WholeBodyWAM: Generalizing Pre-trained World–Action Priors to Humanoid Loco-Manipulation via WBC-Grounded Coordination

Zhuo Li[1,4†] Yiming Yao[2,4*] Jim Tan[3,4*] Mengjie Jing[1,4] Zhipeng Dong[1,4] Fei Chen[1,4‡]

[1]The Chinese University of Hong Kong [2]The University of Hong Kong
[3]Peking University [4]Φ-Institute

[†]Project Lead [‡]Corresponding Author [*]Equal Contribution

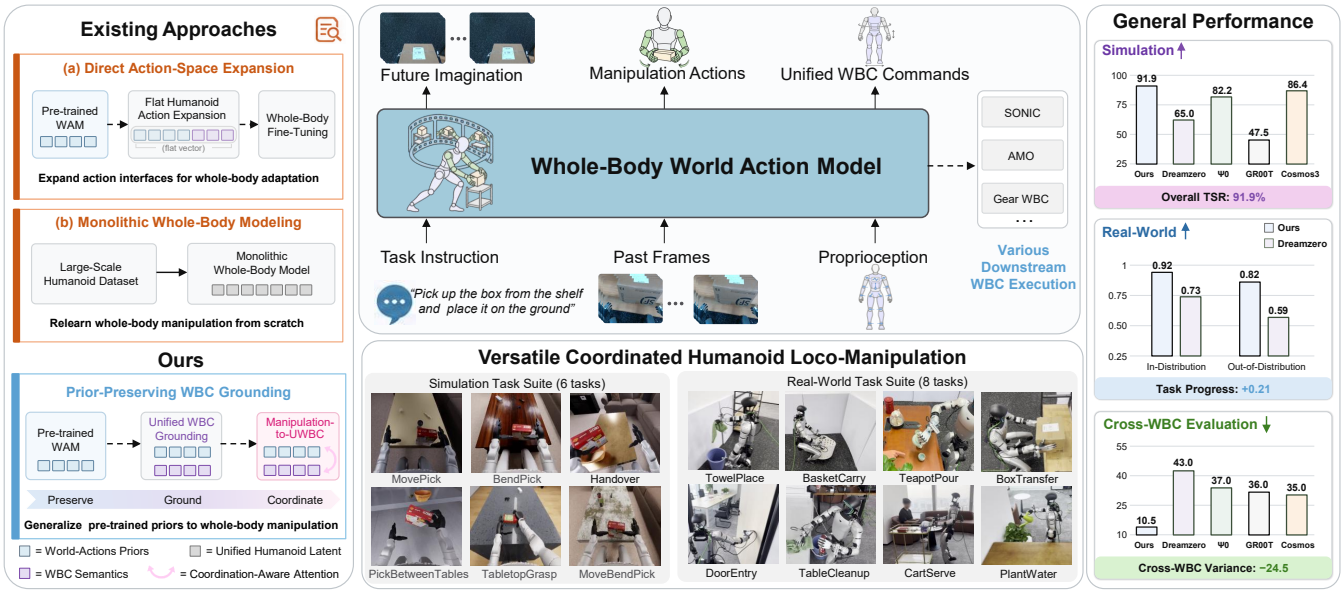

Fig. 1. **Overview of WholeBodyWAM.** Unlike existing approaches that expand action interfaces for whole-body adaptation or learn dedicated humanoid world–action mappings from scratch, WholeBodyWAM preserves and generalizes the world–action priors of pre-trained WAMs to humanoid loco-manipulation through WBC-grounded coordination. Given language, visual history, and proprioception, it jointly predicts future visual dynamics, manipulation intent, and unified whole-body controller (WBC) commands. Evaluations on six simulation and eight real-world tasks demonstrate higher overall task success rates and lower cross-controller performance variance than the evaluated baselines, supporting effective loco-manipulation across diverse tasks and heterogeneous WBCs.

***Abstract*— World Action Models (WAMs) offer a promising approach to general-purpose robot manipulation by jointly modeling visual dynamics and actions. However, most WAM studies focus on tabletop or arm-centric manipulation, while humanoid loco-manipulation remains less explored. To address this gap, we introduce WholeBodyWAM, which jointly predicts future visual dynamics, manipulation actions, and whole-body control intents for generalizable humanoid loco-manipulation. It preserves pre-trained world–action priors while grounding heterogeneous whole-body controller (WBC) semantics and coordinating whole-body behavior. Extensive experiments show that WholeBodyWAM achieves an overall simulation task success rate of 91.9%, with a 0.23 improvement in real-world out-of-distribution task progress and a 70% reduction in success-rate variance across WBCs relative to the respective baselines. These results suggest a path toward scalable humanoid whole-body intelligence by extending pre-trained world–action priors through structured WBC grounding and coordination, rather than relearning whole-body behavior from scratch. Project page: `https://wholebodywam.github.io/`.**

I. Introduction

World Action Models (WAMs) jointly model robot actions and future visual dynamics, providing reusable priors for general robot manipulation [1]–[3]. Most existing studies investigate local hand–object interactions in tabletop or arm-centric settings [3]–[6], while wide-range, whole-body loco-manipulation on bipedal humanoids remains underexplored. This gap stems from three coupled challenges: *Whole-body coordination complexity.* Successful humanoid loco-manipulation depends not only on generating plausible hand–object interactions but also on modeling when and how whole-body motion should coordinate with manipulation. *Heterogeneous WBC semantics.* Existing WBCs expose command ontologies with different physical meanings [7]–[9], resulting in similar motions requiring different combinations of task-space commands and joint targets. This heterogeneity hinders reusable action semantics and consistent action–dynamics modeling across controllers. *Humanoid data scarcity.* Despite

the availability of large-scale robot data for WAM pretraining, paired humanoid loco-manipulation demonstrations remain scarce [10]–[12]. Such scarcity constrains direct supervision for modeling the coordination between manipulation intent and the whole-body adjustments required for execution.

