RoboMirror: Understand Before You Imitate for Video to Humanoid Locomotion

Zhe Li[1,2,†,⋆], Boan Zhu[2,3,⋆], Yangyang Wei[2,4,⋆], Shuanghao Bai[5], Yuheng Ji[6], Yibo Peng[2], Tao Huang[7], Pengwei Wang[2], Zhongyuan Wang[2], S.-H. Gary Chan[3], Chang Xu[9], Cheng Chi[2✉], Jianfei Yang[1✉], and Shanghang Zhang[2,8✉]

[1] Nanyang Technological University
[2] Beijing Academy of Artificial Intelligence
[3] The Hong Kong University of Science and Technology
[4] Harbin Institute of Technology
[5] Xi'an Jiaotong University
[6] Chinese Academy of Sciences
[7] Shanghai Jiaotong University
[8] Peking University
[9] University of Sydney

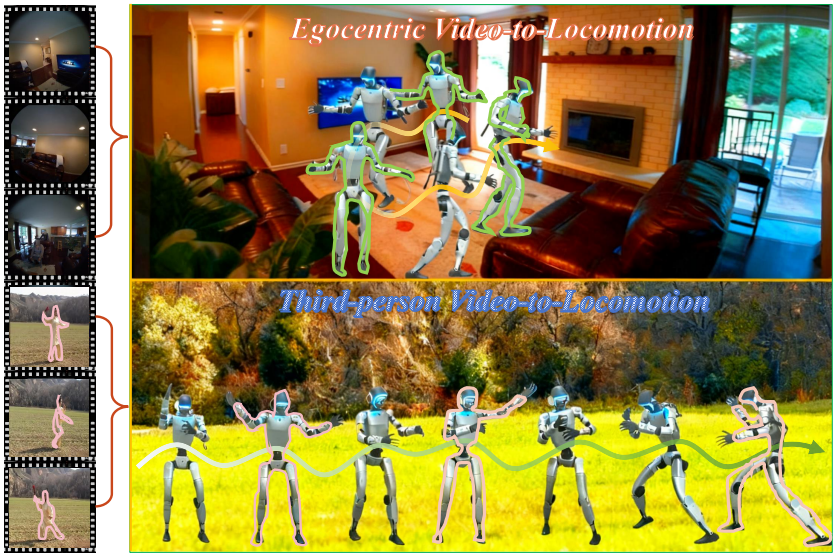

Fig. 1: RoboMirror makes humanoid understand before imitating. It acts like a mirror, which can not only infer and replicate the actions being performed by the shooter from egocentric videos based on the changes in the surrounding environmental perspective (as shown in the upper part of the figure), but also understand the actions first and then imitate them from third-person videos (as shown in the lower part of the figure), without the need for pose estimation and retargeting during inference.

⋆ Equal Contribution † Project Lead ✉ Corresponding Author

Abstract. Humans learn locomotion through visual observation, interpreting visual content first before imitating actions. However, state-of-the-art humanoid locomotion systems rely on either curated motion capture trajectories or sparse text commands, leaving a critical gap between visual understanding and control. Text-to-motion methods suffer from semantic sparsity and staged pipeline errors, while video-based approaches only perform mechanical pose mimicry without genuine visual understanding. We propose RoboMirror, the first retargeting-free video-to-locomotion framework embodying "understand before you imitate". Leveraging VLMs, it distills raw egocentric/third-person videos into visual motion intents, which directly condition a diffusion-based policy to generate physically plausible, semantically aligned locomotion without explicit pose reconstruction or retargeting. Extensive experiments validate RoboMirror's effectiveness, it enables telepresence via egocentric videos, drastically reduces third-person control latency by 80%, and achieves a 3.7% higher task success rate than baselines. By reframing humanoid control around video understanding, we bridge the visual understanding and action gap.

Keywords: Video to Humanoid Locomotion · Retargeting-free · Vision-Language Models · Diffusion Policy

1 Introduction

Humans intuitively learn locomotion by observing, a process predicated on understanding visual percepts before imitating them. Yet, current humanoid paradigms fail to replicate this "understand-then-act" architecture. Prevailing video-based methods degenerate into kinematic mimicry, a brittle process focused on reconstructing poses rather than inferring intent. Other dominant modalities are even less grounded: Motion Capture (MoCap) bypasses visual perception entirely, while text commands compress rich visual context into sparse symbols. A critical gap thus remains: current approaches either fail to genuinely understand rich visual data or bypass it altogether, fundamentally decoupling perception from control.

This failure to understand is exemplified by the "kinematic mimicry" paradigm [8, 9, 43, 44]. Its "pose estimate-retarget-track" pipeline is not only brittle, plagued by error accumulation and latency, but architecturally incapable of semantic understanding. By forcing the controller to track low-level kinematics, it precludes the learning of high-level visual intent. The alternative paradigms are flawed at the modality level. MoCap-driven systems [34, 35, 48] are inherently non-perceptive, while text-to-motion generation [6, 20, 22, 23] suffers from the inherent sparsity of language, which cannot encode the rich dynamics and goals abundant in video.

In this work, we argue for a different interface: ***video-to-locomotion***. Unlike sparse modalities like text or reference poses, video constitutes a substantially more information-dense medium. As presented in Figure 1, both egocentric and third-person videos encapsulate rich environmental cues, including scene attributes, temporal dynamics, and action goals. ***Our key insight is that this rich visual evidence must be internalized to condition a policy directly,***

rather than being reduced to brittle kinematics via pose estimation. We therefore regard humanoid locomotion as a generative problem: given this internalized visual context, synthesize physically plausible and semantically grounded motion. This direct "understand-then-act" mapping unlocks two capabilities that prior approaches cannot provide: (1) telepresence, where first-person video demonstration guides the robot to perform the corresponding locomotion, creating an "as if I were there" experience; and (2) robust, retargeting-free third-person imitation, avoiding the error accumulation that plagues conventional pipelines.

