Learning Structured Policies for General Humanoid Loco-Manipulation

Shafeef Omar[1], Sebastian Sanokowski[1], Cordelia Schmid[2,3], and Majid Khadiv[1]

***Abstract*—Building generalist humanoid policies that span diverse whole-body manipulation skills is bottlenecked by exploration: directly searching the high-dimensional action space of a humanoid for long-horizon behaviors is intractable, and tracking-based RL controllers, while reliable executors at high frequency, do not by themselves produce the multi-modal, task-conditioned plans such generality requires. Diffusion models, on the other hand, capture multi-modal motion distributions well but lack any grounding in physical feasibility on a specific embodiment. We exploit this complementarity by placing diffusion at the motion level rather than the action level, structuring exploration over motions while leaving high-frequency control to the tracker: a diffusion generator, pretrained on dynamically feasible retargeted demonstrations, restricts the search space to a learned manifold of plausible motions, while a generalist RL tracking controller realizes them on the robot. To close the residual gap between what the motion generator proposes and what the controller can execute, we further post-train the diffusion model using RL with the tracker rolled out in the loop, finetuning against trackability and kinematic-fidelity rewards. The result is a single pipeline that learns from imperfect kinematic demonstrations and generalizes across loco-manipulation skills, including behaviors unseen at pretraining.**

***Index Terms*—Diffusion Models, Reinforcement Learning, Post-Training, Generalist Humanoid Policies**

I. Introduction

Humanoid robots share their morphology with humans, which provides a natural supervisory signal for learning and substantially reduces the exploration burden inherent in high-degree-of-freedom systems. At the same time, legged platforms must continuously stabilize against gravity, unlike wheeled or fixed-base manipulators. This imposes a hard requirement that is absent from large vision-language-action models such as $\pi_{0.5}$ [1]: a fast low-level controller running at 50–100 Hz or higher must always sit beneath any high-level planner, and large models cannot be directly queried at such control rates. Reinforcement learning (RL) has become the dominant tool for this layer for legged robots, producing controllers that run reliably at 1 kHz or more and remain robust across terrains [2]–[4].

This work was funded in part by SIEMENS AG and the Technical University of Munich – Institute for Advanced Study and by the Huawei–TUM joint laboratory.

[1]Munich Institute of Robotics and Machine Intelligence (MIRMI), Chair of AI Planning in Dynamic Environments, Technical University of Munich (TUM), Munich, Germany. {shafeef.omar, sebastian.sanokowski, majid.khadiv}@tum.de

