BEING-M0.7: A LATENT WORLD-ACTION MODEL FOR HUMANOID ROBOTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Humanoid loco-manipulation requires coordinated locomotion and manipulation informed by future scene evolution and whole-body motion, yet learning these capabilities is constrained by scarce robot demonstrations. Human video and motion datasets offer scalable supervision, but many contain only video or motion rather than paired video-motion data. Moreover, human motion does not directly specify executable robot actions. We present **Being-M0.7**, a latent world-action model that transfers visual–motion priors learned from mixed-modality human data to humanoid control through pre-training, robot mid-training, and action post-training. We curate a corpus from more than 10,000 hours of raw human-centric data, integrating video-only, motion-only, and paired video–motion streams to learn complementary visual dynamics and whole-body kinematic structure. Joint prediction of future latent visual states and motion encourages visual representations to encode future kinematics. Robot mid-training adapts this coarse-grained prior to robot viewpoints and body dynamics. During action post-training, an action expert combines visual predictive representations from the frozen, adapted prior with current images and proprioception through gated cross-attention, grounding predictive context in executable whole-body commands. Being-M0.7 achieves the highest aggregate success rate among the compared baselines on SIMPLE and matches the strongest baseline on real-world Unitree G1 loco-manipulation tasks.

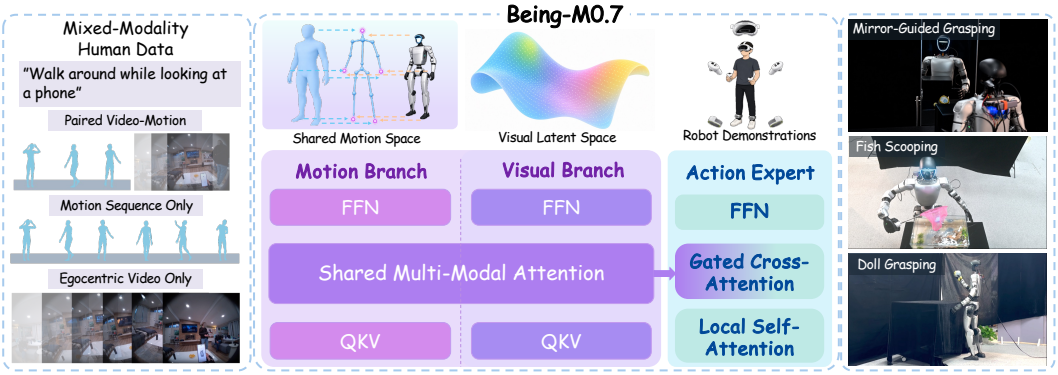

Figure 1: **Being-M0.7 at a glance.** Being-M0.7 is a latent World-Action Model pretrained on egocentric videos and human motions with a Mixture of Transformers (MoT) structure, then grounded into humanoid control through action expert post-training on robot trajectories for diverse loco-manipulation tasks. The rightmost panels show three real-world tasks executed by the humanoid.

1 INTRODUCTION

Humanoid loco-manipulation aims to enable a humanoid robot to move through human-centric environments while performing object-centric tasks with coordinated whole-body motion (He et al., 2025; Ben et al., 2025; Fu et al., 2025; Wei et al., 2026a; Jiang et al., 2026; Shi et al., 2026). Unlike fixed-base manipulation, the robot must jointly decide where to move, how to orient its body, how to

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.

coordinate its hands and feet, and how to maintain a stable posture while interacting with the scene. This makes the problem intrinsically temporal and whole-body: a useful policy should reason about current observations as well as future scene evolution, body motion, and task progress.

Recent robot foundation models and world-action models have made encouraging progress (Brohan et al., 2022; Zitkovich et al., 2023; O'Neill et al., 2024; Octo Model Team et al., 2024; Black et al., 2024; Kim et al., 2025; Du et al., 2023; Agarwal et al., 2025; Ye et al., 2026; Luo et al., 2026a;b), but their extension to humanoid loco-manipulation is constrained by the scarcity of robot demonstrations. Collecting synchronized egocentric video, proprioception, and executable whole-body commands requires demanding teleoperation and is limited by safety, hardware availability, and environment diversity (He et al., 2025; Ben et al., 2025; Fu et al., 2025). Human egocentric videos and motion capture data offer a scalable source of supervision for visual dynamics and whole-body behavior (Grauman et al., 2022; Ma et al., 2024; Guo et al., 2022; guo et al., 2025). However, these sources differ in modality coverage, with many providing only video or motion. Moreover, human motion does not directly specify executable robot actions. Combining video-only, motion-only, and paired video–motion data can expand the scale of pre-training beyond scarce robot demonstrations. The challenge is to transfer these human visual–motion priors to executable humanoid actions.

