

# Do You Have Freestyle? Expressive Humanoid Locomotion via Audio Control

## Supplementary Material

### Appendix Overview

This appendix provides additional details and results, organized as follows:

- **Section 1:** Elaboration on some details during training, including dataset details, motion filter and retargeting, simulator, domain randomization, regularization, reward functions, curriculum learning, and adaptive sigma.
- **Section 2:** Details about evaluation, including metrics about motion tracking and motion-audio alignment.
- **Section 3:** Additional experiments, including audio-motion alignment evaluation, ablation studies on  $\Delta\text{MoE}$  and diffusion policy.
- **Section 4:** Extra qualitative experiment results and visualizations, including in the simulation and in the real-world.

### 1. Implementation Details

This section details the state representation for policy training, including proprioceptive states, privileged information, and network hyperparameters. As summarized in Table 1, the proprioceptive state components are shared between the teacher and student policies, with a critical distinction: the student policy leverages an extended observation history to compensate for the absence of privileged information, substituting temporal context for direct auxiliary signals.

Our proprioceptive information includes joint positions, joint velocities, root angular velocity, root projected gravity, and the aforementioned information from four historical frames, which is elaborated in Table 1. For privileged information, it forms the observation of the critic network together with proprioceptive information. Unlike prior works where both teacher and student policies receive explicit reference motion as part of observations, our framework restricts these target signals exclusively to the teacher. By contrast, the student policy additionally takes proprioceptive states from 25 historical frames, motion latents for content representation, and audio latents for style representation as inputs. The audio latents first feed into a pretrained adaptor to infuse kinematic information. Full details of the target state are provided in Table 2. Both policies output 23-dimensional target joint positions.

The teacher policy is trained via PPO [14], taking privileged information, motion tracking targets, and proprioceptive states as inputs, which are concatenated and processed by  $\Delta\text{MoE}$ . The first expert takes all zeros as conditions to predict action  $\mathbf{a}_1$ . The second expert only receives proprioceptive states as conditions, with all remaining positions filled with zeros. This pattern continues such that the fourth expert accepts all conditions to output action  $\mathbf{a}_4$ . The final

action is obtained through a weighted sum of the outputs of all experts, where the weights  $w_i$  are generated by a gating network. The student policy is trained with DAgger, lacking access to privileged information and explicit reference motion, instead relying on extended observation histories and audio latent representations to enable a retargeting-free, audio latent-driven pipeline. First, audio features are extracted from the input audio. Then, our pretrained adaptor is utilized to infuse kinematic information into the audio features, enabling them to guide humanoid action generation more effectively. The inputs of student policy are concatenated and fed as conditions to a diffusion model with an MLP backbone, where AdaLN injects conditional signals throughout the denoising process. A final MLP layer projects the backbone output to the 23-dimensional action space, with conditional signals further integrated for alignment. Detailed hyperparameters for both policies are listed in Table 3.

**Motion Filter and Retargeting** Following [15], we quantify stability by computing the ground-projected distance between the center of mass (CoM) and center of pressure (CoP) for each frame, with a predefined stability threshold. Let  $\bar{\mathbf{p}}_t^{\text{CoM}} = (p_{t,x}^{\text{CoM}}, p_{t,y}^{\text{CoM}})$  and  $\bar{\mathbf{p}}_t^{\text{CoP}} = (p_{t,x}^{\text{CoP}}, p_{t,y}^{\text{CoP}})$  represent the 2D ground projections of CoM and CoP at frame  $t$ , respectively. We define  $\Delta d_t = \|\bar{\mathbf{p}}_t^{\text{CoM}} - \bar{\mathbf{p}}_t^{\text{CoP}}\|_2$  as this distance. A frame is considered stable if  $\Delta d_t < \epsilon_{\text{stab}}$ . A motion sequence is retained if its first and last frames are stable, and the longest consecutive unstable segment has fewer than 100 frames.

**Simulator** Following established protocols in motion tracking policy research [4, 6, 9], we adopt a three-stage evaluation pipeline: first, large-scale reinforcement learning training in IsaacGym; second, zero-shot transfer to MuJoCo to assess cross-simulator generalization; third, physical deployment on the Unitree G1 humanoid platform to validate real-world performance.

**Reference State Initialization** Task initialization is critical for reinforcement learning (RL) training. We observe that naively initializing episodes at the start of reference motions often leads to policy failure, especially for complex motions. This can cause the environment to overfit to simpler frames, neglecting the most challenging motion segments.

To address this, we adopt the Reference State Initialization (RSI) framework [13]. Specifically, we uniformly sample time-phase variables over  $[0,1]$  to randomize the starting point within the reference motion that the policy must track.

Proprioceptive States	
State Component	Dim.
DoF position	$23 \times (1+4)$
DoF velocity	$23 \times (1+4)$
Last action	$23 \times (1+4)$
Root angular velocity	$3 \times (1+4)$
Projected gravity	$3 \times (1+4)$
Total dim	$75 \times 5$
Privileged Information	
Root linear velocity	$3 \times (1+4)$
Reference body position	81
Body position difference	81
Randomized base CoM offset	3
Randomized link mass	22
Randomized stiffness	23
Randomized damping	23
Randomized friction coefficient	1
Randomized control delay	1
Total dim	250

Table 1. Proprioceptive states and privileged information.