Recent studies have begun exploring WAMs for humanoid loco-manipulation [2], [13], [14]. These approaches fall into two broad categories. *1) Direct action-space expansion:* One straightforward strategy is to expand the action interface of a pre-trained tabletop WAM and fine-tune the model on whole-body demonstrations for humanoid adaptation. DiT4DiT [2] demonstrates related fine-tuning of video–action models for humanoid manipulation. Although this strategy preserves transferable world–action priors, it leaves heterogeneous WBC semantics and complex whole-body coordination to be learned implicitly from limited fine-tuning supervision. Consequently, the model may retain high-level manipulation intent yet produce poorly grounded WBC commands and temporally inconsistent whole-body behavior, limiting learning efficiency and generalization on coordination-intensive tasks. *2) Monolithic whole-body modeling:* An alternative approach learns humanoid-native whole-body WAMs that unify locomotion and manipulation within controller-compatible latents [13], [14]. Although these models support coherent whole-body motion, such approaches can entangle policy prediction with controller-specific conventions (e.g., SONIC motion latents) and remain dependent on costly large-scale whole-body supervision.

These limitations raise a central question: *How can pre-trained world–action priors be generalized to humanoid loco-manipulation without learning whole-body behavior from scratch?*

Our key insight is that pre-trained WAMs already capture rich manipulation priors, but what is fundamentally missing for humanoid loco-manipulation is structured grounding to heterogeneous WBCs that regulates when and how whole-body behavior should coordinate. Based on this insight, we propose WholeBodyWAM, a WBC-grounded world action model that generalizes learned world–action priors to coordinated humanoid loco-manipulation through three complementary designs.

First, a *structured action representation* retains the manipulation pathway of the pre-trained WAMs and introduces a distinct whole-body control stream, with visual dynamics and both action streams jointly generated by a shared Diffusion Transformer (DiT). Second, the *Unified Whole-Body Controller Interface (UWBC)* assigns consistent physical meanings to shared commands while accommodating controller-specific extensions. Third, *Coordination-Aware Self-Attention (CASA)* uses task-directional arm manipulability to modulate an attention bias that strengthens manipulation-to-UWBC information flow when greater whole-body coordination is indicated.

Together, these designs preserve world–action priors of the pre-trained WAMs, ground heterogeneous WBC commands, and coordinate whole-body behavior for generalizable humanoid loco-manipulation. In summary, our contributions are threefold:

- We present WholeBodyWAM, a world action model that generalizes pre-trained world–action priors to humanoid loco-manipulation through coordinated WBC grounding.
- We develop UWBC as a shared semantic interface for heterogeneous WBCs and CASA for manipulation-informed whole-body coordination.
- Experiments on six simulation and eight real-world tasks demonstrate improved task performance, OOD generalization, and cross-WBC robustness.

II. Related Work

A. World Action Models

World Action Models couple visual dynamics with action generation to enable generalizable robot manipulation [1]–[4], [13], [14]. DreamZero [1] jointly models future video and actions. Cosmos Policy [4] encodes actions, future states, and values as latent frames within a pre-trained video diffusion model, while Fast-WAM [3] co-trains video and action prediction but removes future-video tokens at inference. Recent work has begun to extend this paradigm from tabletop manipulation toward humanoid loco-manipulation. DiT4DiT [2] applies video–action modeling to bimanual manipulation on a humanoid platform. MotionWAM [13] predicts a unified humanoid motion latent, while ω-0 [14] generates controller-compatible whole-body action latents. In contrast to expanding action interfaces or relearning humanoid world–action mappings, WholeBodyWAM adapts pre-trained WAMs by learning the WBC grounding and coordination needed for humanoid execution. This allows limited humanoid supervision to extend a reusable world–action prior to whole-body configurations.

B. Humanoid Loco-Manipulation

Learning-based humanoid loco-manipulation has advanced from whole-body motor control to task-level policy learning. OmniH2O [15], HOVER [16], and HugWBC [9] provide robust tracking and versatile whole-body control interfaces for externally specified commands. Building on these foundations, VLA-based methods such as WholeBodyVLA [10], Ψ_0 [12], HEX [17], and OpenHLM [18] learn autonomous whole-body policies from vision–language priors and human or cross-embodiment supervision. WAM-based MotionWAM [13] and ω-0 [14] further incorporate predictive world modeling into whole-body action generation. These policies generally use action representations tailored to specific downstream WBC interfaces. In contrast, WholeBodyWAM explicitly learns a shared semantic interface over heterogeneous WBC commands through UWBC. This enables reusable WBC grounding priors to be adapted across different downstream controllers.

III. Method

A. Problem Formulation

WholeBodyWAM formulates humanoid loco-manipulation as joint prediction of future world dynamics and structured

