expected discounted return:

$$\max_{\boldsymbol{\phi}_k} \mathbb{E}_{\pi_{\boldsymbol{\theta}_0,\boldsymbol{\phi}_k}} \left[\sum_{t=0}^{T-1} \gamma^t r_t^k\right]. \tag{2}$$

The expectation is over length-T rollouts in $\mathcal{M}^k$ under the adapted policy, with reward r_t^k at step t. Only the task pathway is trained; the pretrained backbone remains frozen. Adaptation starts from the pretrained controller without changing its initial action mean. Let $\boldsymbol{\phi}_k^{(0)}$ denote the initial task-pathway parameters and $\boldsymbol{\mu}$ the policy's action mean. For any valid inputs $(\mathcal{H}_t, \mathcal{G}_t, \boldsymbol{o}_t^k)$, initialization satisfies

$$\boldsymbol{\mu}_{\boldsymbol{\theta}_0,\boldsymbol{\phi}_k^{(0)}}(\mathcal{H}_t, \mathcal{G}_t, \boldsymbol{o}_t^k) = \boldsymbol{\mu}_{\boldsymbol{\theta}_0}(\mathcal{H}_t, \mathcal{G}_t). \tag{3}$$

3.2 Learning Paradigm

Motion tracking provides dense whole-body supervision throughout adaptation. We retain the pretrained tracking objective and motion interface, and use the task pathway to incorporate perception for terrain interaction and object interaction.

Policy optimization. We update the task pathway with proximal policy optimization (PPO). Each update collects a fixed number of steps per environment and then runs several epochs over shuffled minibatches. The gradient is assembled from four terms: the clipped policy-gradient surrogate, a clipped value regression, an entropy bonus $\mathbb{H}[\pi]$ that keeps the action distribution from collapsing, and a behavior-preservation loss $\mathcal{L}_{\mathrm{BFM}}$ that keeps the action mean close to the frozen BFM,

$$\begin{aligned} \mathcal{L} &= \mathcal{L}^{\mathrm{clip}} + c_v\, \mathcal{L}^{\mathrm{value}} - c_e\, \mathbb{H}[\pi] + \lambda_{\mathrm{BFM}}\, \mathcal{L}_{\mathrm{BFM}}, \\ \mathcal{L}_{\mathrm{BFM}} &= \mathbb{E}_t\left[\left\| \boldsymbol{\mu}_{\boldsymbol{\theta}_0,\boldsymbol{\phi}_k}(\mathcal{H}_t, \mathcal{G}_t, \boldsymbol{o}_t^k) - \boldsymbol{\mu}_{\boldsymbol{\theta}_0}(\mathcal{H}_t, \mathcal{G}_t) \right\|_2^2\right]. \end{aligned} \tag{4}$$

The preservation loss is the squared distance between the adapted action mean and the action mean of the frozen BFM under the same proprioception.

Reward design. The tracking reward is a weighted sum of global-frame tracking terms, penalty terms, and survival terms. Root-relative pose alone does not specify the robot's global trajectory. We therefore track root position together with whole-body pose along the same reference motion, constraining both global displacement and body configuration.

Domain randomization. We randomize physical parameters, including friction coefficients, default joint positions, and the base center of mass. We apply periodic random perturbations to robot velocities to simulate external pushes, and randomize depth camera extrinsics. For object-interaction tasks, we additionally randomize object scale, density, and friction.

Table 1: Reward terms of the two interaction domains. Entries in the two domain columns are the configured weights and σ is the kernel standard deviation; the environment further scales every term by the control period (0.02 s). A dash marks a term the domain does not use.