Hyperparameter	Value
Optimizer	Adam
$\beta_1, \beta_2$	0.9, 0.999
Learning Rate	$1 \times 10^{-3}$
Batch Size	8192
Teacher Policy	
GAE Discount factor ( $\gamma$ )	0.99
GAE Decay factor ( $\gamma$ )	0.95
Clip Parameter	0.2
Entropy Coefficient	0.01
Max Gradient Norm	1
Learning Epochs	5
Mini Batches	4
Value Loss Coefficient	1.0
Value MLP Size	[512, 256, 128]
Actor MLP Size	[768, 512, 128]
Experts	4
Student Policy	
MLP Layers	4 + 1 (final layer)
MLP Size	[1792, 1792, 1792, 23]

Table 3. Hyperparameters for teacher and student policy training.

The robot’s state, including root position, orientation, linear and angular velocities, and joint positions and velocities,

Teacher Policy	
State Component	Dim.
Proprioceptive states	$75 \times 5$
DoF position	23
Keypoint position	81
Root Velocity	3
Root Angular Velocity	3
Root Orientation	3
Total dim	489
Student Policy	
Motion Latent	64
Audio Latent	256
Proprioceptive States	$75 \times (25+1)$
Total dim	2270

Table 2. Reference information in the teacher and student policies.

is then initialized to the reference motion’s values at the sampled phase. This approach enhances motion tracking performance, particularly for highly dynamic whole-body motions, by enabling the policy to learn diverse movement segments in parallel rather than being constrained to strictly sequential learning.

**Domain Randomization and Regularization** To improve the robustness and generalization of the pretrained policy, we utilize the domain randomization techniques and regularization items, which are listed in Table 4.

Term	Value
Dynamics Randomization	
Friction	$\mathcal{U}(0.2, 1.5)$
PD gain	$\mathcal{U}(0.75, 1.25)$
Link mass (kg)	$\mathcal{U}(0.9, 1.1) \times \text{default}$
Ankle inertia ( $\text{kg} \cdot \text{m}^2$ )	$\mathcal{U}(0.9, 1.1) \times \text{default}$
Base CoM offset (m)	$\mathcal{U}(-0.05, 0.05)$
ERFI [2] ( $\text{N} \cdot \text{m}/\text{kg}$ )	$0.05 \times \text{torque limit}$
Control delay (ms)	$\mathcal{U}(0, 40)$
External Perturbation	
Random push interval (s)	[5, 10]
Random push velocity (m/s)	0.5

Table 4. Domain randomization settings.

**Motion Tracking Rewards** As shown in Table 5, we define the reward function as the sum of task rewards and regularization, which are meticulously designed to improve both the performance and motion realism of the humanoid robot. Following [15], we enforce penalties for joint positions exceeding soft limits, which are symmetrically derived from hard limits via a fixed scaling ratio ( $\alpha = 0.95$ ). Specifically, the midpoint  $m$  and range  $d$  of hard limits are first computed as:

$$m = \frac{q_{\min} + q_{\max}}{2}, \quad (1)$$

$$d = q_{\max} - q_{\min}, \quad (2)$$

where  $q_{\min}$  and  $q_{\max}$  denote the hard limits of joint position  $q$ . The soft limits are then determined by:

$$q_{\text{soft-min}} = m - 0.5 \cdot d \cdot \alpha, \quad (3)$$

$$q_{\text{soft-max}} = m + 0.5 \cdot d \cdot \alpha. \quad (4)$$

This computation extends to joint velocity  $\dot{q}$  and torque  $\tau$  for their respective soft limits.

Category	Term	Expression & Weight
Reward	Joint position	$\exp\left(-\frac{\ \mathbf{q}_t - \hat{\mathbf{q}}_t\ _2^2}{\sigma_{\text{pos}}^2}\right), 1.0$
	Joint velocity	$\exp\left(-\frac{\ \dot{\mathbf{q}}_t - \hat{\dot{\mathbf{q}}}_t\ _2^2}{\sigma_{\text{vel}}^2}\right), 1.0$
	Body position	$\exp\left(-\frac{\ \mathbf{p}_t - \hat{\mathbf{p}}_t\ _2^2}{\sigma_{\text{pos}}^2}\right), 1.0$
	Body rotation	$\exp\left(-\frac{\ \theta_t - \hat{\theta}_t\ _2^2}{\sigma_{\text{rot}}^2}\right), 0.5$
	Body velocity	$\exp\left(-\frac{\ \mathbf{v}_t - \hat{\mathbf{v}}_t\ _2^2}{\sigma_{\text{vel}}^2}\right), 0.5$
	Body angular velocity	$\exp\left(-\frac{\ \omega_t - \hat{\omega}_t\ _2^2}{\sigma_{\text{ang}}^2}\right), 0.5$
	Body position VR 3 points	$\exp\left(-\frac{\ \mathbf{p}_{\text{vr},t} - \hat{\mathbf{p}}_{\text{vr},t}\ _2^2}{\sigma_{\text{pos, vr}}^2}\right), 1.6$
	Body position feet	$\exp\left(-\frac{\ \mathbf{p}_{\text{feet},t} - \hat{\mathbf{p}}_{\text{feet},t}\ _2^2}{\sigma_{\text{pos, feet}}^2}\right), 1.0$
	Max Joint position	$\exp\left(-\frac{\ \mathbf{q}_t - \hat{\mathbf{q}}_t\ _{\infty}}{\sigma_{\text{max, jpos}}}\right), 1.0$
	Contact Mask	$1 - \frac{\ \mathbf{c}_t - \hat{\mathbf{c}}_t\ _1}{2}, 0.5$
Regularization	Joint position limits	$\mathbb{I}(\mathbf{q} \notin [\mathbf{q}_{\text{soft-min}}, \mathbf{q}_{\text{soft-max}}]), -10.0$
	Joint velocity limits	$\mathbb{I}(\dot{\mathbf{q}} \notin [\dot{\mathbf{q}}_{\text{soft-min}}, \dot{\mathbf{q}}_{\text{soft-max}}]), -5.0$
	Joint torque limits	$\mathbb{I}(\tau \notin [\tau_{\text{soft-min}}, \tau_{\text{soft-max}}]), -5.0$
	Slippage	$\ \mathbf{v}_{\text{feet},xy}\ _2^2 \cdot \mathbb{I}(\ \mathbf{F}_{\text{feet}}\ _2 \geq 1), -1.0$
	Feet contact forces	$\min(\ \mathbf{F}_{\text{feet}} - 400\ _2^2, 0), -0.01$
	Feet air time	$\mathbb{I}[T_{\text{air}} > 0.3], -1.0$
	Stumble	$\mathbb{I}(\ \mathbf{F}_{\text{feet},xy}\  > 5 \cdot \mathbf{F}_{\text{feet},z}), -2.0$
	Torque	$\ \tau\ _2^2, -10^{-6}$
	Action rate	$\ \mathbf{a}_t - \mathbf{a}_{t-1}\ _2^2, -0.02$
	Collision	$\mathbb{I}_{\text{collision}}, -30$
	Termination	$\mathbb{I}_{\text{termination}}, -200$