To this end, we present **Being-M0.7**, a latent world-action model for humanoid loco-manipulation, together with a training framework that transfers visual–motion priors learned from mixed-modality human data to executable robot actions. We curate a pre-training corpus from more than 10,000 hours of raw human-centric data, organizing it into video-only, motion-only, and paired video–motion streams. Video-only data provides visual dynamics across diverse interaction contexts, motion-only data supplies whole-body kinematic structure, and paired data links visual evolution to body motion. The model learns from the available modalities through latent visual prediction and motion prediction. Joint visual–motion prediction encourages the visual representations to focus on underlying kinematic changes rather than the visual details needed for pixel-level reconstruction, providing the downstream policy with motion-informed predictive context for coordinated whole-body action generation.

This coarse-grained visual–motion prior captures future scene evolution and body motion, but does not directly specify executable robot commands for precise interaction. Moreover, a prior model learned from human data must adapt to robot viewpoints and body dynamics. We therefore use robot mid-training to adapt the visual–motion prior model to robot trajectories, followed by action post-training to learn executable whole-body commands. During post-training, a separate action expert combines visual predictive representations from the frozen, adapted prior model with current images and proprioception at each sampled action query. Through gated cross-attention, the expert learns how much predictive information to incorporate from the prior model's visual hidden states. These visual representations also capture information about future kinematics learned through joint visual–motion modeling. Together, these three stages form a framework for transferring scalable human visual–motion learning to closed-loop humanoid loco-manipulation tasks.

Our contributions are summarized as follows:

- We propose Being-M0.7, a latent world-action model for humanoid loco-manipulation, with a unified pre-training, mid-training, and post-training framework. Joint visual–motion pre-training from large-scale human video and motion data encourages visual predictive representations to encode future kinematics, while robot mid-training and gated action post-training adapt the coarse-grained visual–motion prior and ground it in executable whole-body control.
- We curate a mixed-modality pre-training corpus from more than 10,000 hours of raw human-centric data, integrating video-only, motion-only, and paired video–motion streams. These sources complement costly robot demonstrations with visual and whole-body motion supervision.
- We evaluate Being-M0.7 on the SIMPLE benchmark and real-world Unitree G1 tasks, achieving the highest aggregate simulation success rate among the compared baselines and matching the strongest baseline on real-world tasks. Ablations support the contributions of mixed-modality pre-training, robot mid-training, and predictive context to downstream control.

2 Related Work

Humanoid Loco-manipulation. Learning-based humanoid control combines whole-body motion priors with task-conditioned interaction. Such interaction requires coordinating locomotion, bal-

ance, and manipulation as object geometry and contact conditions change. A useful motion prior must therefore be coupled with perceptual feedback to support task execution in the current scene. Reference-based methods use retargeted demonstrations or generated trajectories to guide coordinated behavior (Fu et al., 2025; Zhao et al., 2025; Weng et al., 2025), while reference-free approaches learn from task objectives and perception (Lin et al., 2026; Xu et al., 2025). Reference guidance supplies a concrete target for whole-body coordination, whereas task-driven learning must discover suitable behavior through the interaction objective. This distinction highlights two complementary needs: retaining useful movement structure and allowing the policy to adjust that structure to the object and task at hand. Structured interaction representations and hierarchical control further support geometric generalization (Lin et al., 2026; Yin et al., 2025). These approaches motivate learning reusable motion knowledge beyond individual skills. Being-M0.7 learns this knowledge through mixed-modality video–motion pre-training, followed by robot adaptation and action grounding.

Egocentric Motion and Joint Prediction. Text-conditioned motion models learn semantic whole-body behavior from large motion corpora (Tevet et al., 2022; Jiang et al., 2023). Egocentric methods additionally connect motion to the agent's viewpoint: EgoEgo and EgoAllo recover body motion from first-person observations (Li et al., 2023; Yi et al., 2025), while UniEgoMotion combines reconstruction, forecasting, and generation (Patel et al., 2025). EgoAgent jointly predicts future states and actions (Chen et al., 2025), and EgoTwin couples egocentric video generation with head-centered human motion (Xiu et al., 2025). Together, these directions motivate learning how visual observations evolve with body motion, linking perception with coordinated movement. For downstream control, the value of such prediction lies in providing temporal context for action selection as the interaction unfolds. Transferring this knowledge to a humanoid additionally requires an interface between human motion structure and robot commands, with feedback to accommodate differences in embodiment and execution. Our focus is transferring predictive knowledge to humanoid execution: we jointly learn visual-latent and motion dynamics, use a shared human–humanoid motion representation, and train an action expert on the adapted prior model's visual context.

