KIMODO-POLICY: FROM TEXT-TO-MOTION GENERATORS TO HUMANOID VISION-LANGUAGE-ACTION POLICIES

Anonymous authors
Paper under double-blind review

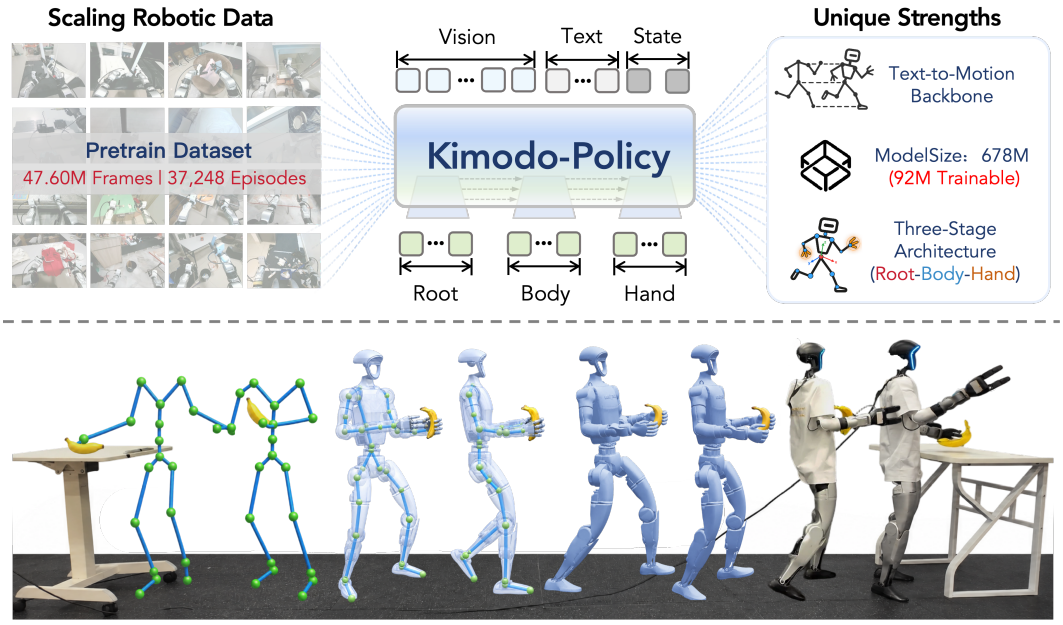

Figure 1: **Overview of Kimodo-Policy.** Top: diverse VLA pretraining adapts T2M priors through three-stage root–body–hand generation for visually grounded control. Bottom: skeletal history conditions action-chunk prediction; real-world examples illustrate coordinated whole-body execution.

ABSTRACT

Humanoid vision–language–action (VLA) models translate multimodal task understanding into coordinated whole-body behavior. Existing approaches equip pretrained vision–language or video generation backbones with action experts, but their pretraining does not directly establish whole-body motion priors. Unified or upper–lower-body action designs also lack dedicated root, body, and hand modeling despite their distinct requirements. These limitations increase the burden of learning precise, coordinated actions from limited task-specific humanoid demonstrations. We introduce Kimodo-Policy, which initializes visually grounded control with pretrained text-to-motion (T2M) priors over inter-joint coordination and temporal structure. Its three-stage root–body–hand architecture combines dedicated generators with explicit dependencies: root trajectories condition body generation, while future body features and visual observations guide hand control. We pretrain the policy on diverse humanoid demonstrations, conditioning actions on visual observations, language instructions, and motion history. With 678M total and only 92M trainable parameters, Kimodo-Policy achieves strong performance on HumanoidArena, SIMPLE, and real-world loco-manipulation tasks. Further experiments show performance gains from larger VLA pretraining datasets, greater model capacity, and stronger pretrained T2M priors.

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

1 INTRODUCTION

Vision–language–action (VLA) models integrate visual perception, language understanding, and action generation across manipulation tasks and robot embodiments (Brohan et al., 2023; Kim et al., 2024; Black et al., 2025b; Wang et al., 2026b; 2025). Learning instruction-conditioned behaviors from demonstrations provides a unified mapping from multimodal inputs to robot actions. Extending VLA learning to humanoids could broaden task coverage in human environments, where locomotion, posture, and manipulation must be coordinated (Bjorck et al., 2025; Wei et al., 2026a; Hu et al., 2026). For example, transferring objects between tables requires grasping, turning, walking, and placement while maintaining balance (Fig. 1, bottom). Such tasks require grounding task intent and visual feedback in coherent whole-body actions, motivating the extension of robot policy architectures to humanoid loco-manipulation.

Humanoid policy learning commonly extends the backbone–action expert paradigm from manipulation to whole-body control. Pretrained vision–language models encode instruction and observation semantics (Fig. 2(a)) (Bjorck et al., 2025; Wei et al., 2026a; Hu et al., 2026), while video generators supply spatiotemporal priors for scene evolution (Fig. 2(b)) (Zheng et al., 2026; Li et al., 2026). These representations condition action generation across root, body, and hand motion. Unified predictors generate whole-body sequences (Fig. 2(d)) (Hu et al., 2026; Zheng et al., 2026); hybrid designs combine upper-body joint targets with lower-body locomotion commands executed by dedicated controllers (Fig. 2(e)) (Zhong et al., 2025; Jiang et al., 2025). The former captures cross-component dependencies; the latter couples upper-body control with learned locomotion and stabilization. Both transfer pretrained semantic and visual dynamics knowledge to humanoid loco-manipulation.

Despite these advances, two challenges remain. ❶ **Insufficient Whole-Body Motion Priors.** Compared with fixed-base manipulation, humanoid control requires higher-dimensional inter-joint and temporal coordination. Backbone pretraining (Bjorck et al., 2025; Wei et al., 2026a; Zheng et al., 2026; Li et al., 2026) does not directly equip action generators with priors aligned with target humanoid kinematics, leaving them to learn these priors, often from limited task-specific demonstrations. ❷ **Insufficient Differentiation of Motion Components.** Root, body, and hand control differ in scale and precision. Unified prediction (Hu et al., 2026; Zheng et al., 2026) may create competing objectives; hybrid execution (Zhong et al., 2025; Jiang et al., 2025) does not itself provide dedicated root, body, and hand generators. Both must accommodate component-specific requirements while maintaining whole-body coordination.