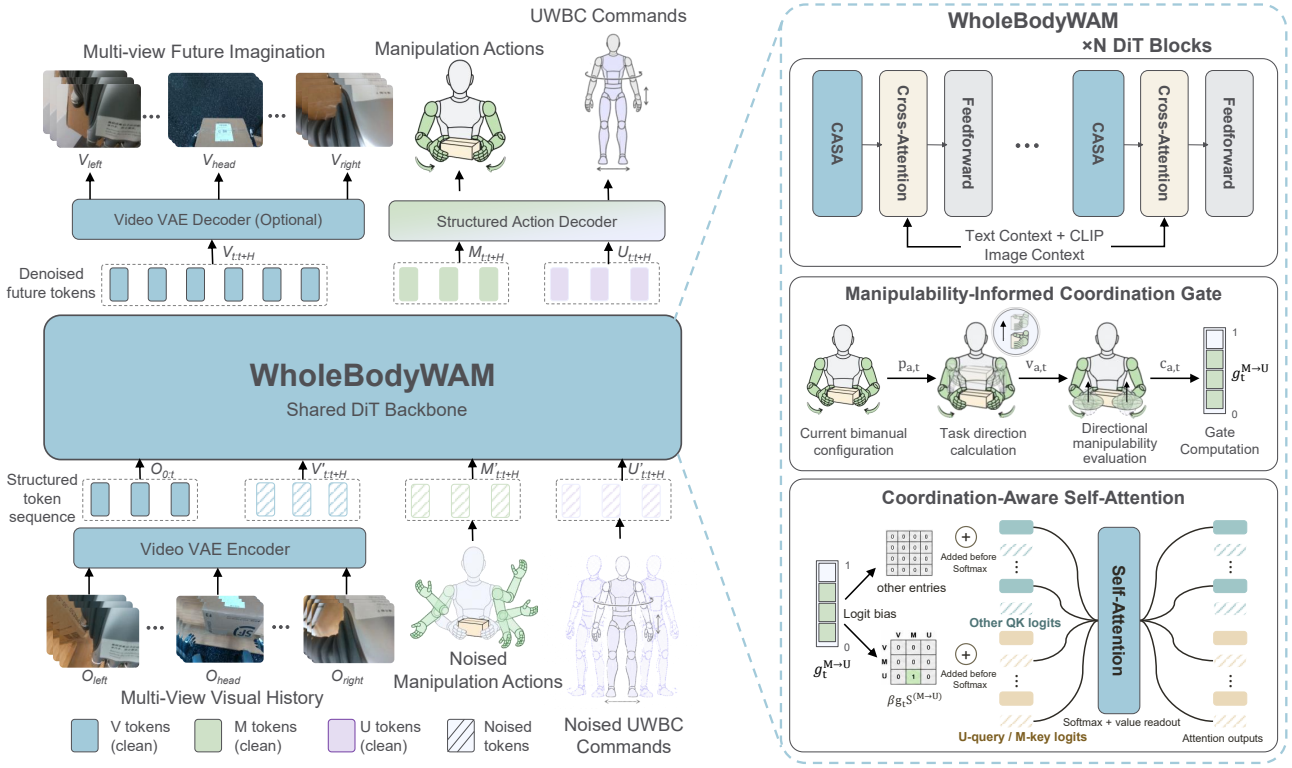

Fig. 2. **Model Architecture. Left:** A shared Diffusion Transformer jointly generates future visual dynamics, manipulation actions, and UWBC commands from a structured token sequence conditioned on visual history and task context. **Right:** CASA adaptively strengthens manipulation-to-UWBC attention based on reductions in task-directional arm manipulability, allowing manipulation intent to guide whole-body coordination.

whole-body actions under the task context:

$$p_\theta(\mathbf{v}_{t:t+H}, \mathbf{m}_{t:t+H}, \mathbf{u}_{t:t+H} \mid \mathbf{c}_t). \tag{1}$$

The context $\mathbf{c}_t$ comprises visual history $\mathbf{o}_{0:t}$, current robot state $\mathbf{s}_t$, and language instruction ℓ. Here H is the prediction horizon, and $\mathbf{v}_{t:t+H}$ denotes future visual dynamics. The manipulation stream $\mathbf{m}_{t:t+H}$ specifies arm joint targets and finger joint angles for task-level manipulation. The stream $\mathbf{u}_{t:t+H}$ carries whole-body commands expressed through the UWBC interface. Following joint world–action modeling [1], the conditional distribution in Eq. (1) can be factorized as:

$$\begin{aligned} &p_\theta(\mathbf{v}_{t:t+H}, \mathbf{m}_{t:t+H}, \mathbf{u}_{t:t+H} \mid \mathbf{c}_t) \\ &\quad = p_\theta^v(\mathbf{v}_{t:t+H} \mid \mathbf{c}_t)\, p_\theta^a(\mathbf{m}_{t:t+H}, \mathbf{u}_{t:t+H} \mid \mathbf{v}_{t:t+H}, \mathbf{c}_t). \end{aligned} \tag{2}$$

The world factor p_θ^v models future scene evolution conditioned on the task context, while the action factor p_θ^a jointly models manipulation actions and UWBC commands conditioned on both the task context and the corresponding future dynamics. This factorization describes the dependency structure of the joint world–action distribution rather than a sequential video-then-action inference procedure. The visual, manipulation, and UWBC streams are instantiated and optimized jointly within a shared generative backbone, as detailed in the following subsection.

B. Model Architecture

As illustrated in Fig. 2, WholeBodyWAM builds upon the pre-trained video Diffusion Transformer (DiT) backbone [1], which provides a shared generative model for visual dynamics and robot actions. We extend the existing world–action token sequence with a structured UWBC stream, allowing visual, manipulation, and whole-body motion to be jointly modeled. The pre-trained T5 text encoder [19] provides instruction features through cross-attention. The pre-trained Wan video VAE [20] compresses observations into latent video tokens, while a lightweight state encoder maps the current robot state to proprioceptive conditioning tokens. Future visual tokens $\mathbf{v}$, manipulation tokens $\mathbf{m}$, and UWBC tokens form three typed streams that share a prediction block at each future step and are jointly processed by the DiT. The manipulation stream retains the visual–manipulation pathway inherited from the pre-trained WAM [1], while distinct UWBC tokens make controller-facing whole-body behavior explicitly addressable. Following the joint flow-matching formulation [1], WholeBodyWAM minimizes a modality-weighted objective [21]:

$$\mathcal{L}(\theta) = \mathbb{E}\left[\sum_k \lambda_k w_k(\tau) \left\| f_\theta^k(\mathbf{x}_\tau, \tau, \mathbf{c}_t) - \dot{\mathbf{x}}^{k,*} \right\|_k^2\right]. \tag{3}$$

Here $\mathbf{x}_\tau$ denotes the noisy joint block at flow time $\tau \in [0,1]$, and $\dot{\mathbf{x}}^{k,*}$ is the target flow velocity for stream $k \in \{v, m, u\}$. The coefficient λ_k balances the stream losses, while $w_k(\tau)$ is a loss weight that varies with flow time.

C. Unified WBC Interface

Existing whole-body controllers expose heterogeneous command interfaces spanning task-space objectives, discrete

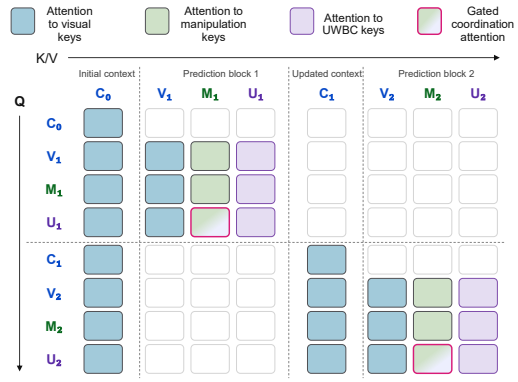

Fig. 3. **Blockwise-Causal Attention with Gated Coordination.** Native self-attention visibility is preserved across visual, manipulation, and UWBC streams, while a gated coordination bias selectively enhances manipulation-to-UWBC attention within each prediction block.