We therefore propose ***RoboMirror***, a retargeting-free framework named to emphasize the mirror-like mapping from video to action. Our core technical insight is to use a Vision-Language Model (VLM) to distill raw video into a visual latent representation, which is then explicitly trained to reconstruct a corresponding motion latent. This reconstructed motion latent serves as the sole conditioning signal for a diffusion-based locomotion policy. This architecture explicitly bridges semantic visual understanding with physics-based control. Concretely, we leverage the robust generalization of VLMs [1, 2, 41] to obtain visual latents from first- or third-person videos. These latents are mapped to motion latents, which are then fed into the diffusion policy to generate smooth, executable actions, bypassing explicit pose estimation and retargeting entirely.

Extensive experiments validate the effectiveness and practicality of RoboMirror. Compared directly to imitation based on pose estimation, our method dramatically accelerates the pipeline from video understanding to on-robot deployment, reducing latency from 9.22 s to 1.84 s. Beyond sheer speed, our method delivers higher-quality control by avoiding retargeting failures, as evidenced by a 3.7% absolute increase in task success rate and lower tracking error relative to baseline. Furthermore, for egocentric videos [28], we demonstrate that robust, semantically grounded locomotion can be synthesized without any explicit human pose supervision, a task where traditional pose-estimation pipelines fail. Crucially, we successfully introduce VLM into the humanoid control loop, enabling video-conditioned policies that can be further extended to fine-grained hand manipulation and low-friction teleoperation. In short, RoboMirror reframes humanoid control around video understanding. By learning to understand first and imitate second, we close the gap between what is seen and how to move, moving from fragile pose mimicry to robust, visually grounded action.

Our contributions can be summarized as follows:

- We propose RoboMirror, the first retargeting-free framework for humanoid locomotion that replaces brittle pose reconstruction with a direct mapping from 2D video to visual motion intent, enabling a true end-to-end, video-to-locomotion policy.
- We introduce a VLM-assisted locomotion policy where a VLM distills visual latents from raw video. These latents are trained to reconstruct motion latents, which in turn serve as a robust, non-kinematic conditioner for a diffusion-based action generator.
- We validate RoboMirror's significant outperformance against "pose estimation-retarget-track" baselines in task success rate and latency. Crucially, we

demonstrate robust locomotion from egocentric video without explicit pose supervision, a task where traditional pipelines fail.

2 Related Work

2.1 Humanoid Whole-body Control

Model-based whole-body controllers deliver precise task execution via accurate dynamics and contact modeling, but entail heavy modeling effort and limited generalization to new skills or unmodeled dynamics [5,37]. Learning-based approaches ease modeling burden yet hinge on carefully crafted, task-specific rewards; despite successes in challenging settings such as locomotion on complex terrains, jumping, and fall recovery [10,12,21,24,27,30,39,40,47,52,53], they require per-task reward engineering and often struggle to produce coordinated, human-like behaviors. To disentangle distinct upper- vs. lower-body objectives, some studies split control into independent policies [16, 51], which can undermine inter-limb coordination and limit generality. Others adopt hierarchical planning to learn complex sequential skills, e.g.table tennis [38], improving modularity but adding design complexity and latency. Whole-body motion tracking reframes the objective by directly using human motion as supervision [7], eliminating task-specific rewards while promoting globally coordinated, expressive motion across diverse skills. This perspective provides a unified control objective and a scalable path toward human-like whole-body behavior.

2.2 Humanoid Motion Tracking

Learning realistic behaviors from human motion has become central to high-fidelity control. DeepMimic [29] introduced a phase-based tracking framework with random state initialization and early termination to stabilize imitation.To mitigate sim-to-real gaps for dynamic skills, ASAP [8] proposed a multi-stage pipeline with a delta-action model. HuB [50] and KungfuBot [44] further leverage sophisticated motion preprocessing and tracking mechanisms to achieve precise imitation of highly dynamic single motions.

For multi-skill policies within a single controller, OmniH2O [9] demonstrated a universal policy that catalyzed subsequent research on broad motion libraries. ExBody2 [14] improves expressiveness via decomposed tracking targets and motion filtering. TWIST [49] and CLONE [17] attain high-quality tracking in teleoperation settings but primarily cover lower-dynamic motions. BumbleBee [42] employs a two-stage strategy, which is clustering motions to train expert policies, then distilling them into a unified policy. GMT [4] achieves robust tracking of aggressive motions by prioritizing root velocity and pose information over global positions. UniTracker [46] supports dynamic tracking but relies on global targets, limiting stability on long sequences. BeyondMimic [25] attains high-fidelity single-motion tracking via carefully designed objectives and precise system identification, followed by distillation into a unified diffusion policy for task-specific

control. KungfuBot2 [7] proposes an orthogonal MoE to realize a general motion tracking policy spanning diverse skills. Building on these insights, we pursue a universal policy that conditions on video to generate humanoid locomotion, moving from fragile kinematic mimicry to visually grounded performance control and transforming humanoids into semantically aligned imitators.

2.3 Modality-driven Humanoid Locomotion

Recent work explores conditioning humanoid locomotion on high-level modalities, with language as a prominent interface. LangWBC [35] pairs a compact auxiliary network with the control policy to generate motions online from instructions, but its limited capacity hinders scaling to complex, diverse motion distributions and offers weak guarantees for generalization to unseen prompts. RLPF [48] fine-tunes an LLM and introduces physical feasibility feedback from a motion-tracking policy to iteratively align semantic intent with executable motion, helping bridge sim-to-real; however, heavy decoder updates risk catastrophic forgetting of pretrained knowledge. RoboGhost [19] proposes a latent-driven, retargeting-free framework that treats locomotion as generation, reducing error accumulation and inference latency, but it considers language as the sole input modality. RoboPerform [18] proposes a retargeting-free audio-to-locomotion framework that unifies music-driven dance and speech-driven co-speech gesture generation for humanoids. LeVERB [45] explores vision-language-guided whole-body control by leveraging visual feedback and linguistic instructions; it depends on retargeting pre-collected motion data and focuses mainly on quasi-static or low-dynamic tasks. In contrast, we directly process raw egocentric/third-person video without motion retargeting, aligning semantic understanding from vision with physically feasible actions using a latent-driven diffusion policy.