Table 5. Reward terms and weights.

**Curriculum Learning** To imitate highly dynamic motions, we follow [15], introduce two curriculum mechanisms: a termination curriculum that gradually reduces tracking error tolerance, and a penalty curriculum that progressively increases the weight of regularization terms to promote more stable and physically plausible behaviors.

- **Termination Curriculum:** The episode is terminated early when the humanoid’s motion deviates from the reference beyond a termination threshold  $\theta$ . During training, this threshold is gradually decreased to increase the difficulty:

$$\theta \leftarrow \text{clip}(\theta \cdot (1 - \delta), \theta_{\min}, \theta_{\max}), \quad (5)$$

where the initial threshold  $\theta = 1.5$ , with bounds  $\theta_{\min} = 0.3$ ,  $\theta_{\max} = 2.0$ , and decay rate  $\delta = 2.5 \times 10^{-5}$ .

- **Penalty Curriculum:** To facilitate learning in the early training stages while gradually enforcing stronger regularization, we introduce a scaling factor  $\alpha$  that increases progressively to modulate the influence of the penalty term:

$$\alpha \leftarrow \text{clip}(\alpha \cdot (1 + \delta), \alpha_{\min}, \alpha_{\max}), \quad \hat{r}_{\text{penalty}} \leftarrow \alpha \cdot r_{\text{penalty}}, \quad (6)$$

where the initial penalty scale  $\alpha = 0.1$ , with bounds  $\alpha_{\min} = 0.0$ ,  $\alpha_{\max} = 1.0$ , and growth rate  $\delta = 1.0 \times 10^{-4}$ .

**Adaptive Sigma** Inspired by [15], we employ adaptive sigma in the reward function. Task-specific rewards enforce alignment of joint states, rigid body states, and foot contact masks. All except the foot contact term adopt a bounded exponential form:

$$A = \exp\left(-\frac{x^2}{\sigma^2}\right),$$

where  $x$  denotes tracking error and  $\sigma$  controls error tolerance. This form outperforms negative error terms by stabilizing training and simplifying reward weighting.

## 2. Evaluation Details

**Motion Tracking Metrics** For motion tracking evaluation, we employ metrics standard in prior work [6, 8]: Success Rate (Succ), Mean Per Joint Position Error ( $E_{\text{MPJPE}}$ ), and Mean Per Keybody Position Error ( $E_{\text{MPKPE}}$ ).

- **Success Rate (Succ):** Evaluates whether the humanoid successfully follows the reference motion without falling. A trial fails if the average trajectory deviation exceeds 0.5 meters at any point, or if the root pitch angle exceeds a predefined threshold.
- **Mean Per Joint Position Error ( $E_{\text{MPJPE}}$ , in rad):** Quantifies joint-level tracking accuracy via the average error in degree-of-freedom (DoF) rotations between reference and generated motions.
- **Mean Per Keybody Position Error ( $E_{\text{MPKPE}}$ , in m):** Assesses keypoint tracking performance using the average positional discrepancy between reference and generated keypoint trajectories.

**Motion-Audio Alignment Metrics** We evaluate our audio adaptor using motion-audio alignment metrics: retrieval accuracy (R@1, R@2, R@3), Multimodal Distance (MMDist), and Beat Alignment Score (BAS) [7].

- **Retrieval Accuracy (R-Precision):** These metrics measure the relevance of audio to corresponding motion in a retrieval setup. R@1 denotes the fraction of audio queries for which the correct motion is retrieved as the top match, reflecting the model’s precision in identifying the most relevant motion. R@2 and R@3 extend this notion, indicating recall within the top two and three retrieved motions, respectively.
- **Multimodal Distance (MMDist):** This quantifies the average feature-space distance between audios and their corresponding motions, typically extracted via a pretrained retrieval model. Smaller MMDist values indicate stronger semantic alignment between audio and motion.
- **Beat Alignment Score (BAS):** This metric evaluates the temporal alignment quality between kinematic beats and music beats. Audio beats are detected from audio signals using Librosa [12], yielding a timestamp sequence  $B_y = \{t_y^j\}$  where  $t_y^j$  denotes the time of the  $j$ -th music beat. Kinematic beats are identified as the local minima of the motion’s kinetic velocity, capturing the key rhythmic frames of the motion sequence, resulting in a timestamp sequence  $B_x = \{t_x^i\}$  where  $t_x^i$  denotes the time of the  $i$ -th kinematic beat. The BAS metric is defined as the average of exponential-weighted distances between each kinematic beat and its nearest music beat. This exponential formulation emphasizes closer alignments while mitigating the impact of large discrepancies, and it is normalized via a parameter  $\sigma$  to adapt to sequences with fixed FPS. The formal definition is:

$$\text{BAS} = \frac{1}{m} \sum_{i=1}^m \exp \left( -\frac{\min_{t_y^j \in B_y} \|t_x^i - t_y^j\|^2}{2\sigma^2} \right), \quad (7)$$

where  $m$  is the number of kinematic beats in  $B_x$ . Consistent with our experimental setup (30 FPS), we fix  $\sigma = 3$  across all evaluations.