control modes, and joint-space references with different physical semantics [7], [8]. Naively concatenating these quantities into a flat action vector obscures their control semantics and tightly couples the learned policy to a specific controller interface. We therefore introduce the Unified Whole-Body Controller Interface (UWBC) to organize heterogeneous whole-body commands according to shared physical semantics while supporting controller-specific extensions. As shown in Table I, UWBC defines shared continuous command fields with consistent physical meanings across controllers. Task-space whole-body commands, together with lower-body and waist joint positions and velocities, form a 46-D block of shared continuous commands. Individual WBCs may require additional controller-specific commands, which are accommodated by ten residual slots at indices $[46, 56)$. The command vector expressed through UWBC is defined as

$$\begin{aligned} \mathbf{u}_t &= [\mathbf{u}_t^{\mathrm{sh}}, \mathbf{u}_t^{r}] \in \mathbb{R}^{56}, \\ \mathbf{u}_t^{\mathrm{sh}} &= [\mathbf{u}_t^{\mathrm{task}}, \mathbf{q}_t^{\mathrm{lower}}, \dot{\mathbf{q}}_t^{\mathrm{lower}}] \in \mathbb{R}^{46}. \end{aligned} \tag{4}$$

Here $\mathbf{u}_t^{\mathrm{sh}}$ collects the shared slots, $\mathbf{u}_t^{\mathrm{task}} \in \mathbb{R}^{16}$ contains task-space commands, and $\mathbf{q}_t^{\mathrm{lower}}, \dot{\mathbf{q}}_t^{\mathrm{lower}} \in \mathbb{R}^{15}$ specify desired lower-body and waist joint positions and velocities. The residual slots $\mathbf{u}_t^{r} \in \mathbb{R}^{10}$ encode controller-specific commands. For a downstream controller, UWBC fields are activated according to their registered physical semantics and controller capability profile, while unsupported or undefined fields are masked.

D. Coordination-Aware Self-Attention

Coordination in humanoid loco-manipulation is inherently directional and state dependent, requiring compensatory body motions when arm manipulability along the task direction becomes insufficient. We therefore introduce a **coordination-aware self-attention mechanism** that selectively strengthens manipulation-to-UWBC information flow.

Gated attention logit bias. The native self-attention logits are modified as follows:

$$\mathbf{L}'_t = \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{M}^{\mathrm{nat}} + \beta g_t \mathbf{S}^{M\to U}. \tag{5}$$

TABLE I

UWBC INTERFACE SLOT SPECIFICATION: SHARED PHYSICAL-SEMANTIC FIELDS AND CONTROLLER-SPECIFIC RESIDUAL SLOTS.

Index Range	Element Index	Mapped Physical Semantics
		Shared slots $[0, 46)$ (46-D)
[0,3)	0–2	Desired base linear velocity (v_x, v_y, v_z)
[3,6)	3–5	Desired base angular velocity $(\omega_x, \omega_y, \omega_z)$
[6,8)	6–7	Desired heading $(\sin\psi, \cos\psi)$
[8,10)	8–9	Desired pelvis planar position (x, y)
[10,11)	10	Desired pelvis height
[11,13)	11–12	Desired pelvis tilt (roll/pitch)
[13,16)	13–15	Desired torso-to-pelvis orientation (roll/pitch/yaw)
[16,31)	16–30	Desired lower-body/waist joint positions
[31,46)	31–45	Desired lower-body/waist joint velocities
		Residual slots $[46, 56)$ (10-D)
[46,56)	46–55	Registered controller-specific commands, including discrete control modes

Here $\mathbf{Q}$ and $\mathbf{K}$ denote the query and key matrices, d denotes the head dimension, and $\mathbf{M}^{\mathrm{nat}}$ represents the native self-attention mask. The coordination gate $g_t \in [0, 1]$ scales a fixed bias of strength $\beta > 0$. As illustrated in Fig. 3, the selector $S_{ij}^{M\to U} = 1$ only when i is a UWBC query and j is a manipulation key in the same prediction block.

Coordination gate computation. The gate g_t evaluates reduced task-directional manipulability in the current arm configurations to determine whether whole-body coordination is needed. First, for each arm $a \in \{L, R\}$, the task direction is calculated from consecutive observed end-effector positions during motion:

$$\mathbf{v}_{a,t} = \frac{\mathbf{p}_{a,t} - \mathbf{p}_{a,t-1}}{\Delta t}, \qquad \hat{\mathbf{d}}_{a,t} = \frac{\mathbf{v}_{a,t}}{\|\mathbf{v}_{a,t}\|_2}. \tag{6}$$

Here $\mathbf{p}_{a,t}$ denotes the observed end-effector position, Δt denotes the observation interval, and $\mathbf{v}_{a,t}$ denotes the observed end-effector velocity. The unit vector $\hat{\mathbf{d}}_{a,t}$ represents the estimated task direction. The manipulability of each arm configuration along this task direction is then calculated as

$$c_{a,t} = \left[\hat{\mathbf{d}}_{a,t}^\top \left(C_a^{v,\lambda}(\mathbf{q}_{a,t}^{\mathrm{obs}})\right)^{-1} \hat{\mathbf{d}}_{a,t}\right]^{-1/2}. \tag{7}$$

Here $C_a^{v,\lambda} \in \mathbb{R}^{3\times 3}$ denotes the damped manipulability matrix incorporating joint speed limits, and $\mathbf{q}_{a,t}^{\mathrm{obs}}$ denotes the observed arm joint configuration. We construct a voxelized manipulability map offline following [22]. At runtime, the reference $c_{a,t}^{\mathrm{ref}}$ is the maximum sampled radial capability along the estimated task direction among configurations in the current end-effector pose voxel. The gate for each arm is then calculated from the relative reduction in manipulability:

$$g_{a,t} = 1 - \mathrm{clip}\left(\frac{c_{a,t}}{c_{a,t}^{\mathrm{ref}} + \epsilon_m}, 0, 1\right), \tag{8}$$

where $\epsilon_m > 0$ is a small constant that prevents division by zero. Finally, the global gate for both arms takes the larger relative reduction:

$$g_t^{M\to U} = \max(g_{L,t}, g_{R,t}). \tag{9}$$

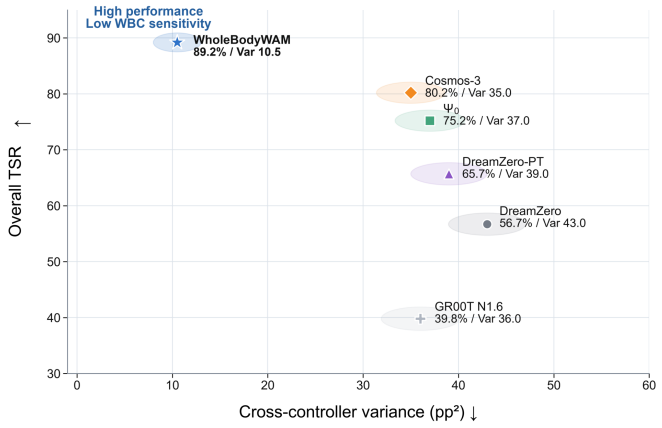

Fig. 4. **Cross-WBC evaluation.** Mean TSR and population variance across SONIC, AMO, and GEAR WBC, with separate fine-tuning for each controller. WholeBodyWAM achieves the highest mean TSR (89.2%) and lowest variance (10.5 pp^2).