3 Method

This section presents the core components of our framework, which is depicted in Figure 2. We start with an overview of our main framework and its motivation in Section 3.1, offering a high-level description of the architectural design and underlying rationale. Section 3.2 elaborates on the method for reconstructing motion latent from VLM-derived latent, which leverages a diffusion model for accurate and robust reconstruction. Furthermore, Section 3.3 introduces our MoE-based residual teacher policy and the latent-guided diffusion-based student policy, along with a detailed exposition of their inference procedures. Other implementation details are provided in the Appendix.

3.1 Overview

We present a novel framework that enables robots to understand visual content from videos and execute corresponding physical actions in a coherent manner. At its core, our framework replaces simplistic cross-modal alignment with motion

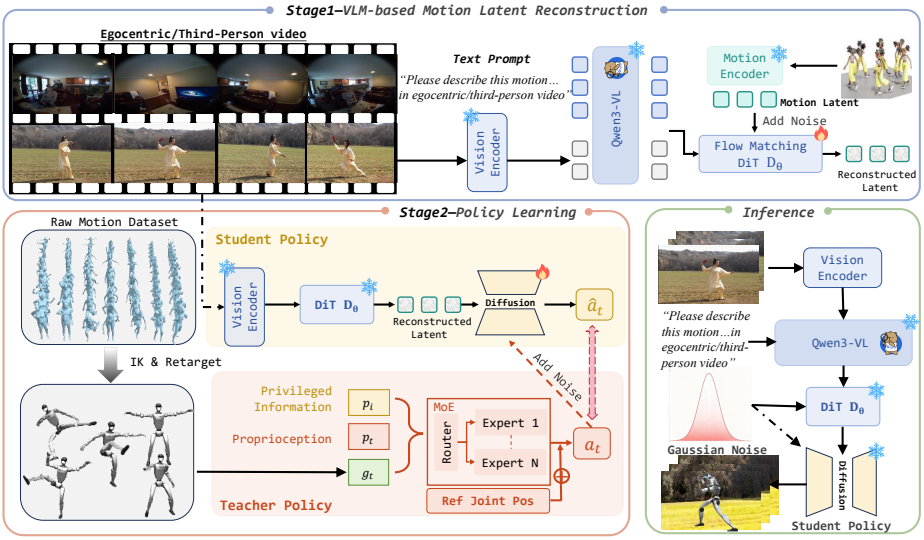

Fig. 2: Overview of RoboMirror. It adopts a two-stage framework: initially, it leverages Qwen3-VL to process egocentric or third-person video inputs, generating motion latents through diffusion models with DiT $\mathcal{D}_\theta$. Subsequently, in the policy learning stage, a MoE-based teacher policy is trained with RL, while a diffusion-based student policy learns to denoise actions under the guidance of reconstructed motion latents. During inference, it can first understand and then imitate the motion in the video without motion obtainment and retargeting.

latent reconstruction, an intentional choice rooted in the principle that robust, semantically meaningful vision-motion mapping arises from generative reconstruction rather than superficial feature matching. As illustrated in Figure 2, our approach comprises three sequentially linked components: a VLM-based multi-view video understanding module, a VLM-conditioned diffusion model for motion latent reconstruction, and a diffusion-based policy training module. We address a critical challenge: generating physically executable actions for humanoid robots directly from videos of varying perspectives (first- or third-person), entirely eliminating reliance on error-prone pose estimation.

The pipeline unfolds in a natural "understand-reconstruct-control" progression. It begins by feeding input videos into a pretrained VLM, which extracts a high-level semantic latent l_{VLM} encoding both action details and contextual scene information. Unlike existing methods that naively repurpose l_{VLM} as a control signal, we treat this latent as a semantic anchor for a diffusion-based motion latent reconstructor. This design stems from a key insight: "reconstruction outperforms alignment." When a model learns to reconstruct kinematically coherent motion latents from VLM semantics, it achieves not only more robust cross-modal alignment but also inherently embeds physical plausibility constraints, avoiding the semantic disconnection plaguing direct alignment strategies.

The reconstructed motion latent l_{motion} then fuses with the robot's proprioceptive states and real-time observations to condition a diffusion-based deployment policy π. Through iterative denoising, this policy outputs actions directly executable on humanoid platforms. By obviating explicit motion estimation and alignment procedures, our reconstruction-driven paradigm enables an end-to-end video-to-action mapping, uniquely suited for practical deployment scenarios where reliability and efficiency are paramount.

3.2 Motion Latent Reconstruction from Vision-Language Model

Leveraging the strong image and video understanding capabilities as well as robust generalization of vision-language models, we adopt Qwen3-VL [2] as our video understanding module. We first train a VAE [15] on our motion dataset to reconstruct motion sequences. For video processing, we feed first-person or third-person videos into Qwen3-VL with task-specific prompts. For egocentric videos, the prompt is *"Please describe the motion of the first-person individual in the egocentric video"* while for third-person videos it is *"Please describe the motion in the video"*. This design ensures the latent representations output by Qwen3-VL are enriched with high-quality semantic information about the motion content.

For infusing kinematic information into video latents and reconstructing motion representations, we employ a flow-matching based diffusion model, denoted as $\mathcal{D}_\theta$, which takes VLM-derived video latents as conditional signals to reconstruct VAE-learned motion latents. As illustrated in Fig. 2, $\mathcal{D}_\theta$ consists of stacked transformer blocks with adaptive layer normalization, enabling effective conditioning on video semantics while preserving motion kinematic structure [3]. The model operates on noised motion latents ϵ_{motion}, with cross-attention blocks attending to the video latents $\mathbf{l}_{\text{VLM}}$ to enforce semantic consistency, and self-attention blocks capturing temporal dependencies within motion sequences.