## 2.1. Deployment Details

**Sim-to-Sim Transfer** As noted in Humanoid-Gym [3], MuJoCo delivers more realistic dynamics than Isaac Gym. Aligning with standard protocols in motion tracking policy research [6, 10], we conduct reinforcement learning training in Isaac Gym to capitalize on its high computational efficiency. To evaluate policy robustness and generalization capability, we perform zero-shot transfer to the MuJoCo simulator. This sim-to-sim transfer serves as an intermediate validation step before deploying the policy on a physical humanoid robot to verify the real-world motion tracking efficacy of our framework.

Method	IsaacGym			MuJoCo			BAS $\uparrow$
	Succ $\uparrow$	$E_{mpjpe}$ $\downarrow$	$E_{mpkpe}$ $\downarrow$	Succ $\uparrow$	$E_{mpjpe}$ $\downarrow$	$E_{mpkpe}$ $\downarrow$	
BEAT2							
Baseline	0.98	0.08	0.06	0.94	0.16	0.14	0.163
Ours	0.99	0.05	0.04	0.96	0.10	0.09	0.197
FineDance							
Baseline	0.86	0.26	0.23	0.58	0.35	0.32	0.176
Ours	0.93	0.18	0.16	0.67	0.26	0.24	0.214

Table 6. Ablation study on whether to use adaptor to inject kinematic information into the audio modality. It can be observed that the adaptor successfully aligns the audio and motion, improving the tracking performance and success rate.

**Sim-to-Real Deployment** Real-world experiments are conducted on a Unitree G1 humanoid robot, integrated with an onboard Jetson Orin NX module for computation and communication. The control policy processes motion tracking targets to generate target joint positions, then transmits control commands to the robot’s low-level controller at 50Hz, with a communication latency of 18–30ms. The low-level controller operates at 500Hz to guarantee stable real-time actuation. Communication between the high-level policy and low-level interface is implemented via Lightweight Communications and Marshalling (LCM) [5].

## 3. Additional Experiments

**Audio-Motion Alignment** To evaluate whether the co-speech gestures or dance motions generated by the robot adhere to rhythmic patterns, we compute the BAS for successful cases in the test set. Specifically, we retrieve the joint velocity of the robot’s motors and calculate the BAS value by correlating it with music beats. The results are presented in Table 6, where the Baseline corresponds to the outcome of concatenating music latents with other observations and motion latents as inputs to the student policy. It can be observed that when music is treated as an external condition to further modulate the content, the generated actions exhibit superior rhythmic alignment.

**Denosing Steps in Student Policy** We evaluate DDIM sampling with different denoising steps, measuring average per-action step time. Table 7 shows that increasing steps leads to higher latency, which is critical for real-world humanoid robot deployment as latency degrades execution outcomes.

**Noise Scale in Student Policy** We ablate the noise scale  $\beta_{\max}$  for DDIM sampling to study its impact on performance and latency. Table 8 shows that  $\beta_{\max} = 0.20$  achieves optimal success rate.

Method	Avg Time (s) $\times 10^{-3}$
DDIM-2 sampling	5.3
DDIM-4 sampling	11.6
DDIM-6 sampling	13.4
DDIM-8 sampling	17.6
DDIM-10 sampling	18.9

Table 7. Average inference time across DDIM sampling steps.

Noise Scale ( $\beta_{\max}$ )	Denoising Steps	Success Rate (%)
0.10	2	92.0
0.15	2	92.0
0.20	2	93.0
0.25	2	91.0
0.30	2	91.0

Note: Fixed settings: cosine noise schedule, DDIM sampling ( $\eta = 0$ ),  $\beta_{\max}$  denotes the maximum  $\beta_t$  over 50 training timesteps.

Table 8. Fine-grained ablation on noise scale.

Sampling Strategy	Denoising Steps	Success Rate (%)	Latency (s $\times 10^{-3}$ )
DDIM ( $\eta = 0$ )	2	93.0	5.3
DDIM ( $\eta = 0.5$ )	2	86.0	5.3
DDPM (Stochastic)	2	65.0	8.6

Note: Fixed settings: cosine noise schedule,  $\beta_{\max} = 0.20$ ,  $\eta$  controls DDIM stochasticity.

Table 9. Fine-grained ablation on sampling strategies in the FineDance dataset.

Method	IsaacGym			MuJoCo		
	Succ $\uparrow$	$E_{mpjpe}$ $\downarrow$	$E_{mpkpe}$ $\downarrow$	Succ $\uparrow$	$E_{mpjpe}$ $\downarrow$	$E_{mpkpe}$ $\downarrow$
$\epsilon$ -prediction	0.72	0.46	0.43	0.49	0.58	0.56
$x_0$ -prediction	0.93	0.18	0.16	0.67	0.26	0.24

Table 10. Tracking performance across optimization objectives in the FineDance dataset.

**Noise Schedule Strategies in Student Policy** We compare three sampling strategies: DDIM ( $\eta = 0$ , deterministic), DDIM ( $\eta = 0.5$ , semi-stochastic), and DDPM (stochastic). Table 9 shows that deterministic DDIM achieves the highest success rate and lowest latency. Stochastic strategies reduce performance and increase latency.