This ensures that reduced task-directional manipulability in either active arm can trigger manipulation-to-UWBC coordination.

IV. Experiments

We conduct comprehensive experiments to investigate the following questions: (1) Does WholeBodyWAM improve task success over WAM and VLA baselines? (2) Does UWBC improve stability and portability across heterogeneous WBCs? (3) Can manipulation priors generalize to new whole-body coordination in OOD settings? (4) How does each mechanism contribute to the performance of WholeBodyWAM?

A. Experimental Setup

Training and implementation. WholeBodyWAM is initialized from the pre-trained 14B WAM [1], retaining its world–action backbone. We collect 15K whole-body loco-manipulation demonstrations in SIMPLE [23] using a PICO 4 Ultra headset and wrist trackers. Each demonstration is converted into a SMPL motion sequence [24] and retargeted to the G1 embodiment using GMR [25], yielding desired robot motion references. We then calculate supervision for each UWBC shared slot from the robot motion references according to its predefined physical semantics, coordinate convention, and unit. After temporal alignment, the UWBC targets are incorporated into the post-training dataset. We post-train WholeBodyWAM on this dataset using LoRA [26], with rank $r = 4$ and scaling parameter $\alpha = 4$. The model is optimized for 50,000 steps with a learning rate of 10^{-5} and a batch size of 1 per GPU. At inference, we use 16 denoising steps with a second-order UniPC flow-matching sampler and a fixed noise-schedule shift of 5.

Platforms. We conduct all real-world experiments on a Unitree G1 humanoid equipped with two BrainCo Revo 2 dexterous hands. Visual observations are captured by one head-mounted and two wrist-mounted Intel RealSense D435i RGB cameras. For simulation, we evaluate humanoid loco-manipulation on the SIMPLE benchmark [23].

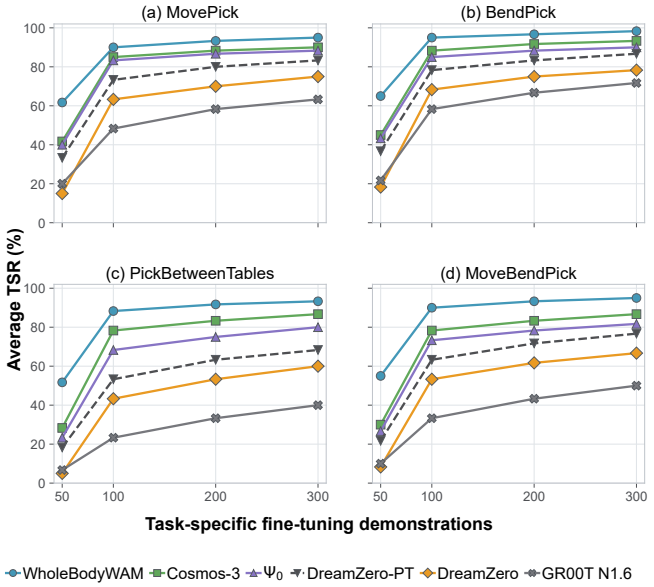

Fig. 5. **Downstream fine-tuning efficiency.** We compare downstream fine-tuning efficiency across methods. WholeBodyWAM achieves its largest advantage with few task-specific demonstrations, consistent with effective reuse of pre-trained world–action priors through coordinated WBC grounding.

Baselines. (1) *DreamZero* [1] jointly models video and actions for generalizable bimanual manipulation. To adapt its bimanual checkpoint to humanoid tasks, we resize the pre-trained action head to match the G1 embodiment before downstream task fine-tuning. (2) *DreamZero-PT* [1] follows the same model and humanoid adaptation setup as DreamZero, with additional post-training on the same dataset as WholeBodyWAM before downstream task fine-tuning. (3) *Cosmos-3* [27] unifies world modeling and action generation, with demonstrated generalization in tabletop manipulation. For humanoid adaptation, we use the Cosmos3-Nano baseline from SIMPLE with a resized humanoid action head and fine-tune it on downstream tasks. (4) Ψ_0 [12] is a native whole-body VLA that learns humanoid loco-manipulation from egocentric human videos and humanoid demonstrations. We directly fine-tune its pre-trained model on downstream task demonstrations. (5) *GR00T N1.6* [28] is a humanoid VLA pre-trained on diverse robot data. We adapt it through direct fine-tuning on downstream task demonstrations.

Metrics. We adopt three quantitative metrics. (1) *Task Success Rate (TSR)* is the percentage of trials completing all task objectives within the evaluation horizon. (2) *Task Progress (TP)* [14] measures normalized progress along the ordered task stages before the first failure or irreversible deviation. (3) *Cross-WBC Variance* is the population variance of mean TSR across different downstream controllers, measured in pp^2, with lower values indicating greater robustness. All metrics use 20 independent trials for each combination of method and task. TSR and TP are averaged over trials, while cross-WBC variance is computed from the resulting controller-wise TSRs.

B. Simulation Experiments

Overall performance. Table II reports performance across six tasks and three SIMPLE perturbation levels. WholeBody-

TABLE II

SIMULATION TASK SUCCESS RATE (%) WITH SONIC. EACH TASK ENTRY REPORTS L0/L1/L2 RESULTS UNDER THE SIMPLE PROTOCOL.