Recent text-to-motion (T2M) methods learn structured motion distributions conditioned on text or spatial constraints (Petrovich et al., 2022; Tevet et al., 2023; Chen et al., 2023; Zhang et al., 2023; Guo et al., 2024; Xie et al., 2024). Building on these advances, **Kimodo-Policy** combines a pre-

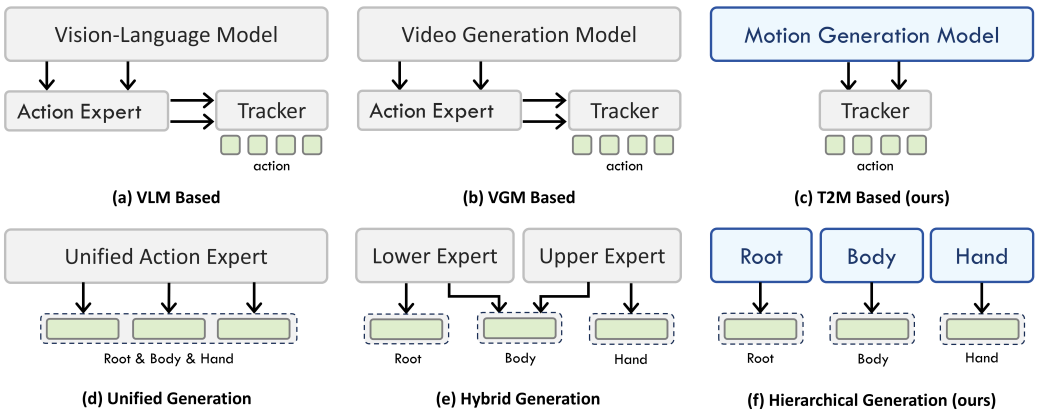

Figure 2: **Comparison of Humanoid Policy Architectures.** Panels (a)–(c) contrast VLM-based, VGM-based, and our T2M-based policy backbones. Panels (d)–(f) illustrate unified whole-body generation, hybrid generation of upper-body joint targets and lower-body locomotion commands, and our hierarchical root–body–hand generation, respectively.

trained T2M backbone with three-stage root–body–hand generation for humanoid VLA control. ❶ **Strong Whole-Body Motion Priors.** We transfer Kimodo's pretrained inter-joint and temporal priors to high-dimensional, visually grounded control (Fig. 2(c)) (Rempe et al., 2026), easing motion learning from limited task-specific humanoid demonstrations. ❷ **Strong Component-Specific Motion Modeling.** Dedicated generators (Fig. 2(f)) predict root translation and heading, root-conditioned body motion, and hand controls conditioned on future body features and visual observations. Their dependencies support component specialization and whole-body coordination. We pretrain on diverse humanoid demonstrations to generate actions conditioned on visual observations, language instructions, and motion history. With 678M total and only 92M trainable parameters, Kimodo-Policy shows consistent gains from more VLA pretraining data, larger models, and stronger T2M priors. Our main contributions are:

- We introduce **Kimodo-Policy**, a humanoid VLA framework transferring pretrained T2M priors to whole-body control. Pretraining on diverse humanoid demonstrations conditions closed-loop actions on visual observations, language instructions, and motion history.
- We propose dedicated root, body, and hand generators in a three-stage architecture to meet distinct control requirements. Explicit cross-stage conditioning preserves motion dependencies, supporting specialized motion generation and coordinated whole-body control.
- We demonstrate strong performance on HumanoidArena, SIMPLE, and real-world humanoid tasks using 678M parameters (only 92M trainable), with consistent gains from larger VLA pretraining datasets, greater model capacity, and stronger T2M priors.

2 Related Work

2.1 Humanoid Vision-Language-Action Models

Humanoid VLA models typically use pretrained vision–language models (VLMs) or video generation models (VGMs). VLM-based methods transfer visual-semantic knowledge to actions (Bjorck et al., 2025; Wei et al., 2026a; Zhong et al., 2025; Hu et al., 2026): GR00T N1 (Bjorck et al., 2025) couples a VLM with a diffusion action model, while Ψ_0 (Wei et al., 2026a) transfers representations from human egocentric demonstrations to a humanoid action expert. VGM-based methods exploit visual dynamics priors (Zheng et al., 2026; Li et al., 2026): MotionWAM (Zheng et al., 2026) generates whole-body motion from intermediate video features; WholeBodyWAM (Li et al., 2026) grounds world–action priors in whole-body controllers. Unified generators (Hu et al., 2026; Zheng et al., 2026) share prediction across motion components; hybrids (Zhong et al., 2025; Jiang et al., 2025) combine upper-body targets with locomotion commands. HAF (Gu et al., 2026) stages generation with a shared action expert, expanding the active action set and refining earlier predictions. Kimodo-Policy instead transfers a T2M backbone with pretrained inter-joint and temporal priors, reducing the need to learn motion regularities from humanoid demonstrations. Dedicated root, body, and hand generators meet distinct control requirements, while cross-stage conditioning supports coordinated, visually grounded whole-body behavior.

2.2 Text-to-Motion (T2M)

T2M includes TEMOS's variational language–motion modeling (Petrovich et al., 2022), pose-space MDM (Tevet et al., 2023) and latent-space MLD (Chen et al., 2023) diffusion, and autoregressive T2M-GPT (Zhang et al., 2023) and masked MoMask (Guo et al., 2024) quantized token generation. OmniControl (Xie et al., 2024) adds spatial constraints; Kimodo (Rempe et al., 2026) trains controllable root–body diffusion hierarchies at scale for human and humanoid skeletons. Humanoid-VLA (Ding et al., 2025) and ZeroWBC (Yang et al., 2026) learn autoregressive motion priors in pretrained language models and VLMs, respectively, before visual integration or task adaptation. MotionWAM (Zheng et al., 2026) adapts a pretrained video world model via staged video–motion training; WholeBodyWAM (Zhang et al., 2026) integrates a pretrained motion expert with video and action experts. Kimodo-Policy directly reuses a diffusion T2M backbone pretrained at scale in robot kinematic space. Dedicated root–body–hand generators ground inter-joint and temporal priors in vision and motion history for closed-loop loco-manipulation, reducing relearning from robot demonstrations while accommodating distinct control requirements.