Term	σ	Terrain	Object
body height (pelvis)	0.3	0.5	0.5
body position (global)	0.3	1.0	1.0
body rotation (global)	0.4	1.0	1.0
body linear velocity	1.0	1.0	1.0
body angular velocity	3.14	1.0	1.0
object position (global)	0.3	–	1.0
object rotation (global)	0.6	–	1.0
object contact	–	–	0.3
action rate (ℓ_2)	–	-0.1	-0.1
joint limit	–	-10.0	-10.0
survival	–	1.0	1.0

Perceptual observations. The task observation $\boldsymbol{o}_t^k$ is a depth image. A noisy ray-caster camera captures the scene within its field of view at 60 Hz. Each depth image is cropped and resized to 18×32 pixels, clipped to the $[0, 2.5]$ m range, and normalized to $[0, 1]$.

3.3 Model Architecture

The pretrained Transformer backbone stays frozen, and the only trainable modules inside its augmented blocks are the residual adapters. Figure 2 shows the arrangement: the left panel places the adapters in the frozen stack, and the right panel expands one adapter into its two residual branches.

Frozen backbone. The frozen ScaleBFM actor comprises four Transformer blocks with a hidden width of 256, each applying self-attention, reference cross-attention, and a feed-forward network. A frozen task embedder maps six motion targets, masked by the control mode, to six reference tokens that every block uses as the keys and values of its reference cross-attention, and the action mean is read from a learnable readout token over the interleaved proprioception and action history.

Perception tokenization. For terrain interaction and object interaction, the depth observation $\boldsymbol{o}_t^k$ from Section 3.2 is encoded into a single task token. A small convolutional encoder compresses the depth frame into a compact feature, and a single linear tokenizer projects that feature into the backbone's token space. The encoder and the tokenizer are trainable and are optimized together with the adapters, while the pretrained backbone stays frozen.

Adapters. We insert each adapter between the reference cross-attention and the frozen feed-forward network of a block. The block state attends to the task token through a cross-attention of its own, and the attended state passes through a feed-forward branch. Both modules are trained from scratch. Each branch returns through a learned scalar gate initialized to zero, so the adapted network is exactly the pretrained BFM at initialization.

4 Experiments

We evaluate whether a frozen behavior foundation model acquires perception-conditioned interaction skills through the task pathway, on two task families: terrain interaction, in which the robot tracks motion over irregular ground, and object interaction.