**Optimization Objective in Student Policy** We ablate two supervision targets for the diffusion policy:  $\epsilon$ -prediction and  $x_0$ -prediction. Table 10 shows that  $x_0$ -prediction achieves significantly better tracking performance compared to  $\epsilon$ -prediction.

**Experts Number in  $\Delta$ MoE** We conduct ablation experiments on the number of experts in our  $\Delta$ MoE. Since the number of experts in  $\Delta$ MoE determines the dimensionality

$N$	IsaacGym			MuJoCo		
	Succ $\uparrow$	$E_{mpjpe}$ $\downarrow$	$E_{mpkpe}$ $\downarrow$	Succ $\uparrow$	$E_{mpjpe}$ $\downarrow$	$E_{mpkpe}$ $\downarrow$
3	0.90	0.23	0.21	0.63	0.30	0.28
4	0.93	0.18	0.16	0.67	0.26	0.24
5	0.91	0.22	0.18	0.66	0.30	0.27
6	0.92	0.21	0.18	0.67	0.27	0.24

Table 11. Tracking performance across different numbers of experts in the FineDance dataset.

Method	IsaacGym			MuJoCo		
	Succ $\uparrow$	$E_{mpjpe}$ $\downarrow$	$E_{mpkpe}$ $\downarrow$	Succ $\uparrow$	$E_{mpjpe}$ $\downarrow$	$E_{mpkpe}$ $\downarrow$
Random	0.93	0.19	0.16	0.67	0.26	0.25
Ours	0.93	0.18	0.16	0.67	0.26	0.24

Table 12. Ablation study on the impact of different condition space partitioning methods on tracking performance in the FineDance Dataset.

of the condition space, we split the condition into  $N - 1$  partitions when training  $\Delta$ MoE with different  $N$  experts. A critical constraint is that each condition partition  $c_i$  must contain complete information. For instance, the dof positions in proprioceptive states must not be split in both  $c_1$  and  $c_2$ .

As shown in Table 11, the optimal performance is achieved when the number of experts is set to 4. Furthermore, we verify that with a fixed number of experts, the partitioning of conditions has a negligible impact on the results, which is presented in Table 12.

## Evaluation on musical understanding and freestyle adaptation

We conduct experiments to evaluate BAS on unseen music and slowed-down music. As shown in Figure 1, we randomly tested five songs and reported the results. BAS of normal music is 0.226, slowed-down is 0.212, and freestyle is 0.193. Interestingly, we find that when the music is slowed down, the robot’s motion amplitude also decreases.

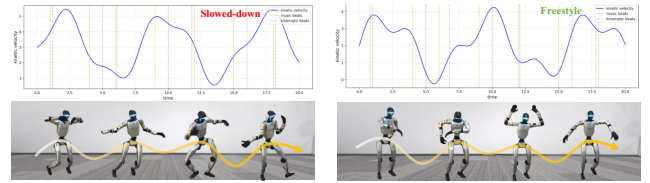


Figure 1. Dancing to slowed-down and freestyle music. Green dash denotes kinematic beats, and orange denotes music beats.

## 4. Qualitative Results

**Simulation** To validate the advantages of the diffusion policy in such conditional control tasks, we visualize two cases in simulation. As shown in the upper part of Figure 2,



the MLP policy exhibits poor tracking performance. In contrast, the diffusion policy achieves superior tracking results by leveraging its enhanced robustness and ability to model distributions.

Furthermore, we verify the freestyle capability of our policy. As illustrated in the lower part of Figure 2, when fed with a piece of music unseen during training to generate actions, the diffusion policy successfully completes the entire motion sequence due to its strong generalization ability, whereas the MLP policy immediately results in a fall.

**Retargeting Method** When training the teacher oracle policy, we investigate diverse retargeting approaches, encompassing PHC [11] and GMR [1]. While GMR demonstrates robust performance in mitigating motion penetration, it gives rise to abrupt motion transitions, as visualized in Figure 3. Thus, we ultimately select PHC as the designated retargeting method for subsequent experimental evaluations. The related video can be found in the supplementary material.

**Real-World** We present real-world deployment for music-to-locomotion and speech-to-locomotion tasks, as shown in Figures 4, 5, and 6. A supplementary video showcasing real-robot deployments is provided in the supplementary material.

## References

- [1] 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. 6
- [2] Luigi Campanaro, Siddhant Gangapurwala, Wolfgang Merkt, and Ioannis Havoutis. Learning and deploying robust locomotion policies with minimal dynamics randomization. In *6th Annual Learning for Dynamics & Control Conference*, pages 578–590. PMLR, 2024. 2
- [3] Xinyang Gu, Yen-Jen Wang, and Jianyu Chen. Humanoid-gym: Reinforcement learning for humanoid robot with zero-shot sim2real transfer. *arXiv preprint arXiv:2404.05695*, 2024. 4
- [4] 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. 1
- [5] Albert S Huang, Edwin Olson, and David C Moore. Lcm: Lightweight communications and marshalling. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4057–4062. IEEE, 2010. 4
- [6] Mazeyu Ji, Xuanbin Peng, Fangchen Liu, Jialong Li, Ge Yang, Xuxin Cheng, and Xiaolong Wang. Exbody2: Advanced expressive humanoid whole-body control. *arXiv preprint arXiv:2412.13196*, 2024. 1, 3, 4
- [7] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++, 2021. 4
- [8] Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Pengju An, Xiaoli Li, Kaichen Zhou, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation. *Advances in Neural Information Processing Systems*, 37:40085–40110, 2024. 3
- [9] Mengzhen Liu, Mengyu Wang, Henghui Ding, Yilong Xu, Yao Zhao, and Yunchao Wei. Segment anything with precise interaction. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3790–3799, 2024. 1
- [10] Mengzhen Liu, Enshen Zhou, Cheng Chi, Yi Han, Shanyu Rong, Liming Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Sapave: Towards active perception and manipulation in vision-language-action models for robotics. *arXiv preprint arXiv:2603.12193*, 2026. 4
- [11] Zhengyi Luo, Jinkun Cao, Alexander W. Winkler, Kris Kitani, and Weipeng Xu. Perpetual humanoid control for real-time simulated avatars. In *International Conference on Computer Vision (ICCV)*, 2023. 6
- [12] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. *SciPy*, 2015: 18–24, 2015. 4
- [13] 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. 1
- [14] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 1
- [15] Weiji Xie, Jinrui Han, Jiakun Zheng, Huanyu Li, Xinzhe Liu, Jiyuan Shi, Weinan Zhang, Chenjia Bai, and Xuelong Li. Kungfubot: Physics-based humanoid whole-body control for learning highly-dynamic skills. *arXiv preprint arXiv:2506.12851*, 2025. 1, 3