3 METHOD: KIMODO-POLICY

Kimodo-Policy adapts Kimodo's pretrained motion prior for visually grounded, closed-loop humanoid control. We describe its root–body–hand architecture and visual conditioning (Sec. 3.1), motion-prior transfer and history-conditioned control (Sec. 3.2), and training and inference (Sec. 3.3). Background on Kimodo and ControlNet is provided in Appendix A.

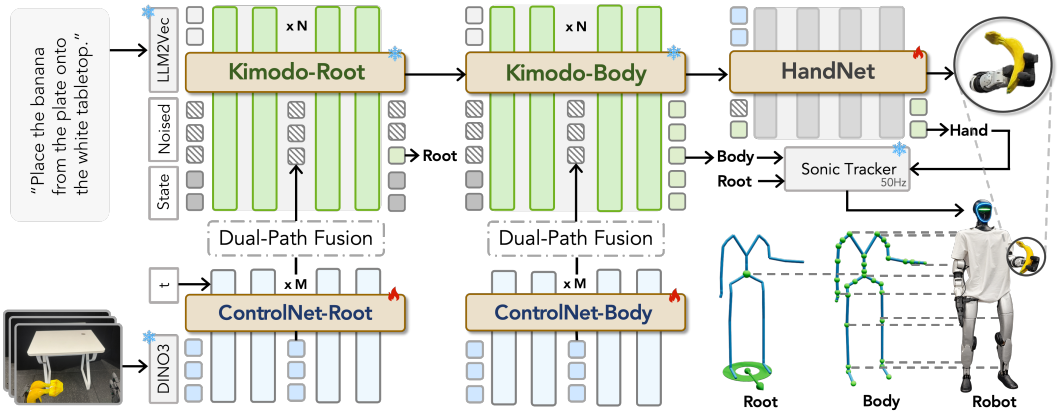

Figure 3: **Kimodo-Policy architecture.** Colored blocks: ▯ Kimodo denoisers, ▯ DINOv3-conditioned (Siméoni et al., 2025) ControlNet, and ▯ HandNet. Dashed boxes show Dual-Path Fusion: motion-query cross-attention and spatial-to-temporal MLPs. Root predictions condition body generation; future body features guide bilateral hand control. The Motion Projector maps root–body outputs to tracker references. Snowflakes/flames mark frozen/trainable modules.

3.1 MODEL ARCHITECTURE

Problem Formulation. We formulate language-conditioned humanoid loco-manipulation as closed-loop prediction of short action chunks. At step k, the policy receives instruction ℓ, the current egocentric RGB image I_k, available motion history $\mathcal{H}_k = [\mathbf{x}_{k-L+1}, \ldots, \mathbf{x}_k]$ in Kimodo's representation, and the current bilateral hand state $\mathbf{s}_k^h$. It predicts a K-frame chunk, with $\mathbf{a}_i = [\mathbf{r}_i, \mathbf{b}_i, \mathbf{h}_i]$ encoding root motion, articulated body motion, and bilateral hand controls:

$$\mathbf{A}_{k+1:k+K} \sim \pi_\theta(\cdot \mid I_k, \ell, \mathcal{H}_k, \mathbf{s}_k^h). \tag{1}$$

Root and body outputs provide kinematic references to a low-level motion tracker (Ze et al., 2025; Guan et al., 2026), while hand outputs specify synchronized binary or continuous closure commands according to the task interface. A chunk prefix is executed before replanning from updated observations, combining visual feedback with temporal context from observed history.

Three-Stage Architecture. As shown in Figure 3, Kimodo-Policy extends Kimodo's (Rempe et al., 2026) pretrained generator with dedicated root, body, and hand stages. LLM2Vec (BehnamGhader et al., 2024) encodes instructions; frozen DINOv3 (Siméoni et al., 2025) extracts visual features. Trainable ControlNet-Root and ControlNet-Body inject these features via Dual-Path Fusion into their corresponding frozen Kimodo denoisers, which also condition on instruction embeddings and motion history. The root stage predicts global translation and heading; their local representation conditions articulated body generation. Trainable HandNet predicts bilateral hand controls from future body features, visual features, and the current bilateral hand state. All root-to-body conditioning is detached, blocking body-stage gradients to Kimodo-Root and ControlNet-Root. HandNet's body and visual inputs are also detached, so hand supervision cannot update Kimodo-Body or either ControlNet branch. These boundaries isolate stage optimization while preserving forward dependencies. Root–body predictions yield whole-body tracking references executed with hand commands.

Dual-Path Fusion. DINOv3 (Siméoni et al., 2025) encodes the current egocentric image into spatial tokens for trainable root and body ControlNet branches. At selected layers of each Kimodo denoiser, future-motion hidden states query these tokens through cross-attention, while a token-axis MLP maps the image-token sequence to the future temporal horizon. Summed outputs pass through a layer-specific zero-initialized projection, producing a residual hint for the corresponding Kimodo layer. Hints are added only to future-motion tokens, leaving the constrained history prefix unchanged. Zero initialization preserves pretrained behavior at adaptation onset; both paths ground the frozen prior through content-selective visual retrieval and temporally aligned conditioning.

HandNet. Hand control specifies opening or closing rather than kinematic pose trajectories, motivating a dedicated bilateral diffusion head. At each denoising step, HandNet jointly processes the current hand state, noisy future hand controls, future body hidden states, and visual tokens to predict two hand-control values per future frame. Conditioning on predicted body features and the current scene aligns hand commands with whole-body action, while the separate head keeps hand labels outside the root–body motion space. It supports binary open/close states and continuous closure changes. During training, body and visual features are detached: hand supervision updates HandNet without gradients to the motion denoiser or ControlNet.

3.2 Adapting Motion Priors for Closed-Loop Control

Motion Prior Transfer. Figure 4 illustrates how a pretrained motion prior supplies temporal structure and whole-body coordination for visually grounded control. Given only "Sit on the sofa," Kimodo (Rempe et al., 2026) generates a coherent sequence ending in a seated posture (top), but lacks grounding in the current scene and robot state. Kimodo-Policy freezes this backbone and injects egocentric visual features through trainable ControlNet branches; motion-history constraints anchor predictions to the ongoing trajectory, enabling the same task from visual observations and state history (bottom). Task-specific learning thus adapts reusable behavior-level structure to the instruction, scene, and state, reducing the need to relearn basic motion patterns from robot demonstrations while preserving coherent closed-loop execution.