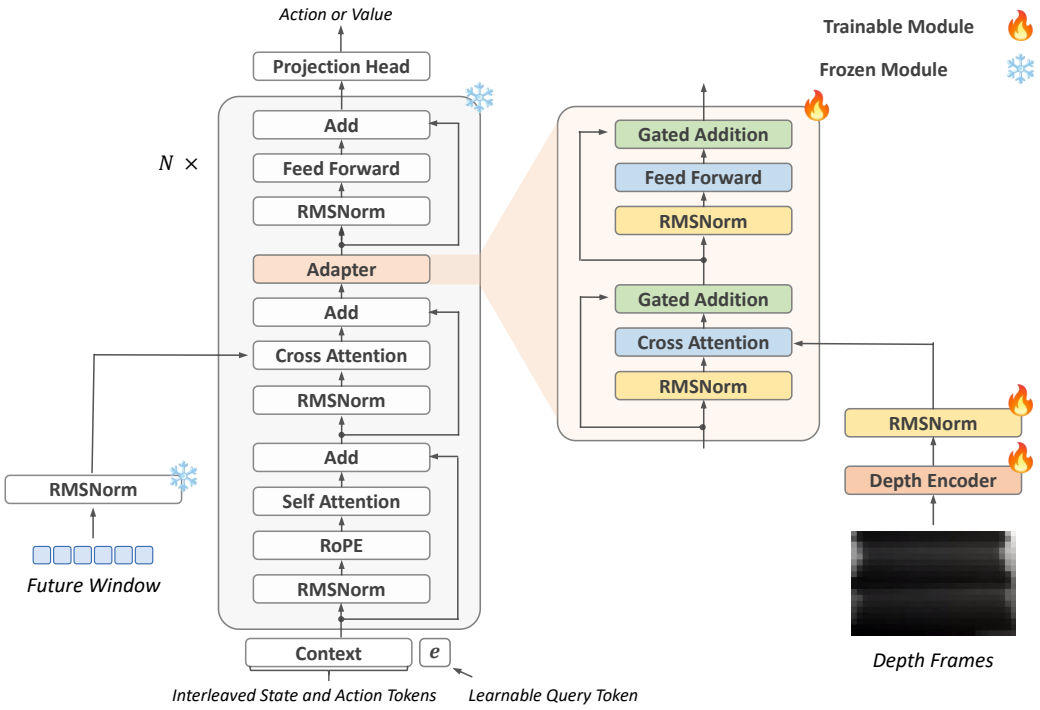


Figure 2: Ada-BFM adapts a pretrained BFM to a new interaction task by freezing the backbone and training a separate task pathway. Left: the pretrained backbone consists of N Transformer blocks and remains fully frozen. The only trainable modules within the augmented blocks are the adapters, inserted after each selected block's reference cross-attention and before its frozen feed-forward network. Right: one adapter expands into two residual branches, each with its own normalization and the gate, and the depth frames enter through the trainable depth encoder that produces the task token.

4.1 Experimental Setup

Environments. We conduct experiments on a Unitree G1 humanoid that stands 1.3 meters tall and has 29 degrees of freedom. All models are trained and evaluated in Isaac Lab.

Evaluation Protocol. BFM pretraining is instantiated as motion tracking, so tracking accuracy under the diverse control modes is a direct evaluation protocol. We therefore report the success rate (SR) together with global and local tracking errors, and evaluate SR on three tasks: flat-ground, terrain interaction, and object interaction. On the first two, a motion sequence is unsuccessful if any selected link deviates from its desired state in the global frame by more than a predefined threshold. Object interaction adds a second condition: the object's deviation from its desired state in the global frame exceeds a predefined threshold continuously over a specified time window.

Dataset. For data interacting with terrain, reference motions come from GRAIL (Xie et al., 2026) and OmniRetarget (Yang et al., 2026). For data interacting with objects, reference motions come from OmniRetarget (Yang et al., 2026). We randomly partition the clips into 80% for training and 20% for testing, and the test split is held fixed as the evaluation manifest shared by every method. To examine whether the adapted policy preserves the motion capability of the original BFM, we additionally evaluate on the BONES Test Set.

Baselines and fairness protocol. We compare against (i) the frozen ScaleBFM and SONIC (Luo et al., 2026) checkpoints; (ii) full fine-tuning of the pretrained BFM; and (iii) keep the BFM frozen and training only the action head; (iv)LoRA (Hu et al., 2022), and (v) Task token Vainshtein et al. (2025) by concatenating perception tokens along the sequence dimension of the existing cross-