[2]INRIA, Paris, France. cordelia.schmid@inria.fr

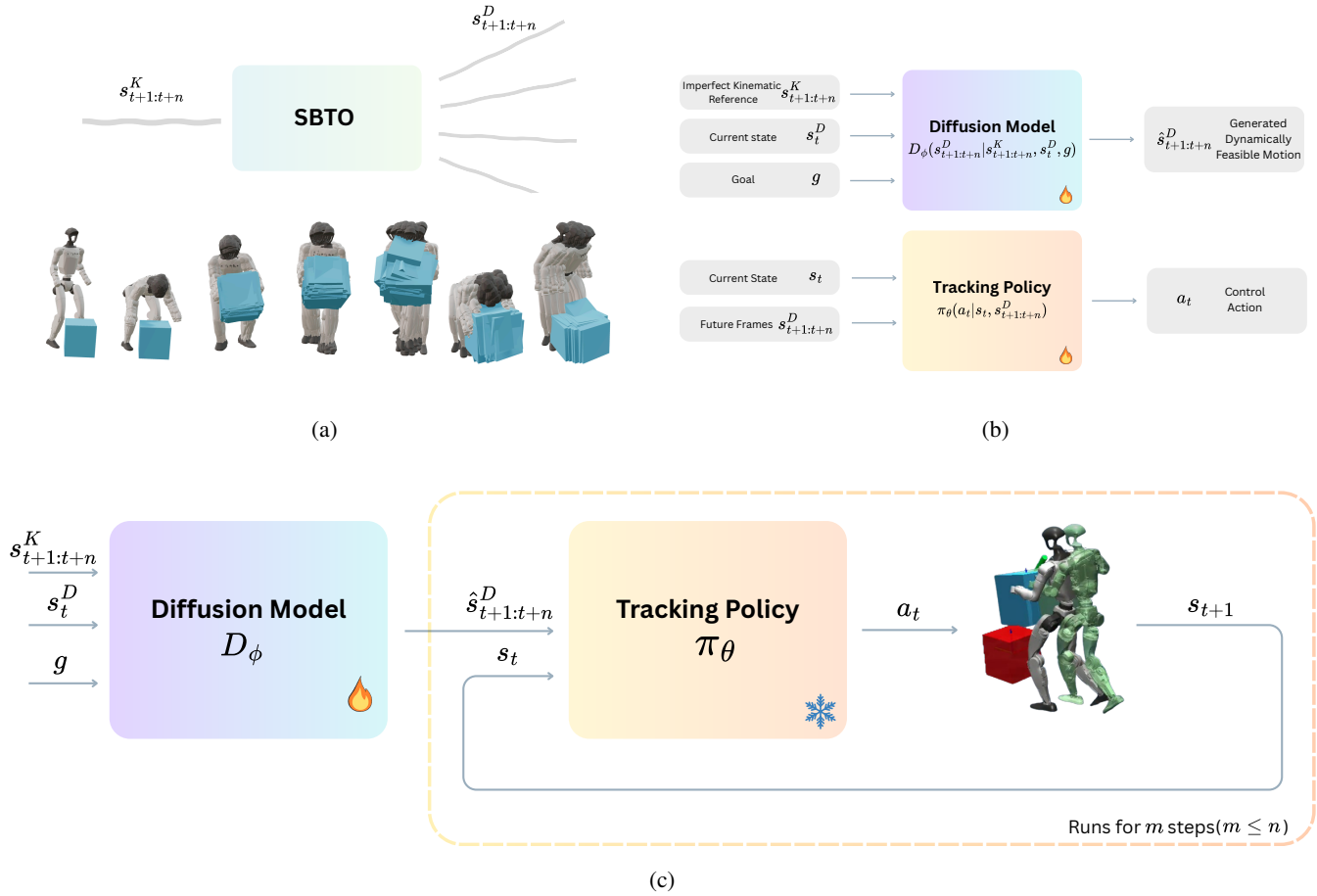
[3]Institute for Advanced Study (IAS), Technical University of Munich, Munich, Germany.

Building on these controllers, RL-based tracking policies have recently made striking progress in mimicking whole-body human motion [4]–[6]. Progress on loco-manipulation, however, has lagged behind pure locomotion. The core obstacle is data: there is no large corpus of physically consistent humanoid loco-manipulation trajectories, and kinematic retargeting from human motion capture alone produces demonstrations that are dynamically infeasible on the robot. Sampling-based optimization methods have emerged as a promising way to close this gap [7], [8]. In particular, Sampling-based Trajectory Optimization (SBTO) [8] can turn a single imperfect kinematic demonstration into a *tube* of dynamically feasible trajectories, yielding both high success rates and strong sample efficiency for downstream tracking controllers (Fig. 1(a)).

Loco-manipulation is inherently multi-modal: the same kinematic demonstration admits many dynamically feasible realizations on the robot, as the SBTO tubes above make explicit. Diffusion models are a natural fit for representing this distribution, and we use one to generate long-horizon motion plans at a low frequency, paired with a high-frequency RL tracking controller for execution. Our goal is a pipeline that starts from a limited set of loco-manipulation demonstrations and generalizes along two axes: it adapts to a pretrained generalist tracking controller, and it accepts new kinematic demonstrations at inference time, including those extracted from in-the-wild videos of humans. A central design choice is how to keep the diffusion generator and the tracker aligned. Drawing on the post-training paradigm that has driven recent progress in large language models, we finetune the generator against a *trackability* reward computed by rolling out the RL tracker in the loop, so the plans it produces are not merely feasible in principle but executable by the specific controller deployed downstream.

Our contributions are:

- A structured generalist pipeline for humanoid loco-manipulation that learns from imperfect kinematic demonstrations and generalizes to new demonstrations at inference time, including those recovered from in-the-wild human video.
- A pretraining data recipe that augments limited demonstrations with SBTO-generated tubes of dynamically feasible trajectories and morphological symmetry.
- A post-training scheme that finetunes the diffusion motion generator with the tracking policy in the loop, closing the plan-execution gap on the deployed controller.