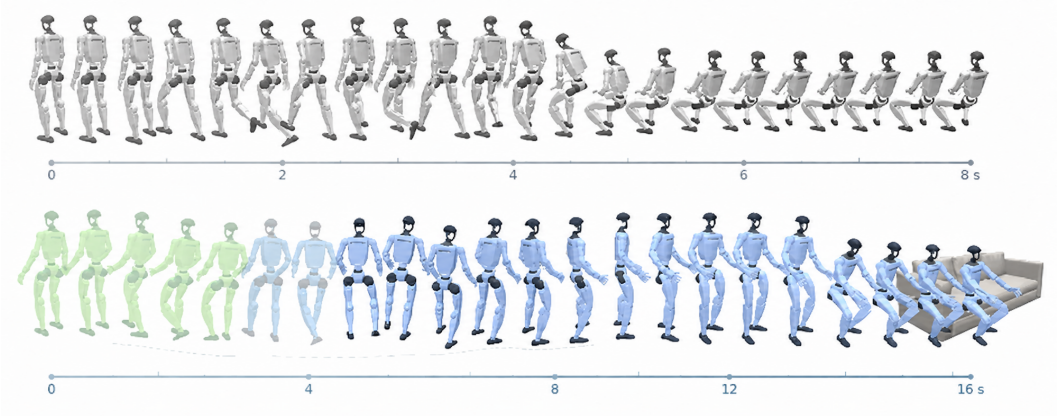

Figure 4: **Motion-Prior Transfer.** Top: frozen Kimodo generates a complete "Sit on the sofa" sequence from text. Bottom: Kimodo-Policy predicts short chunks from vision and motion history for closed-loop control. Light brown: observed history; blue: predicted and executed motion.

Motion-History Conditioning. Fixed T2M sequences cannot respond to deviations between predicted and measured motion or scene changes. We therefore formulate each closed-loop query as constrained motion completion in Kimodo's kinematic representation. Observed history forms the prefix of a history–future window, with feature-wise masks identifying available kinematic values. The prefix is restored before each denoising step; the unconstrained future becomes the next action chunk. Trainable ControlNet branches inject current egocentric image features into Kimodo's root and body denoisers, grounding predictions while retaining the pretrained root–body structure. The tracker executes a chunk prefix; updated history and images condition the next query. This receding-horizon loop preserves the motion prior with continual visual and state-based correction.

3.3 Training and Inference

Training Objective. ControlNet branches and HandNet denoise demonstration windows. At sampled diffusion timesteps, Gaussian noise perturbs normalized root–body features and hand targets; feature-wise masks constrain observed history, with supervision on valid future frames:

$$\mathcal{L} = \lambda_r \mathcal{L}_r + \lambda_b \mathcal{L}_b + \lambda_h \big(\mathcal{L}_h + \lambda_\Delta \mathcal{L}_\Delta\big), \tag{2}$$

The masked MSE terms $\mathcal{L}_r$, $\mathcal{L}_b$, and $\mathcal{L}_h$ reconstruct clean normalized root–body features and hand targets. Binary hand targets use $[-1, 1]$; continuous targets encode closure residuals from the latest valid observed hand state. $\mathcal{L}_\Delta$ computes ℓ_1 error in predicted versus demonstrated consecutive hand-control changes, including the observed-to-first-future transition; continuous residuals are first converted to closure values. All losses mask padding; hand losses also mask unavailable labels, and transitions require two valid adjacent labels. The main MSE weights are $(\lambda_r, \lambda_b, \lambda_h, \lambda_\Delta) = (4, 1, 1, 0.2)$. Predicted root motion and, in the main configuration, root ControlNet features are detached before body generation; HandNet receives detached body and visual features. Kimodo denoisers and the DINOv3 (Siméoni et al., 2025) encoder remain frozen.

Inference. Each policy query encodes the language instruction and current egocentric image, and represents available proprioceptive history in Kimodo's motion space with frame- and feature-wise observation masks. DDIM (Song et al., 2021) iteratively denoises future motion and hand sequences initialized from Gaussian noise. Each step restores observed history, injects visual features into frozen root and body denoisers via ControlNet, conditions body generation on predicted root motion, and predicts hand controls with HandNet using the current hand state, visual features, and future body features. After denoising, we invert the motion representation and map root and joint trajectories, together with bilateral hand commands, to tracker-compatible references. The low-level whole-body tracker executes a chunk prefix; the policy replans from updated images and proprioceptive states, combining structured motion generation with continual closed-loop correction.

4 Experimental Results

We evaluate Kimodo-Policy through single- and multi-task HumanoidArena control (Sec. 4.1), SIMPLE manipulation (Sec. 4.2), and real-world deployment (Sec. 4.3). Analyses cover T2M backbone quality (Sec. 4.4), data and model scaling (Sec. 4.5), and design ablations (Sec. 4.6). Appendix B details the datasets, training protocols, and evaluation environments.

4.1 Evaluation on the HumanoidArena Benchmark

HumanoidArena (Wang et al., 2026a) covers seven language-conditioned tasks spanning human–object interaction (HOI) and human–scene interaction (HSI). We evaluate independent per-task fine-tuning and joint training across SONIC tasks. Appendix C.1 (Fig. 11) visualizes the tasks.

Single-task results. We fine-tune one policy per task with RTC disabled to match baselines. Kimodo-Policy uses direct task fine-tuning; Kimodo-Policy* initializes ControlNet and the hand head from our 419-hour VLA pretraining. Kimodo-Policy achieves the highest overall success rate (SR; 69.51%), exceeding the strongest baseline, DP (Chi et al., 2023) (60.0%), by 9.51 percentage points. It also leads in average HSI SR (84.58%) and ties Kimodo-Policy* for lowest average fall rate (AFR; 4.29%). Kimodo-Policy* achieves 67.14% overall and 84.17% average HSI SR, leading OpenDoor (93.3%) and VisNavi (76.7%). VLA pretraining improves selected tasks, while direct fine-tuning yields higher overall SR.

Table 1: ***Single-task HumanoidArena results.*** A separate policy is fine-tuned per task. For Tables 1–2, values are mean ± standard deviation; * marks VLA-pretrained initialization, and bold/underlined entries mark best/second-best results.