Method	MovePick	BendPick	Handover	PickBetween Tables	Tabletop Grasp	MoveBend Pick	Overall
WAM and VLA baselines							
DreamZero	70/65/55	75/70/60	85/80/70	50/45/35	90/85/75	60/55/45	65.0
DreamZero-PT	80/75/65	85/80/70	90/85/80	60/55/45	95/90/80	70/65/55	73.6
Cosmos-3	90/85/80	95/90/80	**100/95/90**	85/80/70	**100/95**/85	85/80/70	86.4
GR00T N1.6	55/50/40	65/60/50	60/55/45	30/25/15	75/70/60	40/35/25	47.5
Ψ_0	90/85/75	90/85/80	95/90/80	75/70/60	**100/95/90**	80/75/65	82.2
WholeBodyWAM	**95/90/85**	**100/95/90**	**100/95/90**	**95/90/80**	**100/95**/85	**95/90/85**	**91.9**
Ablations							
w/o CASA	90/85/80	95/90/85	**100/95**/85	85/80/75	95/90/85	90/85/75	86.9
w/o UWBC	90/85/75	95/90/80	95/90/85	85/80/70	95/90/85	85/80/70	84.7
w/o SAF	85/80/70	90/85/75	95/90/80	80/75/65	95/90/80	80/75/65	80.8

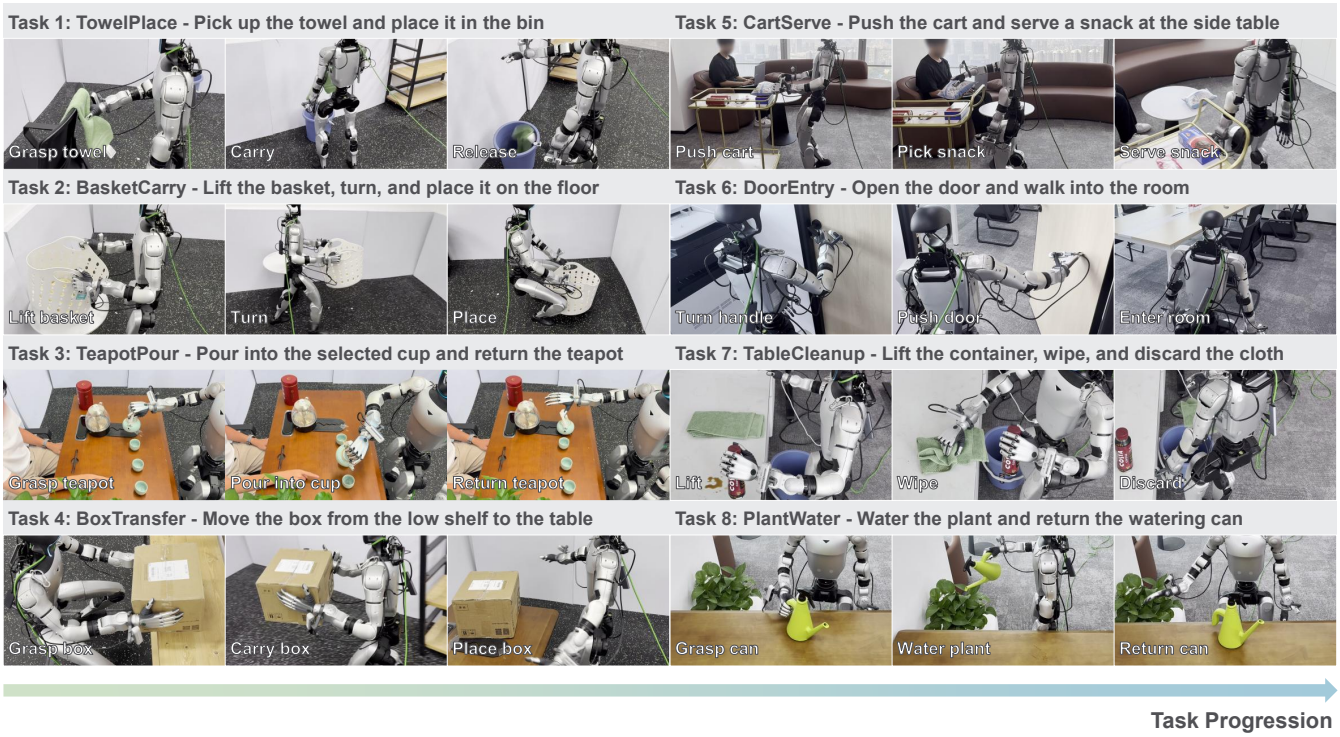

Fig. 6. **Real-world qualitative results.** We show WholeBodyWAM performing eight real-world tasks spanning manipulation-dominant and coordination-intensive behaviors. Task instructions appear above each sequence, with frames ordered from left to right. Lower-left labels identify task stages associated with Task Progress (TP) in Fig. 7.

WAM achieves the highest overall TSR of 91.9%, exceeding the WAM baselines DreamZero, DreamZero-PT, and Cosmos-3 by 26.9, 18.3, and 5.6 percentage points, respectively. The advantage over DreamZero-PT, which uses the same post-training data, indicates that additional humanoid supervision alone does not explain the gains. Compared with Cosmos-3, improvements are larger on coordination-intensive PickBetweenTables and MoveBendPick (10.0 and 11.7 points), while the two methods achieve comparable task-averaged TSRs on Handover and TabletopGrasp. WholeBodyWAM also outperforms the VLA baselines Ψ_0 (82.2%) and GR00T N1.6 (47.5%), exceeding the highest VLA baseline TSR by 9.7 percentage points.

Cross-WBC robustness. We additionally evaluate each method across tasks using AMO [8], SONIC [7], and GEAR WBC [29], with separate task-specific fine-tuning for each controller. Following Table I, SONIC uses joint-space slots $[16, 46)$ for lower-body/waist references. AMO maps planar velocity, heading, absolute height, and torso orientation to $[0, 2)$, $[6, 8)$, 10, and $[13, 16)$, respectively, with its turning flag in residual slots. GEAR WBC uses the same velocity, height, and orientation slots, with its derived yaw rate in slot 5. For each target WBC, fine-tuning activates only fields corresponding to controller commands available in the demonstrations and masks all other registered fields. Figure 4 shows that WholeBodyWAM achieves the highest TSR averaged across the three controllers (89.2%) and the smallest cross-controller variance (10.5 pp^2), compared with

TABLE III

REAL-WORLD TASK SUCCESS RATE (%) UNDER IN-DISTRIBUTION (ID) AND OUT-OF-DISTRIBUTION (OOD) CONDITIONS.