(a) (b) (c)

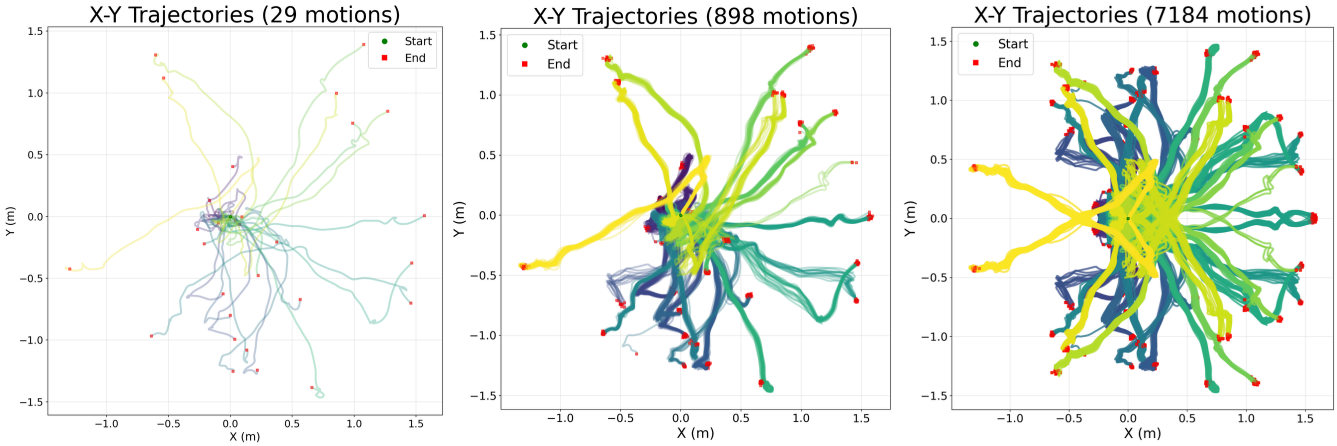
Fig. 1: (a) From a single imperfect reference trajectory, SBTO generates multiple feasible solutions. (b) A diffusion model is pretrained to generate dynamically feasible (SBTO) n step motion, conditioned on imperfect kinematic reference of n steps, current state and goal. A generalist tracking controller is also pretrained separately to generate control actions from dynamically feasible motions. (c) The full pipeline running auto-regressively where the tracking policy runs for $m \leq n$ steps. The diffusion model generates motion for a horizon of n steps, and the tracking policy uses $m \leq n$ steps to generate corresponding control actions sequentially with feedback that are then applied to the robot.

II. Related Work

1) Humanoid motion tracking: DeepMimic [5] established the template of imitating a reference trajectory with a physics-based RL controller. AMP [9] relaxed the per-frame tracking objective into an adversarial style reward, enabling stylized control from unstructured motion data. Subsequent work scaled tracking to large motion corpora with a single policy, most notably PHC [10], and more recently MaskedMimic [11] unified tracking, keyframe, and text-conditioned control under a masked motion-inpainting objective. On real humanoid hardware, HumanPlus [12], OmniH2O [13], ExBody2 [14], BeyondMimic [4], and DynaRetarget [8] have pushed whole-body imitation to physical robots. These systems are strong motion imitators but typically treat the reference as fixed input or the quality of the reference is theirs to cope with, not shape.

2) Motion retargeting: IK-based retargeters remain the default bridge from human motion to robot kinematics, and are convenient but dynamically oblivious [15], [16]. A complementary line implicitly retargets MoCap by absorbing the dynamics gap into the tracking controller itself, e.g. PHC [10] on simulated avatars and HumanPlus [12] / OmniH2O [13] on real humanoids. More recently, explicit physics-aware retargeters refine the reference before tracking [7], [8], producing dynamically feasible motions at higher compute cost. We use SBTO here specifically because of its superior performance.

3) Diffusion generative models for control: Diffusion models capture multi-modal distributions well and have been adopted at multiple levels of the control stack. At the trajectory level, Diffuser [17] frames planning as iterative denoising over full state-action sequences, enabling flexible behavior synthesis under goals and constraints. At the action level, Diffusion Policy [18] established diffusion as a strong behavior-cloning backbone for visuomotor manipulation, and the recipe has since been scaled into generalist vision-language-action models such as π_0 [19] and $\pi_{0.5}$ [1]. In the humanoid setting, BeyondMimic [4] uses guided diffusion for versatile whole-body control. Our use of diffusion is at the trajectory level



(a)

Fig. 2: Visualization of the robot's X-Y trajectories in our dataset to delineate our data augmentation pipeline. Starting from only limited set of 29 demonstrations (left) with all demonstrations aligned at (0, 0), we first generate tube-like diversity of physically consistent motions with SBTO (middle). We then apply symmetry transformations to further augment our dataset (right).

rather than the action level, which lets the tracker stay reusable across behaviors.