Given a ground-truth motion latent l_{motion} from pretrained VAE, a flow-matching timestep $\tau \in [0, 1]$, and sampled Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, the noised motion latent ϵ_{motion} is constructed as:

$$\epsilon_{\text{motion}} = \tau \cdot l_{\text{motion}} + (1 - \tau) \cdot \epsilon.$$

The diffusion model $\mathcal{D}_\theta$ takes l_{VLM}, ϵ_{motion}, and timestep τ as inputs, aiming to predict the velocity vector field $\epsilon - \epsilon_{\text{motion}}$ for flow matching. To optimize $\mathcal{D}_\theta$, we minimize the following velocity-prediction loss:

$$\mathcal{L}_{\text{fm}} = \mathbb{E}_{\tau,\text{motion},\epsilon}\left[\|\mathcal{D}_\theta(l_{\text{VLM}}, \epsilon_{\text{motion}}, \tau) - (\epsilon - \epsilon_{\text{motion}})\|_2^2\right].$$

This training paradigm ensures that $\mathcal{D}_\theta$ learns to reconstruct motion latents with kinematic information from video semantics, inherently achieving robust cross-modal alignment without separate alignment modules.

3.3 Policy Training

MoE-based Residual Motion Tracker We argue that the key to enabling robots to directly observe or infer human motions from videos and perform actions lies in the generalization capability of motion trackers. Specifically, we aim for these trackers to successfully respond to novel prompts while achieving genuine deployment flexibility, which are two critical properties for bridging video understanding and real-world robotic locomotion.

First, we train an oracle teacher policy using the PPO algorithm [33] with privileged simulator-state information. To learn a policy π_t that generalizes across diverse motion inputs, we first train an initial policy π_0 on a highly diverse motion dataset $\mathcal{D}_0$ [14, 19]. Given the relative simplicity of egocentric motion datasets $\mathcal{D}_{\text{first}}$, we restrict the training of π_0 exclusively to third-person view motion datasets $\mathcal{D}_{\text{third}}$.

Subsequently, we evaluate the tracking accuracy of π_0 for each motion sequence $s \in \mathcal{D}_{\text{third}}$, with a specific focus on lower-body motion precision. This evaluation employs the error metric $e(s) = \alpha \cdot E_{\text{key}}(s) + \beta \cdot E_{\text{dof}}(s)$, where $E_{\text{key}}(s)$ denotes the mean position error of key lower-body landmarks and $E_{\text{dof}}(s)$ represents the mean tracking error of lower-body joint angles. Motion sequences with $e(s) > 0.6$ are filtered out, and the remaining data $\mathcal{D}$ are used to train a generalizable teacher policy.

The teacher policy π_t is trained as a simulation oracle via PPO, utilizing real-world unavailable privileged information: ground-truth root velocity, global joint positions, physical properties (e.g., friction, motor strength), proprioceptive state, and reference motion. Besides, to handle challenging motions, we design the policy to focus on learning dynamic information, specifically a corrective offset δ_a rather than directly learning kinematic information such as absolute joint targets p_{target}. This offset is added to the joint positions from the reference kinematic trajectory, yielding our final output action $\hat{a}_t \in \mathbb{R}^{23}$ forming: $\hat{a}_t = p_{\text{target}} + \delta_a$, which is optimized via cumulative rewards to ensure accurate motion tracking and robust behaviors.

Furthermore, we integrate a Mixture of Experts (MoE) module to enhance expressiveness and generalization. The policy includes expert networks and a gating network, which takes the same inputs and computes weights for the experts' outputs to form a weighted sum. The final action is computed as $\hat{a}_t = \sum_{i=1}^{n} p_i \cdot a_i$, with p_i denoting the gating probability for expert i and a_i is its corresponding output. This design boosts generalization, improves tracking accuracy, and enables precise supervision of the student policy.

Diffusion-based Student Policy Unlike prior work where student policies π_s distill knowledge from teachers via explicit reference motion, we regard the student policy as a generation model, which is formulated as a latent-driven generation task [19]. It takes motion latents generated under video latent guidance as input, alongside observation history, to generate humanoid actions. This design enables the robot to more quickly imitate the motion of the subject in the video,

bypassing error-prone steps such as pose estimation and motion retargeting, significantly reducing the time consumption of the entire process.

Following a DAgger-like paradigm, we train the student by rolling it out in simulation, querying the teacher for optimal actions $\hat{a}_t$ at observable states. During training, we inject Gaussian noise ϵ_t into teacher actions and use our reconstructed latents l_{v2m}, which are from video latents, as guiding latents. The noising process follows a Markov chain:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\alpha_t} \cdot x_{t-1}, \alpha_t \mathbf{I}) \tag{1}$$

where $\alpha_t \in (0, 1)$ is a sampling hyper-parameter. Denoising at step t is modeled as $x_{t-1} = \epsilon_\theta(x_t, t)$, with ϵ_θ as the denoiser. Using an x_0-prediction strategy, we supervise via MSE loss: $\mathcal{L} = \|a - \hat{a}_t\|_2^2$, where $a = \frac{x_t - \sqrt{1-\bar{\alpha}_t} \cdot \epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}$. Converged policies require no privileged knowledge or explicit references, enabling real-world deployment.

Inference Pipeline To ensure fluent, smooth motion, we minimize denoising time by adopting DDIM sampling [36] and an MLP-based diffusion model for action generation. The reverse process is:

$$\begin{aligned} x_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1-\alpha_t} \cdot \epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} \right) \\ + \sqrt{1-\alpha_{t-1}} \cdot \epsilon_\theta(x_t, t) \end{aligned} \tag{2}$$

During inference, our pipeline operates as follows: we first input either an egocentric or third-person video into Qwen3-VL to obtain a video latent representation l_{vlm}. This video latent l_{vlm} is then fed into our pretrained diffusion model $\mathcal{D}_\theta$, yielding a motion latent l_{v2m} enriched with kinematic semantics. Finally, we use l_{v2m} as a conditional signal to guide our diffusion-based student policy, which generates deployable actions directly.

Notably, this pipeline eliminates the need for pose estimation from third-person videos or motion guessing from egocentric videos. By leveraging the video understanding capability of VLMs and the latent comprehension ability of our policy, we successfully achieve video-to-locomotion, enabling the robot to imitate both sparse and dense motion modalities present in the input video.

4 Experiments

We evaluate our RoboMirror on both egocentric and third-person videos, aiming to verify its action imitation capabilities for videos of different perspectives respectively. Specifically, for egocentric videos, the model first infers a dense motion latent from the video with sparse motion information, and then performs the imitation task. In the experiments, we train the teacher policy and student policy in the IsaacGym simulation environment, and directly deploy the student policy on the Unitree G1 humanoid robot for real-world testing.