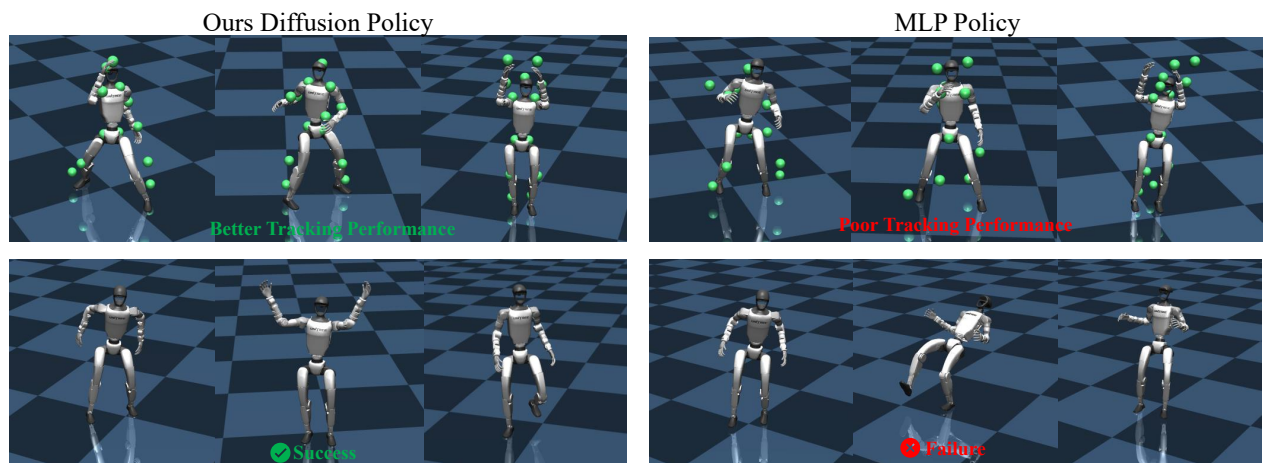


Figure 2. Qualitative results in the MuJoCo. The upper half presents the tracking performance of the MLP policy and the diffusion policy on the same motion; the lower half demonstrates their respective freestyle capabilities when confronted with unseen music.

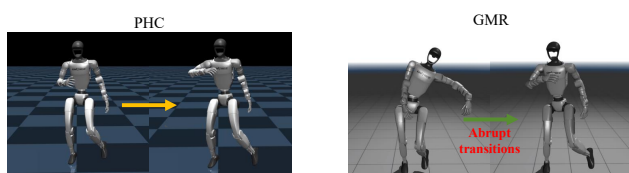
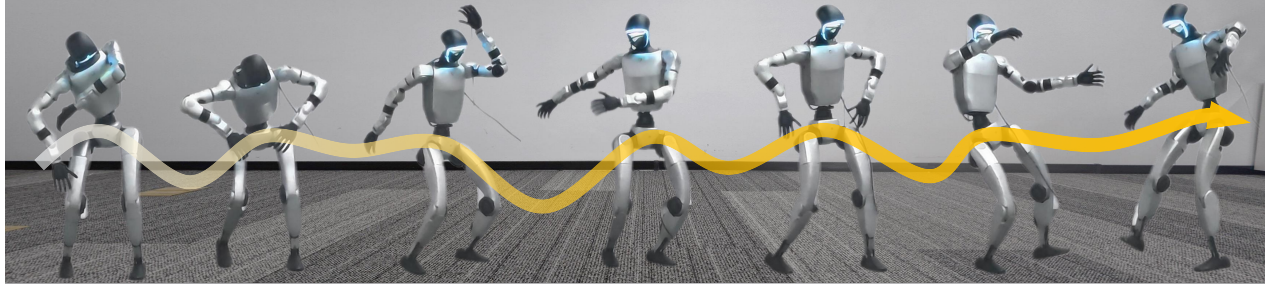
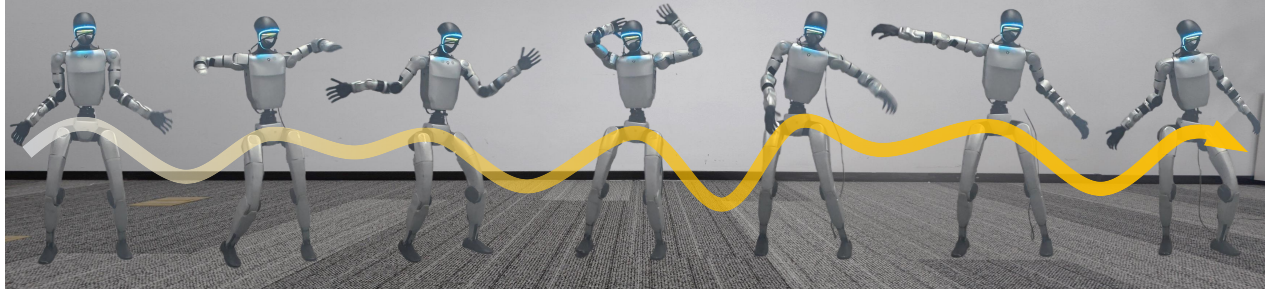


Figure 3. Qualitative results of PHC and GMR retargeting.