4) RL post-training of generative models: Post-training has driven much of the recent progress in large language models: supervised pretraining is followed by an RL phase that aligns the model with human preferences or verifiable rewards, as in InstructGPT [20] and the more recent reasoning-oriented DeepSeek-R1 [21]. The same recipe has been ported to diffusion models. DDPO [22] casts the denoising chain as a multi-step MDP and applies policy gradient to optimize task-specific rewards, demonstrating reward-driven finetuning of text-to-image diffusion. DIME [23] pushes diffusion policies into a maximum-entropy RL formulation for continuous control.

III. Method

Our pipeline consists of four stages: (1) Data Collection, (2) Pretraining Diffusion Trajectory Generator, (3) Pretraining Generalist Tracking, and (4) Post-training Diffusion Trajectory Generator with Tracking Controller in the loop.

A. Data Collection

While teleoperation data is expensive, human motion data is fairly easier to collect and retarget to the humanoid [15]. However, most of the retargeting methods use inverse kinematics (IK) based methods , which are not guaranteed to be physically feasible for the humanoid, due to the inherent difference in the form factor. To tackle this, we further use a physically-consistent sampling based trajectory optimization (SBTO) [8] method to ensure physical consistency. Another benefit of using SBTO for retargeting is the ability to augment our dataset with tube-like diversity as shown Fig. 2. We further augment our dataset with two types of symmetry as shown in Fig. 2. Mirror symmetry for both robot and object and further 90 deg rotations for our object due to the symmetric nature of its shape, effectively 8x the dataset after SBTO.

During this stage, we collect pairs of IK-retargeted ($s^K_{t+1:t+n}$) and SBTO-retargeted ($s^D_{t+1:t+n}$) data, together with the goal (g), which is a vector towards the final object frame with respect to the robot base. These values then go through a preprocessing stage that normalizes and converts the data to relative/global quantities, before passing to the diffusion trajectory generator. We adopt a continuous 6D rotation representation to avoid discontinuities and ambiguities inherent in minimal parameterizations such as Euler angles or quaternions [24].

B. Diffusion Trajectory Generator Pretraining

The retargeted loco-manipulation data from the previous section is used for pretraining by a diffusion-based trajectory generator. We train the diffusion trajectory generator just like a conditional diffusion policy [18], but instead we task the network to learn dynamically-feasible motion from kinematic references instead of directly predicting actions due to the need for the low-level controller to run fast and robust. During the forward noising process, we gradually add noise to the SBTO-retargeted data. And during the reverse denoising process, the diffusion model learns to gradually denoise the SBTO-retargeted data from pure noise, while being conditioned on the IK-retargeted data and the goal.

The simplified loss function used for the diffusion pretraining is given as:

$$\begin{aligned} L(\phi) &= \|\hat{s}^D_{t+1:t+n} - s^D_{t+1:t+n}\|^2 \\ &= \|D_\phi(s^D_t \mid s^K_{t+1:t+n}, g) - s^D_{t+1:t+n}\|^2 \end{aligned} \tag{1}$$

where $\hat{s}^D_{t+1:t+n}$ is the generated motion for n steps by the diffusion model D_ϕ, s^D_t is the current dynamically feasible

TABLE I: Per-episode success and tracking metrics across different motions (n=100 episodes each, mean $\pm$ std)

Dataset	MPJPE (m)	Joint pos. err. (rad)	Joint vel. err. (rad/s)	Object pos. err. (m)	Object ori. err. (rad)	Success rate
Pick & Place (train)	0.069 ± 0.014	0.625 ± 0.060	2.66 ± 0.31	0.067 ± 0.040	0.281 ± 0.286	1.00
Pick & Place (eval)	0.070 ± 0.012	0.619 ± 0.042	2.26 ± 0.19	0.061 ± 0.013	0.529 ± 0.415	1.00
Pushing (eval)	0.117 ± 0.031	0.573 ± 0.040	2.96 ± 0.24	0.247 ± 0.096	0.373 ± 0.134	1.00
Kicking (eval)	0.254 ± 0.088	0.613 ± 0.056	4.48 ± 0.72	0.512 ± 0.112	0.772 ± 0.271	0.74

frame from our dataset, $s^K_{t+1:t+n}$ is the corresponding IK-retargeted motion conditioning for n steps, and g is the final goal conditioning of the object.