4.1 Experimental Setups

Dataset We train our model on the Nymeria [28] and Motion-X [26] datasets. Nymeria is a large-scale real-device dataset capturing diverse human daily activities across indoor and outdoor locations, providing paired text-video-motion data—including egocentric videos, full-body motions, and human-annotated motion narrations. This dataset contains approximately 1K videos, each around 15-20 minutes long. Due to its large scale, we select 100 videos, segment each into 5-second clips, and ultimately obtain 18K video-motion pairs. Both the motions and videos are sampled at 30 FPS.

Motion-X is a large-scale 3D expressive whole-body motion dataset, constructed from massive online videos and eight existing motion datasets, with motions formatted as SMPL-X. We select motion categories paired with videos from it as our third-person dataset. We split the dataset into train and test sets with an 8:2 ratio.

Metrics We adopt two categories of evaluation metrics: motion latent reconstruction and motion tracking. For motion latent reconstruction, we report motion-video retrieval accuracy MV-R@3, MM Dist-V (MM-V), dynamic time warping (DTW), and foot skating to evaluate whether the reconstructed motion latents conform to the semantics of the input videos, where we only evaluate on third-person video datasets. Additionally, we report motion-text retrieval accuracy MT-R@3, MM Dist-T (MM-T), and FID to assess whether the reconstructed motion latents contain sufficient kinematic and semantic information. For motion tracking, evaluated in physics simulators aligning with prior works [9], we use success rate as the core indicator, supplemented by mean per-joint position error (E_{MPJPE}) and mean per-keypoint position error (E_{MPKPE}). Detailed metric definitions are provided in the Appendix.

Implementation Details We first input videos into a pretrained VLM for video understanding to obtain video latents, where we use the Qwen3-VL-4B-Instruct model. For the motion latent reconstruction network, we adopt a 16-layer MLP as the backbone and inject conditions via AdaLN [13] to guide the model in reconstructing motion latents from Gaussian noise. For policy training, teacher policy adopts MoE structure of 5 experts, and the student policy employs a diffusion model with a 4-layer MLP as its backbone; this compact network architecture, combined with DDIM sampling, ensures real-time performance during deployment. More details can be found in the Appendix.

4.2 Evaluation of Motion Latent Reconstruction

To validate the motion latent reconstruction capability of the diffusion model trained with latents from VLMs as conditions, we decode the reconstructed latents into real motions using our pretrained VAE decoder. The reconstruction performance of motion latents is then evaluated by assessing both the generation quality of the motions and their degree of semantic relevance. Specifically, we

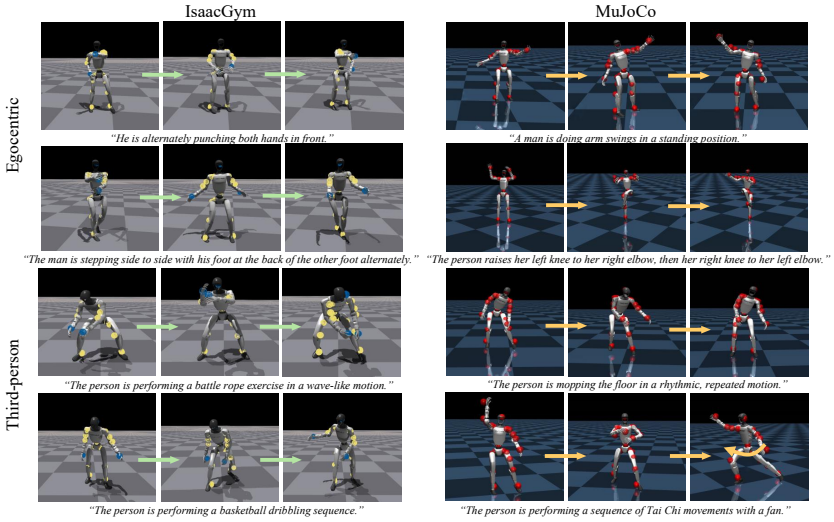

Fig. 3: Qualitative results in the IsaacGym and MuJoCo. The upper half presents the tracking performance of egocentric videos-to-locomotion, and the lower half presents that of third-person videos-to-locomotion.

Methods	Nymeria			Motion-X					
	MT-R@3↑	FID↓	MM-T↓	MT-R@3↑	MV-R@3↑	FID↓	MM-T↓	MM-V↓	DTW↓
Vid2Mot	49.34	62.81	25.67	69.40	68.8	31.27	16.22	6.65	0.33
ViMo [31]	-	-	-	76.92	67.8	21.03	8.48	7.42	0.31
Ours	64.60	42.42	16.97	77.66	72.3	19.53	8.17	6.24	0.23

Table 1: Performance comparison of motion-text and motion-video alignment on the Nymeria and Motion-X datasets.

evaluate the model's reconstruction performance on the Nymeria and Motion-X datasets, with the results summarized in Table 1. For video-motion alignment, we further employ MV-R@3, MM-V, DTW [32], and foot skating to quantitatively measure the temporal consistency and physical plausibility of reconstructed motions with corresponding videos, which is also presented in Table 1. Here, we introduce a simple baseline, named Vid2Mot, which finetunes the VLM via LoRA [11] while freezing the VAE decoder. This design enables the latents output by the VLM to be directly decoded into motions through the decoder.

4.3 Evaluation of Motion Tracking

To further validate the efficacy of our motion tracking policy, we conduct evaluations on egocentric and third-person videos, measuring the Mean Per Joint Position Error (E_{mpjpe}) and Mean Per Keypoint Position Error (E_{mpkpe}) under physics-based simulations in both IsaacGym and MuJoCo. The pipeline proceeds as follows: first, videos and prompts are fed into the VLM for video understanding and latent representation generation; subsequently, these latents are used as

Method	IsaacGym			MuJoCo		
	Succ ↑	E_{mpjpe} ↓	E_{mpkpe} ↓	Succ ↑	E_{mpjpe} ↓	E_{mpkpe} ↓
Nymeria						
Baseline	0.92	0.19	0.17	0.69	0.27	0.24
Ours	0.99	0.08	0.11	0.78	0.19	0.17
Motion-X						
Baseline	0.91	0.23	0.20	0.61	0.36	0.33
Ours	0.95	0.17	0.15	0.70	0.28	0.27

Table 2: Motion tracking performance comparison in simulation on the Nymeria and Motion-X test sets.