Method	AFR (↓)	Human–Object Interaction			Human–Scene Interaction				Overall (↑)
		Football	DoubleDesk	P&Pbox	OpenDoor	SitSofa	Boxing	VisNavi	
ACT (Zhao et al., 2023)	8.57%	16.7±6.2%	18.3±4.7%	56.7±6.2%	78.3±9.4%	73.3±8.5%	56.7±6.2%	33.3±2.4%	47.61%
DP (Chi et al., 2023)	8.33%	**45.0±10.8%**	36.7±4.7%	75.0±4.1%	85.0±10.8%	78.3±14.3%	**76.7±2.4%**	23.3±6.2%	60%
FM (Lipman et al., 2023)	5.71%	13.3±2.4%	38.3±4.7%	73.3±6.2%	70.0±4.1%	15.0±14.7%	70.0±8.2%	38.3±14.3%	45.46%
$\pi_{0.5}$ (Black et al., 2025a)	<u>5.24%</u>	10.0±4.1%	**43.3±6.2%**	71.7±11.8%	66.7±6.2%	73.3±2.4%	70.0±0.0%	23.3±6.2%	51.19%
Ours	**4.29%**	<u>28.3±6.2%</u>	<u>40.0±0.0%</u>	**80.0±4.1%**	<u>90.0±4.1%</u>	**100.0±0.0%**	<u>73.3±13.1%</u>	<u>75.0±10.8%</u>	**69.51%**
Ours*	**4.29%**	26.7±4.7%	28.3±2.4%	<u>78.3±8.5%</u>	**93.3±2.4%**	<u>96.7±2.4%</u>	70.0±8.2%	**76.7±18.9%**	<u>67.14%</u>

Multi-task results. We jointly train one policy on all seven HumanoidArena tasks and evaluate under SONIC (Wang et al., 2026a; Luo et al., 2026), with RTC disabled to match baselines. Table 2 shows Kimodo-Policy* reaching 72.14% overall SR, exceeding the strongest baseline, $\pi_{0.5}$ (Black et al., 2025a) (53.8%), by 18.33 percentage points, with lowest AFR (4.29%). It leads HOI (56.11%) and HSI (84.17%) averages, P&Pbox (90.0%), OpenDoor (95.0%), SitSofa (96.7%, tied), and VisNavi (76.7%). Without VLA pretraining, Kimodo-Policy reaches 67.63% overall and 51.67% HOI SR; overall SR falls only 1.88 percentage points from single-task training. Thus, a shared policy can retain much of its single-task performance with high success and low fall rates.

Table 2: ***Multi-task HumanoidArena results.*** One policy is jointly trained on all seven tasks.

Method	AFR (↓)	Human–Object Interaction			Human–Scene Interaction				Overall (↑)
		Football	DoubleDesk	P&Pbox	OpenDoor	SitSofa	Boxing	VisNavi	
ACT (Zhao et al., 2023)	7.4%	20.0±4.1%	30.0±4.1%	60.0±12.2%	58.3±2.4%	21.7±2.4%	58.3±10.3%	28.3±4.7%	39.5%
DP (Chi et al., 2023)	7.1%	**53.3±6.2%**	35.0±10.8%	0.0±0.0%	83.3±6.2%	83.3±4.7%	21.7±30.6%	50.0±14.7%	46.7%
FM (Lipman et al., 2023)	7.1%	21.7±8.5%	20.0±7.1%	0.0±0.0%	76.7±2.4%	48.3±6.2%	18.3±25.9%	38.3±6.2%	31.9%
$\pi_{0.5}$ (Black et al., 2025a)	6.4%	10.0±7.1%	30.0±7.1%	75.0±4.1%	85.0±4.1%	70.0±14.7%	**71.7±12.5%**	35.0±4.1%	53.8%
Ours	5.0%	35.0±7.1%	**40.0±0.0%**	80.0±4.1%	93.3±6.2%	**96.7±4.7%**	66.7±4.7%	61.7±6.2%	67.6%
Ours*	**4.3%**	40.0±4.1%	38.3±2.4%	**90.0±4.1%**	**95.0±4.1%**	**96.7±4.7%**	68.3±10.3%	**76.7±8.5%**	**72.1%**

Attention map analysis. Figure 5 shows Kimodo-Policy focusing on the ball in Football, box in P&P Box, and punching bag in Boxing, alongside relevant scene elements for door opening, sitting, and navigation. $\pi_{0.5}$ responds less locally, while Diffusion Policy activates broader scene regions. This task-dependent focus suggests alignment with instruction semantics and use of task-relevant cues for shared multi-task action selection.

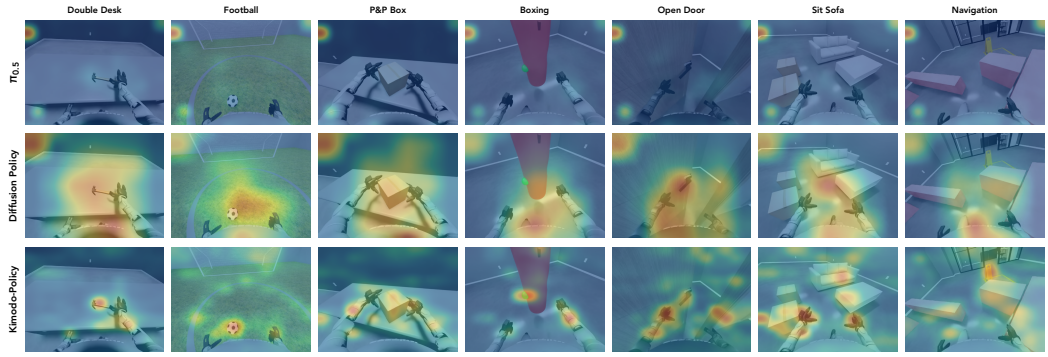

Figure 5: **Task-Conditioned Visual Attention.** Columns show seven HumanoidArena tasks; rows (top to bottom) show $\pi_{0.5}$, Diffusion Policy, and Kimodo-Policy. Kimodo-Policy shifts attention toward task-relevant objects and interaction regions.

4.2 Evaluation on the Simple Benchmark

We further evaluate Kimodo-Policy on the SIMPLE benchmark (Wei et al., 2026b), using the L0–L2 protocol with progressively stronger visual, lighting, and target-pose randomization (Wei et al., 2026b). We fine-tune an independent policy for each of the eight benchmark tasks and report the results in Table 3. This protocol isolates skill acquisition while testing robustness under increasing distribution shifts. See Appendix C.2 (Fig. 12) for task visualizations.