C. Generalist Tracking Policy Pretraining

The tracking policy is a DeepMimic-style [4], [5] policy that tracks motion references. During pretraining of the tracking controller, we formulate the problem as an MDP to maximize the cumulative tracking reward as detailed in Table II (Appendix A). Different from [8] which only observes 1 frame for tracking a single motion, our policy observes a sparse lookahead of 1, 2, 4, 8 frames of the desired trajectory to anticipate movements and enable generalisation to multiple motions. We only use the dynamically-feasbile retargeted motion to train our policy since the tracking policy performance heavily relies on the quality of retargeted motion as explicated in [8], [15]. We adopt adaptive sampling as in [4] for each motion to visit more difficult segments of a motion more frequently and randomly assign each env with different motion references to track. Future work also aims to implement adaptive sampling for sampling different motions.

D. Diffusion Trajectory Generator Post-Training

The main goal of the post-training phase is two-fold: (1) adapt the diffusion trajectory generator with the controller in the loop, and (2) adapt the diffusion trajectory generator to new motions unseen during pretraining. In our case, during pretraining, we only used loco-manipulation tasks that involved pickup up an object and placing it at different locations in different styles. Our goal is to adapt this model to further more styles such as kicking the object to desired goals, sliding the object on the floor etc.

To finetune the diffusion model, we use the Diffusion Policy Policy Optimization (DPPO) [25] framework. Even though the diffusion model was purely trained without any actual physical consequence of its output, the finetuning stage should handle the discrepancy. To this end, we introduce two groups of rewards: (1) Trackability rewards, and (2) Kinematic tracking rewards. These are further detailed in Table III (Appendix B).

IV. Results

We report per-episode tracking metrics for the pretrained pipeline (diffusion generator + tracking controller, no post-training) on one held-in split (*Pick & Place train*) and three held-out evaluations spanning a within-distribution split (*Pick & Place eval*) and two unseen skills (*Pushing on floor*, *Kicking*). Results are summarised in Table I.

A. In-distribution evaluation

In the first two rows of Table I, we notice that the robot shows very similar performance in pick and place tasks, even on demonstrations unseen during training, except for a minor gap in the object rotational error. This points to the fact that the tracking policy, together with the diffusion trajectory generator can successfully perform pick and place tasks of different motion styles.

B. Out-of-distribution evaluation

To further push the limits of our pretrained framework, we evaluate them on novel tasks such as pushing the object on the floor and kicking the object multiple times to repose it. As exepected, since neither the tracking policy, nor the trajectory generator have been trained on these motions, we notice a significant drop in performance. In the pushing task, even though the robot executes the motions with 100% success, the tracking is affected. We believe that the 100% success could be attributed to still similar motion style seen during training.When evaluating on a very different skill such as kicking the object, we find an even significant drop in tracking and success.

V. Conclusion and Future Work

We presented a two-stage framework for humanoid loco-manipulation that combines a diffusion trajectory generator, trained on imperfect kinematic demonstrations, with a generalist DeepMimic-style tracking controller. The pretrained pipeline successfully reproduces in-distribution pick-and-place behaviours across motion styles unseen at training time, and degrades gracefully on out-of-distribution skills such as pushing and kicking, where neither the generator nor the controller has been exposed to the underlying motion class.

The most immediate line of future work is the empirical evaluation of the post-training stage described in Section III-D (Table III), for which we have specified the formulation, reward decomposition, and training objective but have not yet reported results. We expect post-training to address the two failure modes most visible in our out-of-distribution evaluation: (1) the trackability gap between the kinematic trajectories produced by the generator and what the controller can physically execute, which the trackability rewards are designed to close; and (2) the inability of a purely kinematically-pretrained generator to extend to skills with substantially different contact dynamics (e.g. kicking, sliding), which the kinematic-fidelity terms together with controller-in-the-loop rollouts should enable. A thorough study would compare the

pretrained-only baseline reported in Table I against the post-trained generator on both in-distribution and held-out skills, and ablate the trackability vs. kinematic-fidelity reward groups to isolate their contributions.