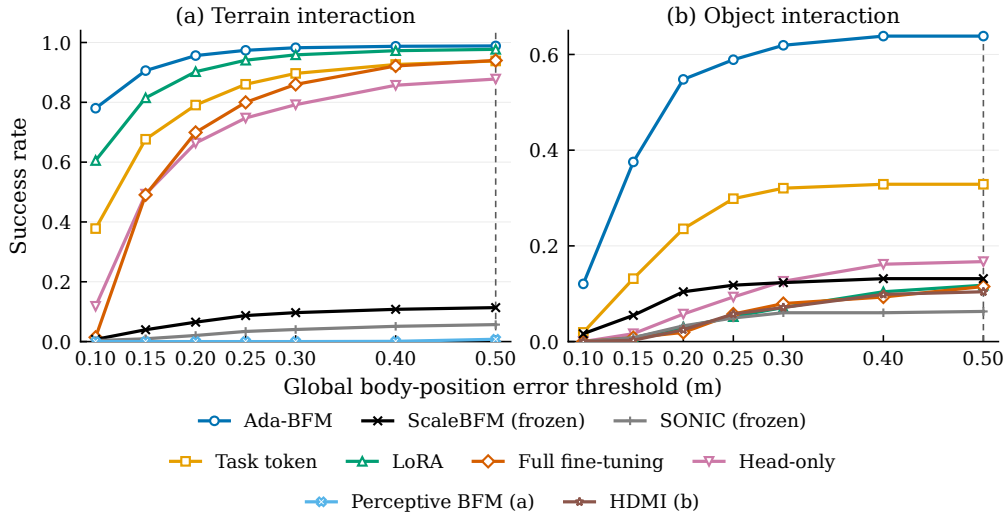


Figure 3: Success rate against the success criterion for the adapted methods, the two frozen checkpoints that fix the reference level, and the published baseline of each setting, Perceptive BFM in (a) and HDMI in (b). Each curve is one method evaluated on the fixed test manifest, and the horizontal axis sweeps the global body-position error threshold that defines success.

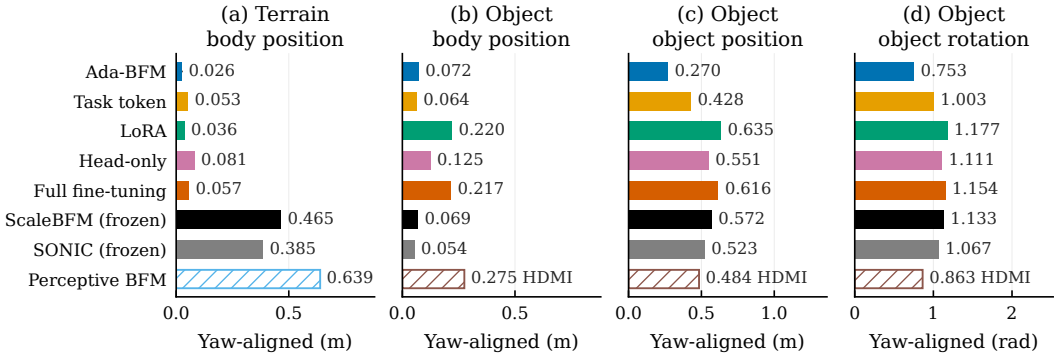


Figure 4: Local tracking errors of the compared policies. (a) Terrain interaction and (b) object interaction report the yaw-aligned body-position error, (c) the yaw-aligned object-position error of object interaction, and (d) its yaw-aligned object-rotation error. The hatched bar is that setting's published baseline, Perceptive BFM in (a) and HDMI in (b) to (d).

attention. We also compare against published methods: Perceptive BFM (Wang et al., 2026) for terrain interaction, and HDMI (Weng et al., 2025a) for object interaction.

4.2 Main Results

We evaluate Ada-BFM against three objectives (Section 1): *Effectiveness*, *Efficiency*, and *Retention*.

Effectiveness. Does a single adapted model within each interaction domain handle diverse tasks and outperform baseline approaches? Figure 1 shows the behaviours behind this question: one terrain-interaction model climbs stairs, steps over curbs, ascends slopes, and clears obstacles, and one object-interaction model carries, turns with, kicks, and pushes a box and lifts objects of varying shapes, materials, and weights. Figure 3 compares all adaptation strategies under the protocol of Section 4.1, sweeping the success criterion on both scene-interaction test sets. The frozen ScaleBFM checkpoint marks the capability boundary of the pretrained BFM on both scene-interaction tasks: it