Method	Towel Place	Basket Carry	Teapot Pour	Box Transfer	Cart Serve	Door Entry	Table Cleanup	Plant Water	Overall
In-Distribution Evaluation									
WholeBodyWAM	**85**	**75**	**85**	**75**	**80**	**80**	**90**	**80**	**81.3**
DreamZero	65	40	75	35	50	45	85	65	57.5
Out-of-Distribution Evaluation									
WholeBodyWAM	**75**	**55**	**80**	**60**	**65**	**70**	**80**	**65**	**68.8**
DreamZero	50	15	65	20	30	25	70	45	40.0

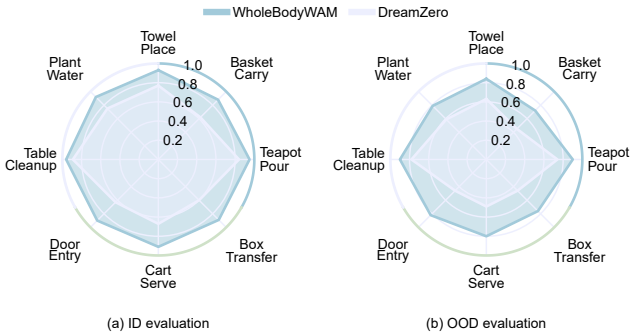

Fig. 7. **Task progress under ID (left) and OOD (right).** WholeBodyWAM maintains strong progress across eight tasks spanning manipulation-dominant and coordination-intensive behaviors.

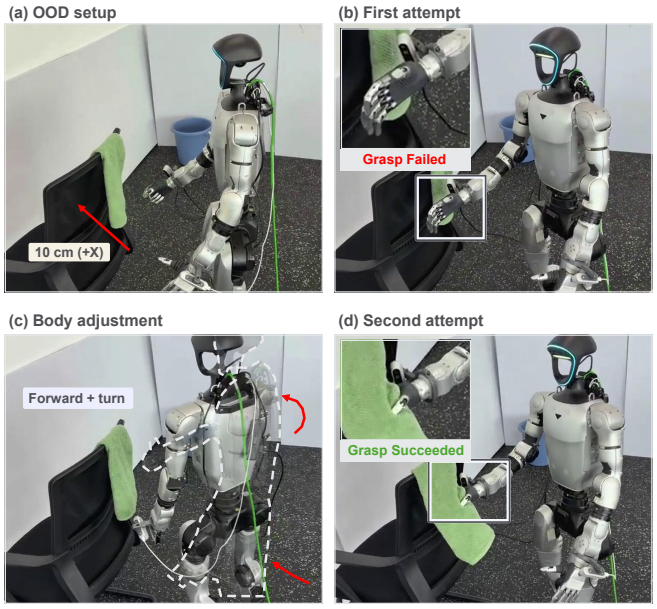

Fig. 8. **Whole-body coordination generalization.** Under a 10 cm displacement of the towel support, WholeBodyWAM recovers from a failed grasp through forward repositioning and body rotation.

80.2% and 35.0 pp[2] for Cosmos-3. For baselines initially trained with a single WBC such as SONIC, adaptation requires learning new command conventions without a shared prior across interfaces. This provides a plausible explanation for their performance degradation and greater sensitivity to WBC choice. By separating internal whole-body intent from controller-specific command conventions, UWBC supports more compatible and stable adaptation across the evaluated WBC interfaces.

Downstream fine-tuning efficiency. Figure 5 compares TSR on four simulation tasks using 50/100/200/300 task-specific fine-tuning demonstrations per task; 100 demonstrations is the main simulation budget. WholeBodyWAM achieves its largest advantage with few task-specific demonstrations. The results support effective reuse of pre-trained world–action priors for humanoid loco-manipulation through structured WBC grounding and coordination.

C. Real-World Experiments

We evaluate WholeBodyWAM and DreamZero on eight real-world tasks (Fig. 6) under ID and OOD conditions. ID follows the training distribution, whereas OOD changes object configurations and language instructions.

ID performance. As shown in Table III, WholeBodyWAM achieves a mean TSR of 81.3%, compared with 57.5% for DreamZero. The gap is small on manipulation-dominant tasks such as TableCleanup (90% versus 85%). This suggests that pre-trained manipulation priors already support effective execution when targets lie within the arm workspace and require little base or torso adjustment. As tasks involve more body subsystems and coordination stages, baseline performance declines, suggesting that treating manipulation and WBC commands as homogeneous action dimensions may inadequately capture complex whole-body coordination behavior. In contrast, WholeBodyWAM maintains strong performance, with gains of 35 points on BasketCarry and DoorEntry and 40 on BoxTransfer. These gains are consistent with the benefits of structured WBC grounding and manipulation-informed coordination on tasks requiring substantial body reconfiguration.

OOD generalization. WholeBodyWAM maintains a mean TSR of 68.8%, compared with 40.0% for DreamZero, with a smaller ID-to-OOD drop (12.5 versus 17.5 points). The advantage reaches 45 points on DoorEntry, which requires substantial whole-body adjustment. Figure 7 shows mean ID/OOD task progress of 0.92/0.82 for WholeBodyWAM versus 0.73/0.59 for DreamZero, with strong performance on manipulation-dominant tasks and larger gains on coordination-intensive tasks. These results suggest that WholeBodyWAM preserves task-level manipulation performance while adapting its execution to new whole-body coordination requirements under distribution shift. Moreover, we observe notable whole-body coordination generalization in OOD settings, as shown

in Fig. 8. In TowelPlace, displacing the towel support 10 cm farther from the robot puts the towel beyond the initial arm workspace. After one unsuccessful local reaching attempt, WholeBodyWAM takes a short forward step and adjusts its torso to extend its reach, successfully grasping the towel on the next attempt. This recovery sequence is absent from our demonstrations. We interpret it as emergent behavior arising from pre-trained WAM manipulation priors coupled with enhanced whole-body coordination.

D. Ablation Studies

We ablate CASA, UWBC, and structured action factorization (SAF) across six simulation tasks (Table II). Removing CASA retains the manipulation and UWBC streams but restores native self-attention, reducing mean TSR from 91.9% to 86.9% (5.0 points). The reduction is larger on PickBetweenTables (8.3 points) and MoveBendPick (6.7 points) than on Handover (1.7 points), consistent with the benefit of state-dependent directional coordination on tasks requiring greater whole-body adjustment. Replacing UWBC with native SONIC commands while retaining the WBC pathway and supervision reduces TSR to 84.7% (7.2 points). This suggests that explicit physical semantics improve controller grounding even when a distinct WBC pathway is available. Finally, removing SAF replaces the manipulation and UWBC streams with a capacity-matched monolithic stream under the same data and optimization budget, reducing TSR to 80.8% (11.1 points). The largest reduction supports preserving the pre-trained manipulation pathway and keeping WBC commands separately addressable within joint generation.