Beyond evaluating the proposed post-training recipe, we see several further directions: (i) extending adaptive sampling from intra-motion segments to inter-motion selection during tracking-policy pretraining, as noted in Section III-C; (ii) scaling the demonstration set to a broader range of loco-manipulation primitives and multi-object scenes; and (iii) exploring parameter-efficient finetuning (e.g. LoRA) of the diffusion generator to enable rapid adaptation to new skills without full retraining.

REFERENCES

[1] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," 2025. [Online]. Available: https://arxiv.org/abs/2504.16054

[2] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter, "Learning agile and dynamic motor skills for legged robots," *Science Robotics*, vol. 4, no. 26, p. eaau5872, 2019. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.aau5872

[3] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science Robotics*, vol. 5, no. 47, p. eabc5986, 2020. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.abc5986

[4] Q. Liao, T. E. Truong, X. Huang, Y. Gao, G. Tevet, K. Sreenath, and C. K. Liu, "Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion," 2025. [Online]. Available: https://arxiv.org/abs/2508.08241

[5] X. B. Peng, P. Abbeel, S. Levine, and M. van de Panne, "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 143:1–143:14, Jul. 2018. [Online]. Available: http://doi.acm.org/10.1145/3197517.3201311

[6] Y. Ze, S. Zhao, W. Wang, A. Kanazawa, R. Duan, P. Abbeel, G. Shi, J. Wu, and C. K. Liu, "Twist2: Scalable, portable, and holistic humanoid data collection system," 2025. [Online]. Available: https://arxiv.org/abs/2511.02832

[7] C. Pan, C. Wang, H. Qi, Z. Liu, H. Bharadhwaj, A. Sharma, T. Wu, G. Shi, J. Malik, and F. Hogan, "Spider: Scalable physics-informed dexterous retargeting," 2025. [Online]. Available: https://arxiv.org/abs/2511.09484

[8] V. Dhedin, I. Taouil, S. Omar, D. Yu, K. Tao, A. Dai, and M. Khadiv, "Dynaretarget: Dynamically-feasible retargeting using sampling-based trajectory optimization," 2026. [Online]. Available: https://arxiv.org/abs/2602.06827

[9] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa, "Amp: Adversarial motion priors for stylized physics-based character control," 2021. [Online]. Available: https://arxiv.org/abs/2104.02180

[10] Z. Luo, J. Cao, A. Winkler, K. Kitani, and W. Xu, "Perpetual humanoid control for real-time simulated avatars," 2023. [Online]. Available: https://arxiv.org/abs/2305.06456

[11] C. Tessler, Y. Guo, O. Nabati, G. Chechik, and X. B. Peng, "Maskedmimic: Unified physics-based character control through masked motion inpainting," 2024. [Online]. Available: https://arxiv.org/abs/2409.14393

[12] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," 2024. [Online]. Available: https://arxiv.org/abs/2406.10454

[13] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. Kitani, C. Liu, and G. Shi, "Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," 2024. [Online]. Available: https://arxiv.org/abs/2406.08858

[14] M. Ji, X. Peng, F. Liu, J. Li, G. Yang, X. Cheng, and X. Wang, "Exbody2: Advanced expressive humanoid whole-body control," 2024. [Online]. Available: https://arxiv.org/abs/2412.13196

[15] J. P. Araujo, Y. Ze, P. Xu, J. Wu, and C. K. Liu, "Retargeting matters: General motion retargeting for humanoid motion tracking," 2025. [Online]. Available: https://arxiv.org/abs/2510.02252

[16] L. Yang, X. Huang, Z. Wu, A. Kanazawa, P. Abbeel, C. Sferrazza, C. K. Liu, R. Duan, and G. Shi, "Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction," 2025. [Online]. Available: https://arxiv.org/abs/2509.26633

[17] M. Janner, Y. Du, J. B. Tenenbaum, and S. Levine, "Planning with diffusion for flexible behavior synthesis," 2022. [Online]. Available: https://arxiv.org/abs/2205.09991

[18] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," 2024. [Online]. Available: https://arxiv.org/abs/2303.04137

[19] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky, "π_0: A vision-language-action flow model for general robot control," 2024. [Online]. Available: https://arxiv.org/abs/2410.24164

[20] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," 2022. [Online]. Available: https://arxiv.org/abs/2203.02155