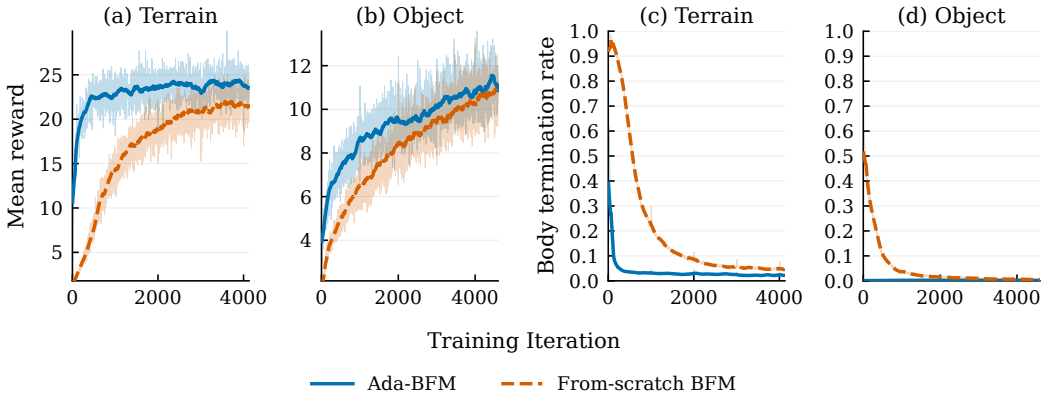


Figure 5: Training comparison between reusing the BFM's motion prior (Ada-BFM) and training from scratch. As shown in (a) and (b), Ada-BFM converges faster than the from-scratch BFM and converges to a higher reward. As shown in (c) and (d), reusing the prior drives the body-position termination rate down to a few percent within the first few hundred iterations, whereas training from scratch spends much of its effort on learning whole-body coordination and interaction.

reaches only 0.114 and 0.132 on the terrain- and object-interaction test sets, respectively. However, as shown in Figure 3, in terrain interaction Ada-BFM attains a success rate of 0.989 on the test set at the 0.5 m global body-position error threshold. It exceeds every baseline. At the 0.1 m threshold, its success rate is 0.780, 0.175 above the second-best method, LoRA. In object interaction, Ada-BFM attains 0.638 at the 0.5 m threshold, 0.310 above the second-best method, task token (0.329). These results confirm that a single adapter per interaction domain adapts the pretrained BFM to that domain's scene-interaction tasks and outperforms every baseline. Figure 4 reports the same policies under local criteria: in terrain interaction the yaw-aligned body-position error keeps the ordering of Figure 3, while in object interaction the body-side error no longer separates the adapted policies from the frozen checkpoints and the object-side errors do.

Efficiency. Does leveraging the behavioral knowledge acquired through pretraining accelerate the acquisition of scene-interaction capabilities? Figure 5 shows how adaptation proceeds. Panels (a) and (b) give the mean training reward in terrain interaction and object interaction, panels (c) and (d) the body-position termination rate on the same tasks. Ada-BFM reaches 90% of its final terrain-interaction reward by iteration 330, a level the from-scratch BFM reaches near iteration $3{,}100$; Ada-BFM passes the reward that run reaches at the end of its $4{,}129$ iterations by iteration 360, and settles more than 2 above it. In object interaction the two runs end within 0.5 of each other over the last tenth of training, so the object-interaction advantage is concentrated in the early phase. Its termination rate falls from 40% to below 5% by iteration 250 in terrain interaction and stays below 0.5% throughout object interaction, where the from-scratch run needs about $4{,}100$ iterations to reach the same rate; the from-scratch policy spends its early training terminating up to 97% of episodes in terrain interaction and 52% in object interaction. This demonstrates that reusing the behavioral knowledge acquired through pretraining is a more efficient route to scene-interaction competence than training from scratch.

Retention. Does the adaptation preserve the broad behavioral repertoire of the pretrained model while minimizing degradation of its original control capabilities? Table 2 reports the BONES retention of each adapted checkpoint. After adaptation on terrain interaction and object interaction, Ada-BFM retains success of 0.928 and 0.878. These values are only 0.040 and 0.090 below the frozen checkpoint's 0.968. The weight-updating baselines lose markedly more under the same adaptation setting: full fine-tuning, head-only, and LoRA fall to 0.734, 0.875, and 0.849 after terrain adaptation, and to 0.408, 0.465, and 0.351 after object adaptation, where each of the three keeps less than half of the frozen success. The task token, which also leaves the backbone frozen, retains slightly more than Ada-BFM on both domains (0.936 and 0.929 against 0.928 and 0.878) but attains far less on the interaction tasks (0.329 against 0.638 in object interaction; Figure 3), so the two methods

Table 2: Retention on the BONES Test Set after adaptation, at the 0.5 m global body-position threshold of that benchmark. Every adapted checkpoint is evaluated with its task branch enabled, on the same 10,000-clip manifest.