Single-task results. Table 3 evaluates VLA-pretrained Kimodo-Policy (Ours*) on eight SIMPLE tasks, aggregating L0–L2 success rates with RTC disabled to match the baselines. It achieves best or joint-best results on six tasks: Handover (83% vs. 80% for the strongest baseline), Mobile P&P (70% vs. 60%), OpenOven (57% vs. 53%), OpenFaucet (53% vs. 33%), CloseDoor and Push Office Chair (both 100%, tied with Ψ_0). It ranks second on XMoveBendPick (80% vs. 93% for Ψ_0) and OpenTrashCan (63% vs. 83%). These results show that the motion-prior-based, visually grounded root–body–hand policy supports handover, mobile manipulation, articulated-object interaction, and pushing under visual, lighting, and target-pose variation.

Table 3: ***Single-Task Results on SIMPLE.*** L0–L2 progressively increase visual and lighting variation and target-pose randomization. Entries report L0/L1/L2 success counts and aggregate success rates (parentheses); bold/underlined values mark the best/second-best aggregate rates.

Method	HANDOVER	MOBILE P&P	XMOVEBENDPICK	CLOSEDOOR	OPENOVEN	OPENFAUCET	PUSH OFFICE CHAIR	OPENTRASHCAN
Ψ_0 (Wei et al., 2026a)	7/7/10 (80%)	7/5/6 (60%)	**10/9/9 (93%)**	**10/10/10 (100%)**	7/5/4 (53%)	3/3/4 (33%)	**10/10/10 (100%)**	**7/9/9 (83%)**
Ours*	**9/8/8 (83%)**	**9/8/4 (70.0%)**	10/7/7 (80%)	**10/10/10 (100%)**	**8/4/5 (57%)**	**5/6/5 (53%)**	**10/10/10 (100%)**	9/4/6 (63%)

4.3 EVALUATION ON THE REAL-WORLD HUMANOID ROBOT

We test a physical humanoid on eight language-conditioned loco-manipulation tasks under unconstrained sensing and execution, covering locomotion, balance, single- and dual-arm coordination, and contact-rich manipulation. Each task has ten trials scored by successful completions; Appendix C.3 (Fig. 13) provides visualizations.

Real-world results. Scratch-trained Kimodo-Policy averages 45% success, above GR00T N1.6 (Bjorck et al., 2025) (31%) and Ψ_0 (Wei et al., 2026a) (35%; Fig. 6). VLA pretraining raises this to 60% (+15 points), improving seven tasks and matching Cart Pushing (60%). Its strongest results are Punching Bag (90%), Cushion Sitting (80%), and Dual-Arm PNP (70%).

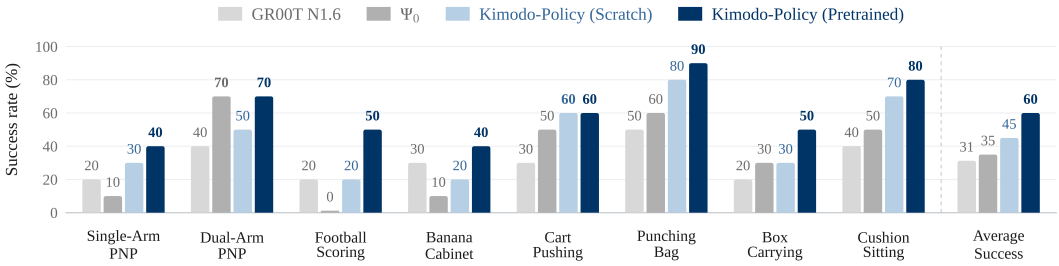

Figure 6: **Real-World Success Rates.** GR00T N1.6, Ψ_0, and Kimodo-Policy with/without VLA pretraining are compared on eight tasks; the rightmost group reports mean success.

4.4 DOES BETTER T2M IMPROVE VLA PERFORMANCE?

To test motion-prior quality, we compare Kimodo-G1-SEED-v1 and RP-v1 T2M backbones (Rempe et al., 2026) under identical policy architectures, adaptation procedures, and HumanoidArena protocols. Both have 282M parameters and 90%/10% train/test splits; SEED-v1 uses 288 h of BONES-SEED data and RP-v1 uses 700 h of Rigplay (Rempe et al., 2026).

Table 4: ***Effect of the T2M backbone on Kimodo-Policy.*** We compare two same-size Kimodo-G1 backbones using the same adaptation protocol. Data scales include the training and test splits; results are mean ± standard deviation over three seeds.

Base Model	Model Size	Data Scale	Human–Object Interaction			Human–Scene Interaction				Overall (↑)
			FOOTBALL	DOUBLEDESK	P&PBOX	OPENDOOR	SITSOFA	BOXING	VISNAVI	
Kimodo-G1-SEED-v1	282 M	288 h	**50.00±16.33%**	33.33±15.46%	75.00±10.80%	60.00±4.08%	90.00±4.08%	**71.67±8.50%**	**70.00±18.71%**	64.29%
Kimodo-G1-RP-v1	282 M	700 h	35.0±7.1%	**40.0±0.0%**	**80.0±4.1%**	**93.3±6.2%**	**96.7±4.7%**	66.7±4.7%	61.7±6.2%	**67.63%**

T2M transfer results. RP-v1 improves overall SR over SEED-v1 from 64.29% to 67.63% (+3.34 points; Table 4), with gains on DoubleDesk (33.33% → 40.0%), P&Pbox (75.0% → 80.0%), OpenDoor (60.0% → 93.3%), and SitSofa (90.0% → 96.7%), but declines on Football (50.0% → 35.0%), Boxing (71.67% → 66.7%), and VisNavi (70.0% → 61.7%). Gains favor human–scene interaction, with task-dependent trade-offs (Appendix D).

4.5 DOES PERFORMANCE SCALE WITH DATA AND MODEL SIZE?

We vary VLA policy-pretraining data duration with the architecture, SonicX7 adaptation, and evaluation fixed, then vary active policy parameters with the pretrained T2M backbone, downstream data, optimization, and evaluation fixed. RTC is disabled throughout to match the benchmark protocol.