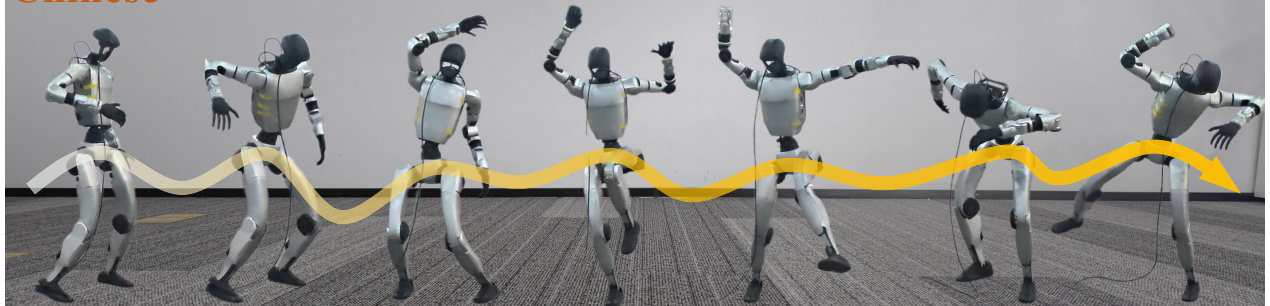
**Korean**



**Jazz**



**Chinese**



**Folk**

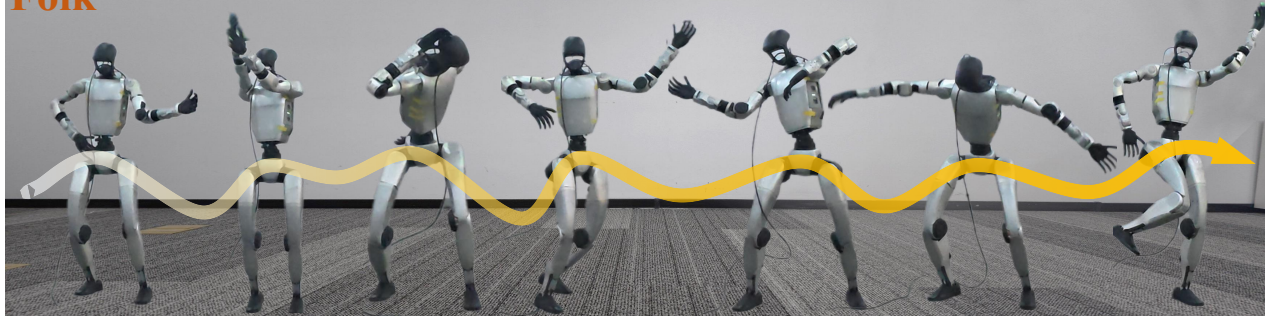


Figure 4. Real-world music-to-locomotion.