V. Conclusion

We present WholeBodyWAM, a WBC-grounded world action model for coordinated humanoid loco-manipulation. WholeBodyWAM preserves pre-trained world–action priors and grounds heterogeneous whole-body commands through UWBC. Coordination-aware self-attention further enhances state-dependent manipulation-to-UWBC interaction. Experiments in simulation and on real robots demonstrate improved task performance, OOD generalization, and cross-WBC robustness while maintaining strong performance on manipulation-dominant tasks. Data-scaling and ablation results further support the value of structured prior reuse. More broadly, our results suggest that scalable humanoid whole-body intelligence can be built by generalizing reusable world–action priors through structured WBC grounding and coordination, rather than relearning whole-body behavior from scratch.

References

[1] S. Ye *et al.*, "World Action Models are Zero-shot Policies," *arXiv preprint arXiv:2602.15922*, 2026.

[2] T. Ma *et al.*, "DiT4DiT: Jointly Modeling Video Dynamics and Actions for Generalizable Robot Control," *arXiv preprint arXiv:2603.10448*, 2026.

[3] T. Yuan, Z. Dong, Y. Liu, and H. Zhao, "Fast-WAM: Do World Action Models Need Test-time Future Imagination?" *arXiv preprint arXiv:2603.16666*, 2026.

[4] M. J. Kim *et al.*, "Cosmos Policy: Fine-Tuning Video Models for Visuomotor Control and Planning," in *International Conference on Learning Representations*, 2026.

[5] L. Li *et al.*, "Causal World Modeling for Robot Control," in *Proc. Robotics: Science and Systems*, 2026.

[6] H. Bi *et al.*, "Motus: A Unified Latent Action World Model," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 35 101–35 113.

[7] Z. Luo *et al.*, "SONIC: Supersizing Motion Tracking for Natural Humanoid Whole-Body Control," *Science Robotics*, vol. 11, no. 117, p. eaed4592, 2026.

[8] J. Li, X. Cheng, T. Huang, S. Yang, R.-Z. Qiu, and X. Wang, "AMO: Adaptive Motion Optimization for Hyper-Dexterous Humanoid Whole-Body Control," in *Proc. Robotics: Science and Systems*, 2025.

[9] Y. Xue, W. Dong, M. Liu, W. Zhang, and J. Pang, "A Unified and General Humanoid Whole-Body Controller for Fine-Grained Locomotion," in *Proc. Robotics: Science and Systems*, 2025.

[10] H. Jiang *et al.*, "WholeBodyVLA: Towards Unified Latent VLA for Whole-Body Loco-Manipulation Control," in *International Conference on Learning Representations*, 2026.

[11] M. Shi *et al.*, "Unlocking In-the-Wild Loco-Manipulation with Robot-Free Egocentric Demonstration," in *Proc. Robotics: Science and Systems*, 2026.

[12] S. Wei *et al.*, "Ψ_0: An Open Foundation Model Towards Universal Humanoid Loco-Manipulation," in *Proc. Robotics: Science and Systems*, 2026.

[13] J. Zheng, T. Ma, Y. Fan, Z. Wang, S. Yang, and J. Liang, "MotionWAM: Towards Foundation World Action Models for Real-Time Humanoid Loco-Manipulation," *arXiv preprint arXiv:2606.09215*, 2026.

[14] Z. Li *et al.*, "ω-0: A Latent Predictive World Action Model for Concurrent Humanoid Loco-Manipulation," *arXiv preprint arXiv:2608.06375*, 2026.

[15] T. He *et al.*, "OmniH2O: Universal and Dexterous Human-to-Humanoid Whole-Body Teleoperation and Learning," in *Proc. Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 270, 2025, pp. 1516–1540.

[16] T. He *et al.*, "HOVER: Versatile Neural Whole-Body Controller for Humanoid Robots," in *Proc. IEEE International Conference on Robotics and Automation*, 2025, pp. 9989–9996.

[17] S. Bai *et al.*, "HEX: Humanoid-Aligned Experts for Cross-Embodiment Whole-Body Manipulation," *arXiv preprint arXiv:2604.07993*, 2026.

[18] Y. Hu *et al.*, "OpenHLM: An Empirical Recipe for Whole-Body Humanoid Loco-Manipulation," *arXiv preprint arXiv:2606.22174*, 2026.

[19] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[20] Wan Team *et al.*, "Wan: Open and Advanced Large-Scale Video Generative Models," *arXiv preprint arXiv:2503.20314*, 2025.

[21] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow Matching for Generative Modeling," in *International Conference on Learning Representations*, 2023.

[22] N. Vahrenkamp, T. Asfour, G. Metta, G. Sandini, and R. Dillmann, "Manipulability Analysis," in *Proc. IEEE-RAS International Conference on Humanoid Robots*, 2012, pp. 568–573.

[23] S. Wei *et al.*, "SIMPLE: Simulation-Based Policy Learning and Evaluation for Humanoid Loco-manipulation," *arXiv preprint arXiv:2606.08278*, 2026.

[24] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Transactions on Graphics*, vol. 34, no. 6, pp. 248:1–248:16, 2015.

[25] J. P. Araújo, Y. Ze, P. Xu, J. Wu, and C. K. Liu, "Retargeting matters: General motion retargeting for humanoid motion tracking," in *Proc. IEEE International Conference on Robotics and Automation*, 2026.

[26] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," in *International Conference on Learning Representations*, 2022.

[27] NVIDIA *et al.*, "Cosmos 3: Omnimodal World Models for Physical AI," *arXiv preprint arXiv:2606.02800*, 2026.

[28] NVIDIA, "GR00T N1.6: An improved open foundation model for generalist humanoid robots," Research release, 2025. [Online]. Available: https://research.nvidia.com/labs/gear/gr00t-n1_6/

[29] NVIDIA, "GR00T-WholeBodyControl: Decoupled WBC," Software documentation, 2026. [Online]. Available: https://nvlabs.github.io/GR00T-WholeBodyControl/references/decoupled_wbc.html