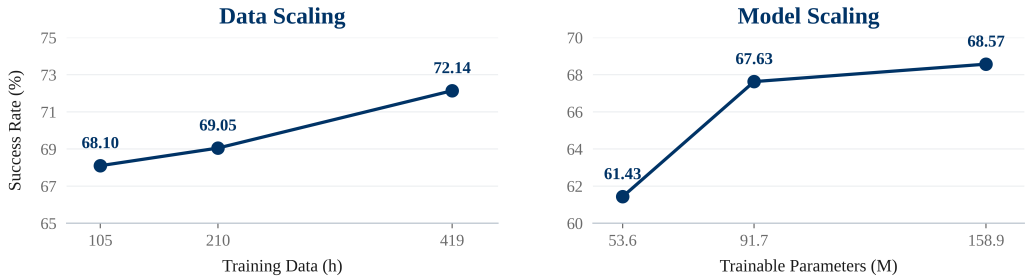

Figure 7: Scaling behavior of Kimodo-Policy. Left: overall success rate as a function of VLA policy-pretraining data duration. Right: overall success rate as a function of active policy parameters.

Scaling results. Increasing VLA pretraining data from 105 h to 210 h and 419 h raises overall success from 68.10% to 69.05% and 72.14%, respectively (+4.04 points). Scaling active policy parameters from 53.6M to 91.7M and 158.9M raises success from 61.43% to 67.63% and 68.57% (+7.14 points). Both controlled sweeps improve success over the evaluated ranges (Fig. 7); the capacity gain is smaller in the upper interval. Appendix E discusses these trends and their interpretation.

4.6 Ablation Study

All ablations train on mixed HumanoidArena tasks in SONIC X7 and evaluate on DoubleDesk, Football, and P&Pbox. Figure 8 reports training loss and mean success across these tasks.

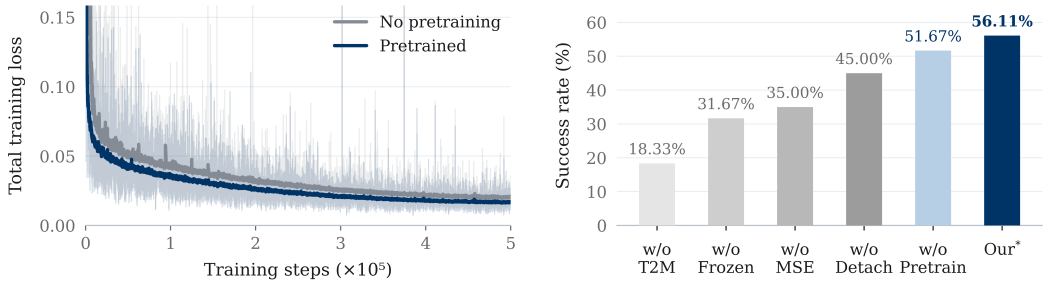

Figure 8: **Training Dynamics and Ablations.** Left: training loss with/without VLA pretraining. Right: mean success on DoubleDesk, Football, and P&Pbox. Gray: component and training-design ablations; light/dark blue: without/with VLA pretraining.

Results Analysis. Without VLA pretraining, Kimodo-Policy achieves 51.67% mean success; removing the T2M backbone, unfreezing it, replacing MSE with Kimodo's original motion objective, or removing root–body gradient detachment yields 18.33%, 31.67%, 35.00%, or 45.00%, respectively. These results support a frozen motion prior, MSE supervision, and gradient separation. VLA pretraining raises success from 51.67% to 56.11%, with generally lower and less variable training loss. Both retain the T2M backbone, indicating complementary benefits (Appendix F).

5 Conclusion

Kimodo-Policy adapts pretrained text-to-motion priors for visually grounded, closed-loop humanoid control through root–body–hand generation and continual visual and state feedback. Experiments show strong results on HumanoidArena, SIMPLE, and eight real-world tasks with 92M trainable parameters out of 678M total. Stronger priors, more VLA data, and greater capacity improve performance within the evaluated settings. Transfer beyond SONIC remains unverified; Appendix G details this limitation and future directions in controller compatibility, multimodal integration, motion pretraining, and direct transfer from human T2M backbones.

AI USE STATEMENT

Large language models (LLMs) supported the preparation of this manuscript through language editing and writing assistance. Their use included rephrasing sentences, correcting grammar, improving readability, and refining the organization and presentation of the text. LLMs also assisted in drafting and revising selected passages using technical descriptions, experimental findings, and arguments supplied and verified by the authors.

All research decisions and methodological choices were made by the authors, who conducted the experiments and data analyses and verified the resulting interpretations and conclusions. The authors reviewed all AI-assisted material and take full responsibility for the scientific content and final manuscript, including any text or ideas developed with AI assistance.

REFERENCES

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. LLM2Vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*, 2024. URL https://arxiv.org/abs/2404.05961.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. URL https://arxiv.org/abs/2503.14734.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025a. URL https://arxiv.org/abs/2504.16054.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. In *Robotics: Science and Systems*, 2025b. URL https://arxiv.org/abs/2410.24164.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. URL https://arxiv.org/abs/2307.15818.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18000–18010, 2023. doi: 10.1109/CVPR52729.2023.01726. URL https://doi.org/10.1109/CVPR52729.2023.01726.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems*, 2023. URL https://roboticsconference.org/2023/program/papers/026/.

Pengxiang Ding, Jianfei Ma, Xinyang Tong, Binghong Zou, Xinxin Luo, Yiguo Fan, Ting Wang, Hongchao Lu, Panzhong Mo, Jinxin Liu, et al. Humanoid-VLA: Towards universal humanoid control with visual integration. *arXiv preprint arXiv:2502.14795*, 2025. URL https://arxiv.org/abs/2502.14795.

Langzhe Gu, Chengkai Hou, Meng Li, Xinhua Wang, Jiaming Liu, Xinyuan Lv, Bowei Zhang, Shuanghao Bai, Guangrun Li, Jingyang He, et al. HAF: Adapting generalist VLAs to humanoid whole-body loco-manipulation via hierarchical action flow and spectral latent RL. *arXiv preprint arXiv:2608.16837*, 2026. URL https://arxiv.org/abs/2608.16837.

Yu Guan, Zekun Qi, Chenghuai Lin, Xuchuan Chen, Dairu Liu, Wenyao Zhang, Jilong Wang, Xinqiang Yu, He Wang, and Li Yi. LIMMT: Less is more for motion tracking. *arXiv preprint arXiv:2606.06953*, 2026. URL https://arxiv.org/abs/2606.06953.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative masked modeling of 3D human motions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1900–1910, 2024. doi: 10.1109/CVPR52733.2024.00186. URL https://doi.org/10.1109/CVPR52733.2024.00186.