Task	Method	BONES SR↑
–	ScaleBFM (frozen)	0.968
Terrain	Full fine-tuning	0.734 $_{\downarrow 0.234}$
	Head-only	0.875 $_{\downarrow 0.093}$
	LoRA	0.849 $_{\downarrow 0.119}$
	Task token	0.936 $_{\downarrow 0.032}$
	Ada-BFM	0.928 $_{\downarrow 0.040}$
Object	Full fine-tuning	0.408 $_{\downarrow 0.560}$
	Head-only	0.465 $_{\downarrow 0.503}$
	LoRA	0.351 $_{\downarrow 0.617}$
	Task token	0.929 $_{\downarrow 0.039}$
	Ada-BFM	0.878 $_{\downarrow 0.090}$

trade retention against interaction competence. The design accounts for the small degradation: the pretrained backbone stays frozen, the residual pathway returns through gates initialized to zero, so the adapted network is exactly the pretrained BFM at initialization, and the preservation loss $\mathcal{L}_{\mathrm{BFM}}$ holds the adapted action mean close to the frozen one during optimization. These results indicate that Ada-BFM preserves the broad behavioral repertoire of the pretrained model while keeping the degradation of its original control capability small.

5 Discussion

The two domains differ in how much of the adaptation problem the pretrained prior already answers. Terrain interaction is tracking under a new observation: the reference motion still specifies the whole-body trajectory to realize, the terrain changes only the geometry on which it is realized, and the reward is the pretrained tracking reward unchanged, so the task pathway modulates a behavior that the frozen backbone already generates. Object interaction is a co-tracking problem instead: its reference carries the object's motion as well as the robot's, its reward adds object position, rotation, and contact terms, and the pretrained behavior contains no object, so the pathway must produce object coordination through frozen weights that never represented one while the preservation loss pulls the action mean back toward the pretrained tracker. The results follow this difference. At the 0.5 m threshold of the evaluation protocol, Ada-BFM reaches a success rate of 0.989 on the terrain manifest, ahead of every baseline, and 0.638 in object interaction, where the local body-position error no longer separates the adapted policies from the frozen checkpoints. Retention splits the same way: the adapted checkpoint keeps 0.928 of BONES success after terrain adaptation against 0.878 after object adaptation, and the policies that do update the pretrained weights lose far more in the object domain (0.408 for full fine-tuning). Adapting to terrain interaction is in this sense the easier and the better-behaved of the two, and closing the object gap may require a pathway that represents the object explicitly, or pretraining that includes object interaction.

6 Conclusion

We presented Ada-BFM, a post-training framework that adapts a pretrained humanoid behavior foundation model to scene-interaction tasks. It reuses the frozen BFM by training only a residual task pathway, which encodes egocentric depth into task tokens behind zero-initialized gated adapters, converges in a fraction of the iterations training from scratch needs, and leaves the pretrained tracker recoverable by switching the pathway off. It extends the capability boundary of the BFM to terrain interaction and object interaction with a single adapter per setting, and it preserves the general motion prior of the original BFM. Ada-BFM is a promising post-training framework that we expect to extend to external information beyond egocentric depth and to downstream tasks beyond the two settings studied here.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks that require disclosure: implementing the method, specifically the training and evaluation code for the task pathway; designing experiments and providing feedback on the research methodology; and proposing and refining hypotheses. We have not used generative AI tools for the remaining tasks that require disclosure—generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, translation, cleaning and reformatting datasets, qualitative and thematic data analysis, and interpreting results. Additionally, we used generative AI tools for the following tasks with recommended disclosure: creating and editing software code for the analysis, plotting, and verification scripts, drafting parts of this paper, and editing it for readability.