Method	Succ ↑	R@3 (%)↑	FID ↓	MM-Dist↓
Nymeria				
Qwen-VL	0.97	61.1	46.73	17.64
Qwen2-VL	0.98	61.9	44.02	17.53
Qwen2.5-VL	0.97	62.6	42.74	17.80
Qwen3-VL	0.99	64.6	42.42	16.97
Motion-X				
Qwen-VL	0.93	76.84	20.69	8.31
Qwen2-VL	0.94	77.08	20.42	8.19
Qwen2.5-VL	0.95	77.58	19.77	8.26
Qwen3-VL	0.95	77.66	19.53	8.17

Table 3: Ablation study on different vision-language models across motion latent reconstruction quality and tracking performance.

conditions to input into our pretrained diffusion model in Section 3.2 for motion latent reconstruction; finally, the student policy takes this representation and generates actions. As shown in Table 2, our method achieves high task success rates on the Nymeria and Motion-X datasets, alongside low joint and keypoint errors—indicating strong alignment between the semantic information of videos and physically executable trajectories. The baseline herein refers to the result of finetuning Qwen3-VL with LoRA, where motion reconstruction is adopted as the optimization objective, and the latents output by Qwen3-VL are directly used to guide the policy in action generation.

4.4 Qualitative Results

We conduct a qualitative evaluation of the motion tracking policy across two deployment scenarios: simulation (IsaacGym) and cross-simulator transfer (MuJoCo). Figure 3 illustrates representative tracking sequences, emphasizing the

Method	Succ ↑	R@3 (%)↑	FID ↓	MM-Dist↓
Nymeria				
Alignment	0.95	51.1	62.42	30.54
Reconstruction	0.99	64.6	42.42	16.97
Motion-X				
Alignment	0.92	56.41	40.19	24.37
Reconstruction	0.95	77.66	19.53	8.17

Table 4: Ablation study on alignment vs reconstruction across motion latent reconstruction quality and tracking performance.

Method	Motion-X			
	Succ ↑	E_{mpjpe} ↓	E_{mpkpe} ↓	Time Cost (s) ↓
Pose-driven	0.94	0.17	0.15	9.22
Latent-driven	0.95	0.16	0.15	1.84

Table 5: Ablation study on comparing the tracking performance of pose-driven and latent-driven approaches for the third-person video-to-locomotion task.

policy's capacity to preserve motion semantics, maintain balance during dynamic transitions, and generalize across distinct physics engines and hardware platforms.

Additionally, in Figure 4, we also provide visualization results of the generated motion decoded from the motion latents reconstructed from DiT $\mathcal{D}_\theta$ via the pretrained decoder. More qualitative results in the simulation and real-world can be seen in the Appendix.

4.5 Ablation Studies

To systematically validate the effectiveness of our method, we conduct a series of ablation studies in this section. These experiments cover three key aspects: 1) different video understanding models, 2) diverse approaches for converting video latents to motion latents, and 3) the advantages of our method over pose estimation-based student policy for third-person video-driven locomotion.

Different Vision-Language Models To verify the impact of different vision-language models on the final results, we employ four distinct models: Qwen-VL, Qwen-2-VL, Qwen2.5-VL, and Qwen3-VL. We evaluate their performance in terms of motion reconstruction quality and their influence on policy tracking performance, with the results summarized in Table 3. As indicated in the table, Qwen3-VL generates superior latent representations that are more suitable for motion latent reconstruction, thereby enabling more effective control of humanoid robots.

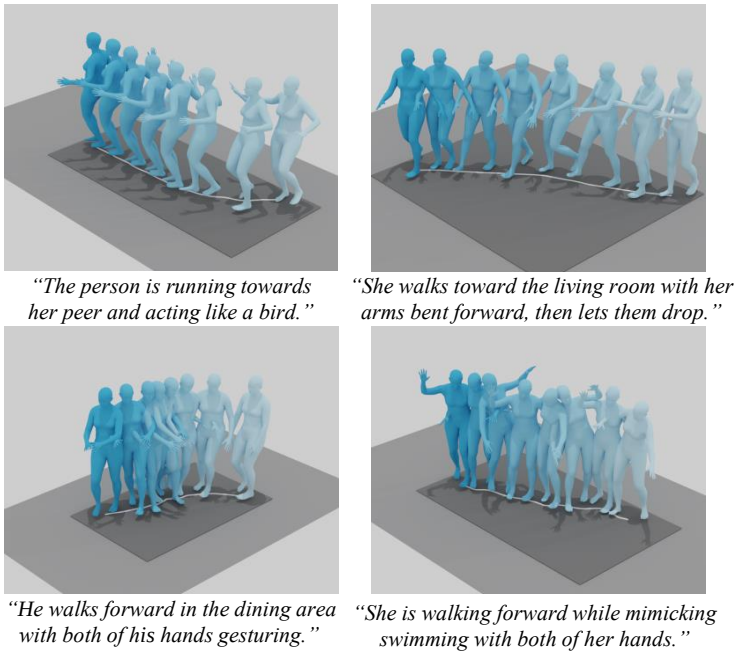

Fig. 4: Qualitative results of generated motions.

Alignment Vs Reconstruction We argue that reconstruction outperforms alignment. Since the latents output by VLM lack kinematic information, directly using them to guide action generation results in unstable motions with ambiguous task relevance. Thus, it is necessary to convert l_{vlm} into l_{motion} embedded with kinematic cues. One approach is to align l_{vlm} with l_{motion} to impart kinematic information, which then guides action generation. However, we hypothesize that reconstructing target motion latents l_{motion} from VLM latents l_{vlm} can more effectively bridge the domain gap between visual semantics and motion dynamics.

To validate this, we conduct ablation experiments on alignment and reconstruction. For the alignment baseline, we train a 4-layer transformer adapter with InfoNCE loss, which pulls positive sample pairs closer while pushing negative pairs apart. As shown in Table 4, reconstructing l_{motion} from l_{vlm} achieves significantly superior performance.