Yingdong Hu, Haodong Zhu, Boyuan Zheng, Yihang Hu, Tong Zhang, Zunhao Chen, Junming Zhao, Ruiqian Nai, and Yang Gao. OpenHLM: An empirical recipe for whole-body humanoid loco-manipulation. *arXiv preprint arXiv:2606.22174*, 2026. URL https://arxiv.org/abs/2606.22174.

Haoran Jiang, Jin Chen, Qingwen Bu, Li Chen, Modi Shi, Yanjie Zhang, Delong Li, Chuanzhe Suo, Chuang Wang, Zhihui Peng, et al. WholeBodyVLA: Towards unified latent VLA for whole-body loco-manipulation control. *arXiv preprint arXiv:2512.11047*, 2025. URL https://arxiv.org/abs/2512.11047.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. URL https://arxiv.org/abs/2406.09246.

Zhuo Li, Yiming Yao, Jim Tan, Mengjie Jing, Zhipeng Dong, and Fei Chen. WholeBodyWAM: Generalizing pre-trained world-action priors to humanoid loco-manipulation via WBC-grounded coordination. *arXiv preprint arXiv:2609.16644*, 2026. URL https://arxiv.org/abs/2609.16644.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, Jinhyung Park, David Sami, Zi Wang, Xingye Da, Runyu Ding, Cyrus Hogg, Lina Song, Edy Lim, Eugene Jeong, Tairan He, Haoru Xue, Wenli Xiao, Simon Yuen, Jan Kautz, Yan Chang, Umar Iqbal, Linxi Fan, and Yuke Zhu. SONIC: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117): eaed4592, 2026. doi: 10.1126/scirobotics.aed4592. URL https://doi.org/10.1126/scirobotics.aed4592.

Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, pp. 480–497, 2022. doi: 10.1007/978-3-031-20047-2_28. URL https://doi.org/10.1007/978-3-031-20047-2_28.

Davis Rempe, Mathis Petrovich, Ye Yuan, Haotian Zhang, Xue Bin Peng, Yifeng Jiang, Tingwu Wang, Umar Iqbal, David Minor, Michael de Ruyter, et al. Kimodo: Scaling controllable human motion generation. *arXiv preprint arXiv:2603.15546*, 2026. URL https://arxiv.org/abs/2603.15546.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.14916.

Taowen Wang, Zikang Xie, Bin Yang, Yunheng Wang, Zizhao Yuan, Yuetong Fang, Yixiao Feng, Yichi Wang, Xingyu Chen, Haodong Chen, et al. HumanoidArena: Benchmarking egocentric hierarchical whole-body learning. *arXiv preprint arXiv:2606.17833*, 2026a. URL https://arxiv.org/abs/2606.17833.

Yunheng Wang, Yuetong Fang, Taowen Wang, Yixiao Feng, Yawen Tan, Shuning Zhang, Peiran Liu, Yiding Ji, and Renjing Xu. DreamNav: A trajectory-based imaginative framework for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2509.11197*, 2025. URL https://arxiv.org/abs/2509.11197.

Yunheng Wang, Yuetong Fang, Taowen Wang, Lusong Li, Kun Liu, Junzhe Xu, Zizhao Yuan, Yixiao Feng, Jiaxi Zhang, Wei Lu, Zecui Zeng, and Renjing Xu. What limits vision-and-language navigation? *arXiv preprint arXiv:2605.13328*, 2026b. URL https://arxiv.org/abs/2605.13328.

Songlin Wei, Hongyi Jing, Boqian Li, Zhenyu Zhao, Jiageng Mao, Zhenhao Ni, Sicheng He, Jie Liu, Xiawei Liu, Kaidi Kang, et al. Ψ_0: An open foundation model towards universal humanoid loco-manipulation. *arXiv preprint arXiv:2603.12263*, 2026a. URL https://arxiv.org/abs/2603.12263.

Songlin Wei, Zhenhao Ni, Jie Liu, Zhenyu Zhao, Junjie Ye, Hongyi Jing, Junkai Xia, Xiawei Liu, Michael Leong, Liang Heng, et al. SIMPLE: Simulation-based policy learning and evaluation for humanoid loco-manipulation. *arXiv preprint arXiv:2606.08278*, 2026b. URL https://arxiv.org/abs/2606.08278.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. OmniControl: Control any joint at any time for human motion generation. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.08580.

Haoran Yang, Jiacheng Bao, Yucheng Xin, Haoming Song, Yuyang Tian, Bin Zhao, Dong Wang, and Xuelong Li. ZeroWBC: Learning natural whole-body humanoid interaction from human egocentric data. *arXiv preprint arXiv:2603.09170*, 2026. URL https://arxiv.org/abs/2603.09170.

Yanjie Ze, Siheng Zhao, Weizhuo Wang, Angjoo Kanazawa, Rocky Duan, Pieter Abbeel, Guanya Shi, Jiajun Wu, and C. Karen Liu. TWIST2: Scalable, portable, and holistic humanoid data collection system. *arXiv preprint arXiv:2511.02832*, 2025. URL https://arxiv.org/abs/2511.02832.

Bowei Zhang, Qiyao Zhang, Shuanghao Bai, Xinhua Wang, Meng Li, Yilei Wang, Leiwang Zhang, Jian Tang, Lu Zhou, Lei Sun, et al. WholeBodyWAM: Learning whole-body world action models with scalable motion priors. *arXiv preprint arXiv:2609.18197*, 2026. URL https://arxiv.org/abs/2609.18197.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2M-GPT: Generating human motion from textual descriptions with discrete representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14730–14740, 2023. doi: 10.1109/CVPR52729.2023.01415. URL https://doi.org/10.1109/CVPR52729.2023.01415.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems*, 2023. URL https://roboticsconference.org/2023/program/papers/016/.

Jia Zheng, Teli Ma, Yudong Fan, Zifan Wang, Shuo Yang, and Junwei Liang. MotionWAM: Towards foundation world action models for real-time humanoid loco-manipulation. *arXiv preprint arXiv:2606.09215*, 2026. URL https://arxiv.org/abs/2606.09215.

Rui Zhong, Yizhe Sun, Junjie Wen, Jinming Li, Chuang Cheng, Wei Dai, Zhiwen Zeng, Huimin Lu, Yichen Zhu, and Yi Xu. HumanoidExo: Scalable whole-body humanoid manipulation via wearable exoskeleton. *arXiv preprint arXiv:2510.03022*, 2025. URL https://arxiv.org/abs/2510.03022.