We have reviewed all AI-assisted work. The AI-assisted code was read by the authors and run through the real training and evaluation pipelines, and every figure and numerical result reported in this paper was checked against the experiment logs and raw data. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022. arXiv:2204.14198.

Qingwei Ben, Feiyu Jia, Jia Zeng, Junting Dong, Dahua Lin, and Jiangmiao Pang. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. *arXiv preprint arXiv:2502.13013*, 2025.

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*, 2026a.

Sirui Chen, Shibo Zhao, Zhen Wu, Jiaman Li, Guanya Shi, and C. Karen Liu. SceneBot: Contact-prompted general humanoid whole body tracking with scene-interaction. *arXiv preprint arXiv:2606.27581*, 2026b.

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.

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*, 2025a.

Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. OmniH2O: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *Proceedings of The 8th Conference on Robot Learning*, volume 270, pp. 1516–1540. PMLR, 2025b. arXiv:2406.08858.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, volume 97, pp. 2790–2799. PMLR, 2019. arXiv:1902.00751.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. arXiv:2106.09685.

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.

Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Residual reinforcement learning for robot control. In *IEEE International Conference on Robotics and Automation*, pp. 6023–6029, 2019. arXiv:1812.03201.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, volume 202, pp. 19730–19742. PMLR, 2023. arXiv:2301.12597.

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. In *International Conference on Learning Representations*, volume 2026, pp. 79697–79725, 2026.

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. *Science Robotics*, 11(117):eadx8924, 2026.

Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. Mitigating the alignment tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 580–606. Association for Computational Linguistics, 2024. arXiv:2309.06256.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023a.

Zhengyi Luo, Jinkun Cao, Alexander Winkler, Kris Kitani, and Weipeng Xu. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10895–10904, 2023b. arXiv:2305.06456.

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117):eaed4592, 2026.

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.

Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. ASE: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions on Graphics*, 41(4):1–17, 2022. doi: 10.1145/3528223.3530110. arXiv:2205.01906.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *International Conference on Learning Representations*, 2024. arXiv:2310.03693.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems*, 2017. arXiv:1705.08045.

Ralf Römer, Yi Zhang, Yuming Li, and Angela P. Schoellig. CLARE: Continual learning for vision-language-action models via autonomous adapter routing and expansion. *IEEE Robotics and Automation Letters*, 11(7):8331–8338, 2026. doi: 10.1109/LRA.2026.3693992. arXiv:2601.09512.

Tom Silver, Kelsey Allen, Josh Tenenbaum, and Leslie Kaelbling. Residual policy learning. *arXiv preprint arXiv:1812.06298*, 2018.

Zhenguo Sun, Bo-Sheng Huang, Yibo Peng, Xukun Li, Jingyu Ma, Yu Sun, Zhe Li, Haojun Jiang, Biao Gao, Zhenshan Bing, Xinlong Wang, and Alois Knoll. MOSAIC: Bridging the sim-to-real gap in generalist humanoid motion tracking and teleoperation with rapid residual adaptation. *arXiv preprint arXiv:2602.08594*, 2026.

Ron Vainshtein, Zohar Rimon, Shie Mannor, and Chen Tessler. Task tokens: A flexible approach to adapting behavior foundation models. *arXiv preprint arXiv:2503.22886*, 2025. URL `https://arxiv.org/abs/2503.22886`.

Huayi Wang, Wentao Zhang, Runyi Yu, Tao Huang, Junli Ren, Feiyu Jia, Zirui Wang, Xiaojie Niu, Xiao Chen, Jiahe Chen, Qifeng Chen, Jingbo Wang, and Jiangmiao Pang. PhysHSI: Towards a real-world generalizable and natural humanoid-scene interaction system. *arXiv preprint arXiv:2510.11072*, 2025.