[21] DeepSeek-AI, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2501.12948

[22] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine, "Training diffusion models with reinforcement learning," 2023. [Online]. Available: https://arxiv.org/abs/2305.13301

[23] O. Celik, Z. Li, D. Blessing, G. Li, D. Palenicek, J. Peters, G. Chalvatzaki, and G. Neumann, "Dime: Diffusion-based maximum entropy reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2502.02316

[24] A. R. Geist, J. Frey, M. Zhobro, A. Levina, and G. Martius, "Learning with 3d rotations, a hitchhiker's guide to so(3)," 2024. [Online]. Available: https://arxiv.org/abs/2404.11735

[25] A. Z. Ren, J. Lidard, L. L. Ankile, A. Simeonov, P. Agrawal, A. Majumdar, B. Burchfiel, H. Dai, and M. Simchowitz, "Diffusion policy policy optimization," in *arXiv preprint arXiv:2409.00588*, 2024.

APPENDIX

A. Tracking Controller Reward Terms

Table II lists the reward terms used to pretrain the generalist tracking controller (Section III-C).

TABLE II: Reward terms used by the generalist tracking controller.

Reward Term	Equation	Weight
Motion Tracking		
Root Position	$\exp\left(-5\lVert \mathbf{p}_t - \hat{\mathbf{p}}_t^{\mathrm{D}} \rVert^2\right)$	0.5
Root Orientation	$\exp\left(-3\lVert \mathbf{q}_t - \hat{\mathbf{q}}_t^{\mathrm{D}} \rVert^2\right)$	0.5
Body Position	$\exp\left(-5\lVert \mathbf{p}_{\mathrm{body},t} - \hat{\mathbf{p}}_{\mathrm{body},t}^{\mathrm{D}} \rVert^2\right)$	1.0
Body Orientation	$\exp\left(-3\lVert \boldsymbol{\theta}_{\mathrm{body},t} - \hat{\boldsymbol{\theta}}_{\mathrm{body},t}^{\mathrm{D}} \rVert^2\right)$	1.0
Body Linear Velocity	$\exp\left(-0.5\lVert \mathbf{v}_t - \hat{\mathbf{v}}_t^{\mathrm{D}} \rVert^2\right)$	1.0
Body Angular Velocity	$\exp\left(-0.05\lVert \boldsymbol{\omega}_t - \hat{\boldsymbol{\omega}}_t^{\mathrm{D}} \rVert^2\right)$	1.0
Joint Pos Tracking	$\exp\left(-5\lVert \boldsymbol{u}_t - \hat{\boldsymbol{u}}_t^{\mathrm{D}} \rVert^2\right)$	2.0
Object Tracking		
Contact Match	$\mathbb{I}(c = \hat{c}^{\mathrm{D}}) \cdot \exp(-0.1\,(\lVert \boldsymbol{F}_t \rVert - 10))$	1.25
Object Position	$\exp\left(-8\lVert \mathbf{p}_{\mathrm{obj},t} - \hat{\mathbf{p}}_{\mathrm{obj},t}^{\mathrm{D}} \rVert^2\right)$	1.0
Object Orientation	$\exp\left(-5\lVert \boldsymbol{\theta}_{\mathrm{obj},t} - \hat{\boldsymbol{\theta}}_{\mathrm{obj},t}^{\mathrm{D}} \rVert^2\right)$	1.0
Object Linear Velocity	$\exp\left(-2\lVert \mathbf{v}_{\mathrm{obj},t} - \hat{\mathbf{v}}_{\mathrm{obj},t}^{\mathrm{D}} \rVert^2\right)$	1.0
Object Angular Velocity	$\exp\left(-0.2\lVert \boldsymbol{\omega}_{\mathrm{obj},t} - \hat{\boldsymbol{\omega}}_{\mathrm{obj},t}^{\mathrm{D}} \rVert^2\right)$	1.0
Regularization		
Action Rate	$-\lVert \mathbf{a}_t - \mathbf{a}_{t-1} \rVert^2$	-0.1
Joint Limit	$-\sum_i \phi(q_i, q_{i,\min}, q_{i,\max})$	-10.0
Self-collisions	$\sum_{c \in \mathcal{C}_{\mathrm{self}}} 1$	-1.0

B. Post-Training Reward Terms

Table III lists the reward terms used to finetune the diffusion trajectory generator (Section III-D).