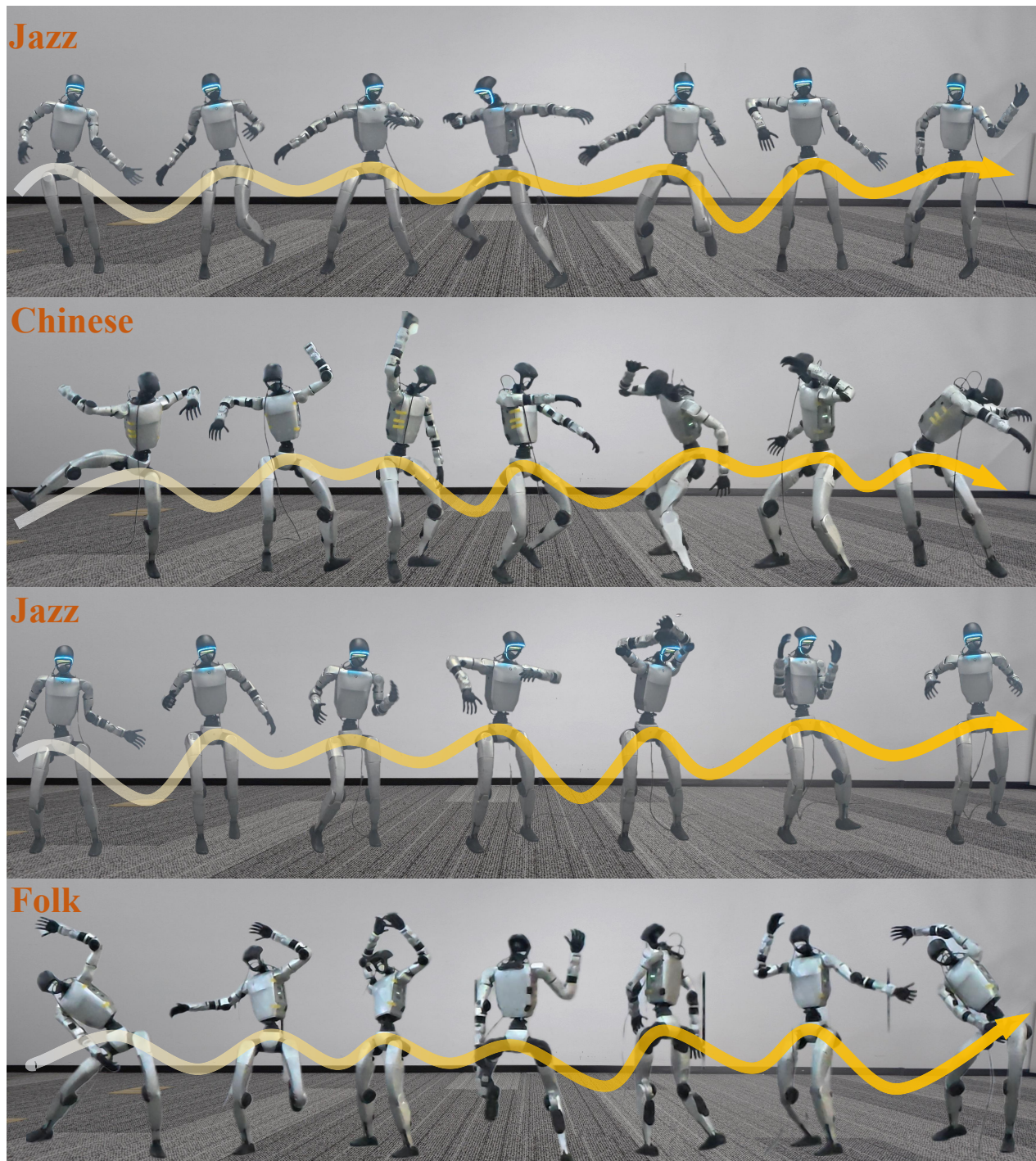


Figure 5. Real-world music-to-locomotion.

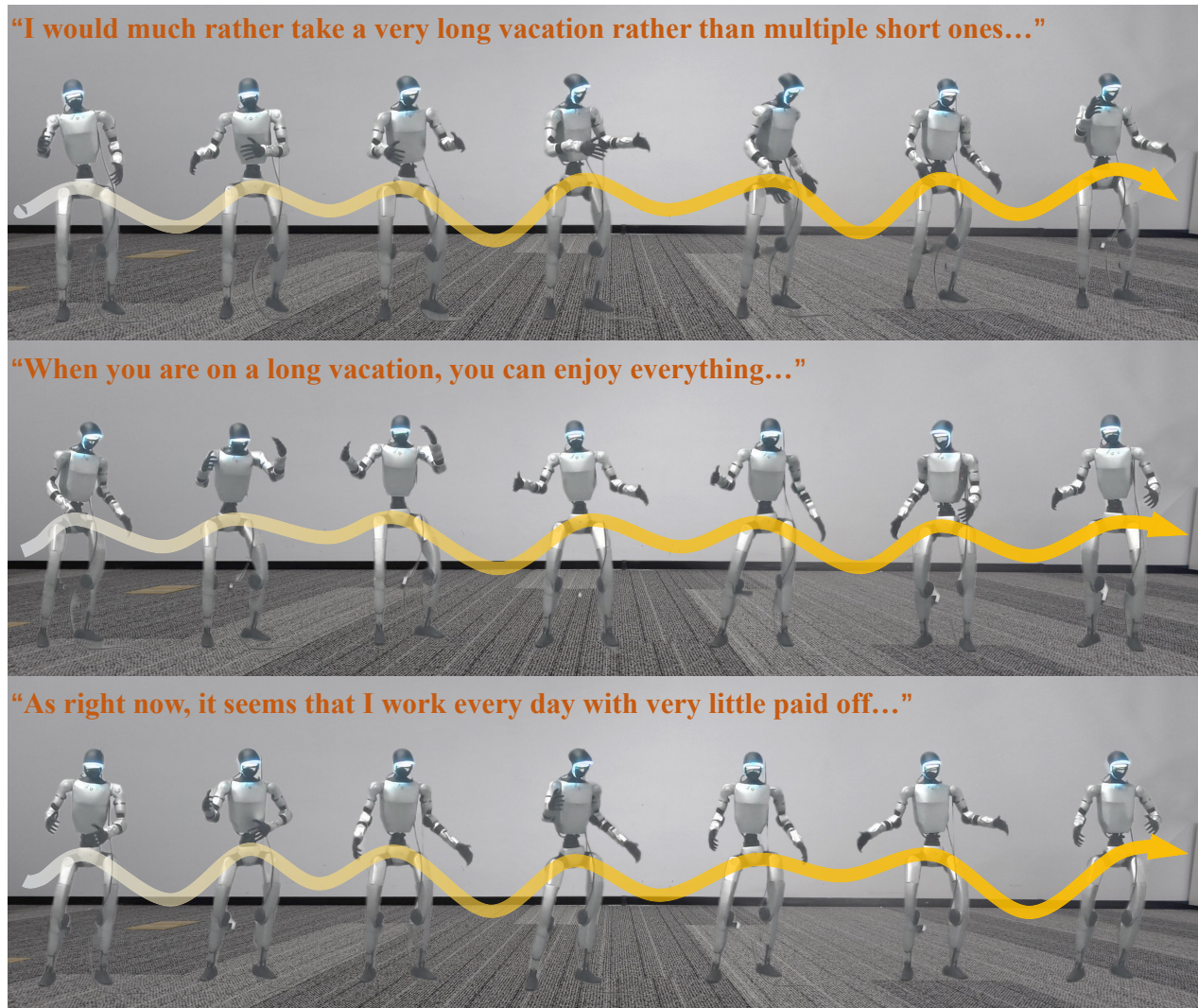


Figure 6. Real-world speech-to-locomotion.