Zifan Wang, Yizhao Li, Teli Ma, Qiang Zhang, Yudong Fan, Hao Xu, Shuo Yang, and Junwei Liang. Perceptive behavior foundation model: Adapting human motion priors to robot-centric terrain. In *Conference on Robot Learning*, 2026. To appear. arXiv:2606.08059.

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*, 2025a.

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*, 2025b.

Zhen Wu, Xiaoyu Huang, Lujie Yang, Yuanhang Zhang, Xi Chen, Pieter Abbeel, Rocky Duan, Angjoo Kanazawa, Carmelo Sferrazza, Guanya Shi, and C. Karen Liu. Perceptive humanoid parkour: Chaining dynamic human skills via motion matching. *arXiv preprint arXiv:2602.15827*, 2026a.

Zhen Wu, Xiaoyu Huang, Lujie Yang, Yuanhang Zhang, Xi Chen, Pieter Abbeel, Rocky Duan, Angjoo Kanazawa, Carmelo Sferrazza, Guanya Shi, et al. Perceptive humanoid parkour: Chaining dynamic human skills via motion matching. *arXiv preprint arXiv:2602.15827*, 2026b.

Tianyi Xie, Haotian Zhang, Jinhyung Park, Zi Wang, Bowen Wen, Jiefeng Li, Xueting Li, Qingwei Ben, Haoyang Weng, Yufei Ye, David Minor, Tingwu Wang, Chenfanfu Jiang, Sanja Fidler, Jan Kautz, Linxi Fan, Yuke Zhu, Zhengyi Luo, Umar Iqbal, and Ye Yuan. GRAIL: Generating humanoid loco-manipulation from 3d assets and video priors. *arXiv preprint arXiv:2606.05160*, 2026.

Pei Xu, Kaixiang Xie, Sheldon Andrews, Paul G. Kry, Michael Neff, Morgan McGuire, Ioannis Karamouzas, and Victor Zordan. AdaptNet: Policy adaptation for physics-based character control. *ACM Transactions on Graphics*, 42(6):1–17, 2023. doi: 10.1145/3618375. arXiv:2310.00239.

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. In *IEEE International Conference on Robotics and Automation*, 2026. arXiv:2509.26633.

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.

Runyi Yu, Xiaoyi Lin, Ji Ma, Yinhuai Wang, Koukou Luo, Jiahao Ji, Huayi Wang, Wenjia Wang, Runhan Zhang, Ping Tan, Ting Wu, Ruoli Dai, Qifeng Chen, and Lei Han. OmniContact: Chaining meta-skills via contact flow for generalizable humanoid loco-manipulation. *arXiv preprint arXiv:2606.26201*, 2026a.

Runyi Yu, Xiaoyi Lin, Ji Ma, Yinhuai Wang, Koukou Luo, Jiahao Ji, Huayi Wang, Wenjia Wang, Runhan Zhang, Ping Tan, et al. Omnicontact: Chaining meta-skills via contact flow for generalizable humanoid loco-manipulation. *arXiv preprint arXiv:2606.26201*, 2026b.

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.

Weishuai Zeng, Kangning Yin, Xiaojie Niu, Shunlin Lu, Weixiang Zhong, Jiahe Chen, Feiyu Jia, Xiao Chen, Zirui Wang, Furui Xu, et al. Scaling behavior foundation model for humanoid robots. *arXiv preprint arXiv:2607.15163*, 2026.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023. arXiv:2302.05543.

Siheng Zhao, Yanjie Ze, Yue Wang, C. Karen Liu, Pieter Abbeel, Guanya Shi, and Rocky Duan. ResMimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning. *arXiv preprint arXiv:2510.05070*, 2025.

Shaoting Zhu, Ziwen Zhuang, Mengjie Zhao, Kun-Ying Lee, and Hang Zhao. Hiking in the wild: A scalable perceptive parkour framework for humanoids. In *Conference on Robot Learning*, 2026. To appear. arXiv:2601.07718.