TABLE III: Post-training reward terms for the diffusion trajectory generator over a chunk of T steps. $(\hat{\cdot})^D$: generated trajectory state; $(\cdot)$: sim state from tracking policy π rolled out on $(\hat{\cdot})^D_{1:t}$; $(\cdot)^K$: kinematic reference state. All squared-norm errors below are averaged over the chunk: $e \equiv \frac{1}{T}\sum_{t=1}^{T} e_t$.

Reward Term	Equation	Weight
Trackability (r_{track}): generated vs. π-executed sim state		
Root Position	$\exp(-5\overline{\lVert \hat{\mathbf{p}}_t^D - \mathbf{p}_t \rVert^2})$	0.50
Root Orientation	$\exp(-3\overline{\lVert \hat{\mathbf{q}}_t^D - \mathbf{q}_t \rVert^2})$	0.50
Body Position	$\exp(-5\overline{\lVert \hat{\mathbf{p}}_{\mathrm{body},t}^D - \mathbf{p}_{\mathrm{body},t} \rVert^2})$	1.00
Body Orientation	$\exp(-3\overline{\lVert \hat{\mathbf{q}}_{\mathrm{body},t}^D - \mathbf{q}_{\mathrm{body},t} \rVert^2})$	1.00
Body Lin. Velocity	$\exp(-0.5\overline{\lVert \hat{\mathbf{v}}_t^D - \mathbf{v}_t \rVert^2})$	1.00
Body Ang. Velocity	$\exp(-0.05\overline{\lVert \hat{\boldsymbol{\omega}}_t^D - \boldsymbol{\omega}_t \rVert^2})$	1.00
Object Position	$\exp(-8\overline{\lVert \hat{\mathbf{p}}_{\mathrm{obj},t}^D - \mathbf{p}_{\mathrm{obj},t} \rVert^2})$	0.50
Object Orientation	$\exp(-5\overline{\lVert \hat{\mathbf{q}}_{\mathrm{obj},t}^D - \mathbf{q}_{\mathrm{obj},t} \rVert^2})$	0.50
Kinematic Fidelity (r_{kin}): generated vs. kinematic reference		
Anchor Position	$\exp(-5\overline{\lVert \hat{\mathbf{p}}_t^D - \mathbf{p}_t^K \rVert^2})$	0.25
Anchor Orientation	$\exp(-3\overline{\lVert \hat{\mathbf{q}}_t^D - \mathbf{q}_t^K \rVert^2})$	0.25
Body Position	$\exp(-5\overline{\lVert \hat{\mathbf{p}}_{\mathrm{body},t}^D - \mathbf{p}_{\mathrm{body},t}^K \rVert^2})$	0.50
Body Orientation	$\exp(-3\overline{\lVert \hat{\mathbf{q}}_{\mathrm{body},t}^D - \mathbf{q}_{\mathrm{body},t}^K \rVert^2})$	0.50
Body Lin. Velocity	$\exp(-0.5\overline{\lVert \hat{\mathbf{v}}_t^D - \mathbf{v}_t^K \rVert^2})$	0.50
Body Ang. Velocity	$\exp(-0.05\overline{\lVert \hat{\boldsymbol{\omega}}_t^D - \boldsymbol{\omega}_t^K \rVert^2})$	0.50
Object Position	$\exp(-8\overline{\lVert \hat{\mathbf{p}}_{\mathrm{obj},t}^D - \mathbf{p}_{\mathrm{obj},t}^K \rVert^2})$	0.25
Object Orientation	$\exp(-5\overline{\lVert \hat{\mathbf{q}}_{\mathrm{obj},t}^D - \mathbf{q}_{\mathrm{obj},t}^K \rVert^2})$	0.25
Smoothness (p_{smooth})		
Joint Position Rate	$\overline{\lVert \hat{\mathbf{q}}_{t+1}^D - \hat{\mathbf{q}}_t^D \rVert^2}$	-0.50
Object Position Rate	$\overline{\lVert \hat{\mathbf{p}}_{\mathrm{obj},t+1}^D - \hat{\mathbf{p}}_{\mathrm{obj},t}^D \rVert^2}$	-0.50
Object Orientation Rate	$\overline{\lVert \hat{\mathbf{q}}_{\mathrm{obj},t}^D - \hat{\mathbf{q}}_{\mathrm{obj},t+1}^D \rVert^2}$	-0.50
Termination Penalty		
Termination	$\mathbf{1}[\text{terminated during chunk}]$	-5.0