Latent-driven Vs Pose-driven For evaluating our method's ability to imitate actions from third-person videos, we assess the advantages of RoboMirror over conventional pose estimation-based methods. Herein, pose estimation-based methods refer to approaches that first estimate poses from video test sets, then retarget the poses to the humanoid, and finally feed the retargeted reference motion as input to the student policy for action generation. As shown in Table 5, RoboMirror achieves superior tracking performance with lower overall pipeline

latency. This is attributed to the fact that both pose estimation and retargeting incur non-negligible time costs and introduce cumulative errors, whereas our framework avoids such inefficiencies.

5 Conclusion

We present RoboMirror, a video-to-locomotion framework rooted in "understand before you imitate." Leveraging VLMs, it extracts semantic motion intents from videos and reconstructs them into kinematically grounded latents, enabling humanoids to generate physically plausible, semantically aligned actions without pose estimation or retargeting. Extensive experiments validate its superiority in task success, latency, and cross-domain generalization. RoboMirror bridges visual understanding and humanoid locomotion, laying groundwork for understanding-driven control.

6 Acknowledgement

This work was supported by the National Natural Science Foundation of China (62476011) and Beijing Natural Science Foundation (L252060).

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., Fang, Y., Fox, D., Hu, F., Huang, S., et al.: Gr00t n1: An open foundation model for generalist humanoid robots. arXiv preprint arXiv:2503.14734 (2025)
4. Chen, Z., Ji, M., Cheng, X., Peng, X., Peng, X.B., Wang, X.: Gmt: General motion tracking for humanoid whole-body control. arXiv preprint arXiv:2506.14770 (2025)
5. Geyer, H., Seyfarth, A., Blickhan, R.: Positive force feedback in bouncing gaits? Proceedings of the Royal Society of London. Series B: Biological Sciences **270**(1529), 2173–2183 (2003)
6. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1900–1910 (2024)
7. Han, J., Xie, W., Zheng, J., Shi, J., Zhang, W., Xiao, T., Bai, C.: Kungfubot2: Learning versatile motion skills for humanoid whole-body control. arXiv preprint arXiv:2509.16638 (2025)
8. He, T., Gao, J., Xiao, W., Zhang, Y., Wang, Z., Wang, J., Luo, Z., He, G., Sobanbab, N., Pan, C., et al.: Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. arXiv preprint arXiv:2502.01143 (2025)

9. He, T., Luo, Z., He, X., Xiao, W., Zhang, C., Zhang, W., Kitani, K., Liu, C., Shi, G.: Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. arXiv preprint arXiv:2406.08858 (2024)
10. He, X., Dong, R., Chen, Z., Gupta, S.: Learning getting-up policies for real-world humanoid robots. arXiv preprint arXiv:2502.12152 (2025)
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
12. Huang, T., Ren, J., Wang, H., Wang, Z., Ben, Q., Wen, M., Chen, X., Li, J., Pang, J.: Learning humanoid standing-up control across diverse postures. arXiv preprint arXiv:2502.08378 (2025)
13. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
14. Ji, M., Peng, X., Liu, F., Li, J., Yang, G., Cheng, X., Wang, X.: Exbody2: Advanced expressive humanoid whole-body control. arXiv preprint arXiv:2412.13196 (2024)
15. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
16. Li, Y., Zhang, Y., Xiao, W., Pan, C., Weng, H., He, G., He, T., Shi, G.: Hold my beer: Learning gentle humanoid locomotion and end-effector stabilization control. In: RSS 2025 Workshop on Whole-body Control and Bimanual Manipulation: Applications in Humanoids and Beyond
17. Li, Y., Lin, Y., Cui, J., Liu, T., Liang, W., Zhu, Y., Huang, S.: Clone: Closed-loop whole-body humanoid teleoperation for long-horizon tasks. arXiv preprint arXiv:2506.08931 (2025)
18. Li, Z., Chi, C., Wei, Y., Zhu, B., Huang, T., Sun, Z., Peng, Y., Wang, P., Wang, Z., Liu, F., et al.: Do you have freestyle? expressive humanoid locomotion via audio control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 956–965 (2026)
19. Li, Z., Chi, C., Wei, Y., Zhu, B., Peng, Y., Huang, T., Wang, P., Wang, Z., Zhang, S., Xu, C.: From language to locomotion: Retargeting-free humanoid control via motion latent guidance. arXiv preprint arXiv:2510.14952 (2025)
20. Li, Z., He, Y., Zhong, L., Shen, W., Zuo, Q., Qiu, L., Dong, Z., Yang, L.T., Yuan, W.: Mulsmo: Multimodal stylized motion generation by bidirectional control flow. arXiv preprint arXiv:2412.09901 (2024)
21. Li, Z., Yang, L.T., Nie, X., Ren, B., Deng, X.: Enhancing sentence representation with visually-supervised multimodal pre-training. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5686–5695 (2023)
22. Li, Z., Yuan, W., He, Y., Qiu, L., Zhu, S., Gu, X., Shen, W., Dong, Y., Dong, Z., Yang, L.T.: Lamp: Language-motion pretraining for motion generation, retrieval, and captioning. arXiv preprint arXiv:2410.07093 (2024)
23. Li, Z., Yuan, W., Shen, W., Zhu, S., Dong, Z., Xu, C.: Omnimotion: Multi-modal motion generation with continuous masked autoregression. arXiv preprint arXiv:2510.14954 (2025)
24. Li, Z., Peng, X.B., Abbeel, P., Levine, S., Berseth, G., Sreenath, K.: Robust and versatile bipedal jumping control through reinforcement learning. arXiv preprint arXiv:2302.09450 (2023)
25. Liao, Q., Truong, T.E., Huang, X., Tevet, G., Sreenath, K., Liu, C.K.: Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. arXiv preprint arXiv:2508.08241 (2025)

26. Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., Zhang, L.: Motion-x: A large-scale 3d expressive whole-body human motion dataset. Advances in Neural Information Processing Systems (2023)
27. Liu, M., Zhou, E., Chi, C., Han, Y., Rong, S., Chen, L., Wang, P., Wang, Z., Zhang, S.: Sapave: Towards active perception and manipulation in vision-language-action models for robotics. arXiv preprint arXiv:2603.12193 (2026)
28. Ma, L., Ye, Y., Hong, F., Guzov, V., Jiang, Y., Postyeni, R., Pesqueira, L., Gamino, A., Baiyya, V., Kim, H.J., et al.: Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In: European Conference on Computer Vision. pp. 445–465. Springer (2024)
29. Peng, X.B., Abbeel, P., Levine, S., Van de Panne, M.: Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. ACM Transactions On Graphics (TOG) **37**(4), 1–14 (2018)
30. Peng, X.B., Ma, Z., Abbeel, P., Levine, S., Kanazawa, A.: Amp: Adversarial motion priors for stylized physics-based character control. ACM Transactions on Graphics (ToG) **40**(4), 1–20 (2021)
31. Qiu, L., Yu, C., Li, Y., Wang, Z., Huang, H., Ma, C., Zhang, D., Wan, P., Han, X.: Vimo: Generating motions from casual videos. arXiv preprint arXiv:2408.06614 (2024)
32. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing **26**(1), 43–49 (2003)
33. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
34. Serifi, A., Grandia, R., Knoop, E., Gross, M., Bächer, M.: Robot motion diffusion model: Motion generation for robotic characters. In: SIGGRAPH asia 2024 conference papers. pp. 1–9 (2024)
35. Shao, Y., Huang, X., Zhang, B., Liao, Q., Gao, Y., Chi, Y., Li, Z., Shao, S., Sreenath, K.: Langwbc: Language-directed humanoid whole-body control via end-to-end learning. arXiv preprint arXiv:2504.21738 (2025)
36. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
37. Sreenath, K., Park, H.W., Poulakakis, I., Grizzle, J.W.: A compliant hybrid zero dynamics controller for stable, efficient and fast bipedal walking on mabel. The International Journal of Robotics Research **30**(9), 1170–1193 (2011)
38. Su, Z., Zhang, B., Rahmanian, N., Gao, Y., Liao, Q., Regan, C., Sreenath, K., Sastry, S.S.: Hitter: A humanoid table tennis robot via hierarchical planning and learning. arXiv preprint arXiv:2508.21043 (2025)
39. Tan, H., Zhou, E., Li, Z., Xu, Y., Ji, Y., Chen, X., Chi, C., Wang, P., Jia, H., Ao, Y., et al.: Robobrain 2.5: Depth in sight, time in mind. arXiv preprint arXiv:2601.14352 (2026)
40. Wang, H., Wang, Z., Ren, J., Ben, Q., Huang, T., Zhang, W., Pang, J.: Beamdojo: Learning agile humanoid locomotion on sparse footholds. arXiv preprint arXiv:2502.10363 (2025)
41. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
42. Wang, Y., Yang, M., Ding, Z., Zhang, Y., Zeng, W., Xu, X., Jiang, H., Lu, Z.: From experts to a generalist: Toward general whole-body control for humanoid robots. arXiv preprint arXiv:2506.12779 (2025)

43. Weng, H., Li, Y., Sobanbabu, N., Wang, Z., Luo, Z., He, T., Ramanan, D., Shi, G.: Hdmi: Learning interactive humanoid whole-body control from human videos. arXiv preprint arXiv:2509.16757 (2025)
44. Xie, W., Han, J., Zheng, J., Li, H., Liu, X., Shi, J., Zhang, W., Bai, C., Li, X.: Kungfubot: Physics-based humanoid whole-body control for learning highly-dynamic skills. arXiv preprint arXiv:2506.12851 (2025)
45. Xue, H., Huang, X., Niu, D., Liao, Q., Kragerud, T., Gravdahl, J.T., Peng, X.B., Shi, G., Darrell, T., Sreenath, K., Sastry, S.: Leverb: Humanoid whole-body control with latent vision-language instruction (2025), https://arxiv.org/abs/2506.13751
46. Yin, K., Zeng, W., Fan, K., Dai, M., Wang, Z., Zhang, Q., Tian, Z., Wang, J., Pang, J., Zhang, W.: Unitracker: Learning universal whole-body motion tracker for humanoid robots. arXiv preprint arXiv:2507.07356 (2025)
47. Yuan, X., Li, Z., Lyu, B., Zuo, K., Lu, Y., Li, G., Yang, J.: Roboforge: Physically optimized text-guided whole-body locomotion for humanoids. arXiv preprint arXiv:2603.17927 (2026)
48. Yue, J., Wang, Z., Wang, Y., Zeng, W., Wang, J., Xu, X., Zhang, Y., Zheng, S., Ding, Z., Lu, Z.: Rl from physical feedback: Aligning large motion models with humanoid control. arXiv preprint arXiv:2506.12769 (2025)
49. Ze, Y., Chen, Z., Araújo, J.P., Cao, Z.a., Peng, X.B., Wu, J., Liu, C.K.: Twist: Teleoperated whole-body imitation system. arXiv preprint arXiv:2505.02833 (2025)
50. Zhang, T., Zheng, B., Nai, R., Hu, Y., Wang, Y.J., Chen, G., Lin, F., Li, J., Hong, C., Sreenath, K., et al.: Hub: Learning extreme humanoid balance. arXiv preprint arXiv:2505.07294 (2025)
51. Zhang, Y., Yuan, Y., Gurunath, P., He, T., Omidshafiei, S., Agha-mohammadi, A.a., Vazquez-Chanlatte, M., Pedersen, L., Shi, G.: Falcon: Learning force-adaptive humanoid loco-manipulation. arXiv preprint arXiv:2505.06776 (2025)
52. Zhou, E., An, J., Chi, C., Han, Y., Rong, S., Zhang, C., Wang, P., Wang, Z., Huang, T., Sheng, L., et al.: Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. arXiv preprint arXiv:2506.04308 (2025)
53. Zhou, E., Chi, C., Li, Y., An, J., Zhang, J., Rong, S., Han, Y., Ji, Y., Liu, M., Wang, P., et al.: Robotracer: Mastering spatial trace with reasoning in vision-language models for robotics. arXiv preprint arXiv:2512.13660 (2025)