Humanoid Foundation and World-Action Models. GR00T (Bjorck et al., 2025), WholeBodyVLA (Jiang et al., 2026), and Ψ_0 (Wei et al., 2026a) explore generalist humanoid policies using heterogeneous supervision, including human videos and robot trajectories. World-action modeling provides a related route through predictive representations that connect anticipation with control. MotionWAM (Zheng et al., 2026) conditions a humanoid motion model on intermediate features from an egocentric video generator. Being-M0.7 also uses intermediate predictive features, with a recipe centered on latent visual–motion learning from paired and unpaired data, robot mid-training, and gated action post-training. In this formulation, video-only and motion-only samples contribute visual dynamics and kinematic supervision, while paired samples connect the two modalities. Motion thus supplies structure during prior learning without requiring explicit motion generation at deployment. The adapted prior model is frozen during action learning; the action expert combines visual context generated by the adapted prior model with current observations through a learned gate. This separation allows human-centric prediction to be learned without robot action labels, while robot supervision determines how the resulting context informs executable commands. The gate makes this contribution conditional on the expert's current observation-dependent features.

3 METHOD

Being-M0.7 learns a latent video–motion prior from human-centric data, adapts it to robot trajectories, and then trains a gated action expert against the frozen prior model. We first formulate the problem in Section 3.1, describe the mixed-modality data and shared representations in Section 3.2, and present the architectures of the prior model and action expert in Section 3.3. We then detail the pre-training, robot mid-training, post-training, and closed-loop inference recipes (Figure 2).

3.1 PROBLEM FORMULATION

Let V be egocentric video, $Z = \mathcal{E}(V)$ its visual latents, M a compact motion sequence, and I the task instruction. Here, $\mathcal{E}$ is a frozen DINO encoder (Siméoni et al., 2025) whose latents serve as both visual inputs and prediction targets, avoiding pixel-level reconstruction. Each sample is a temporal window of T consecutive time steps from a video and/or motion sequence, with observed history indexed by $\mathcal{C} = 1, \ldots, K$ and future targets indexed by $\mathcal{Q} = K+1, \ldots, T$. Details appear in Appendix B.2.

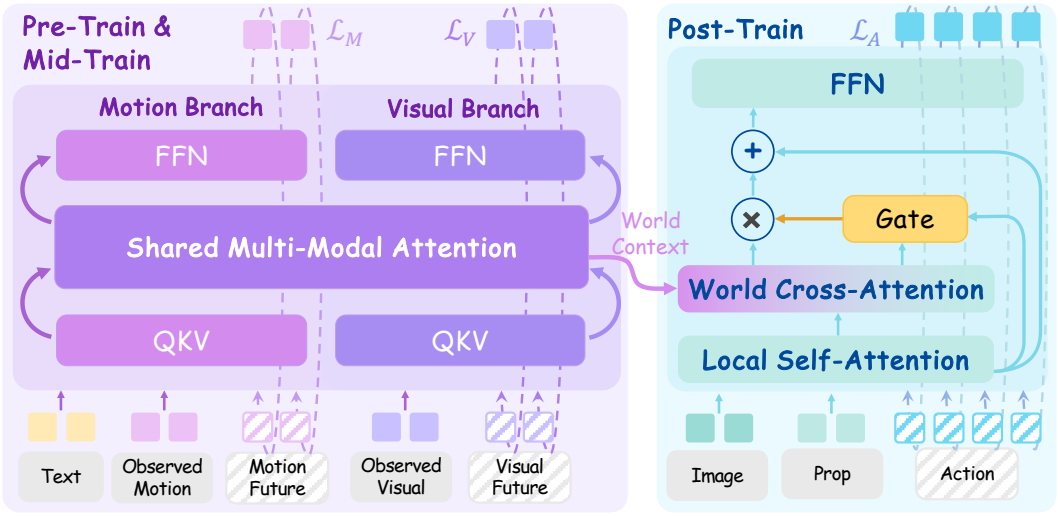

Figure 2: **Being-M0.7 framework.** Mixed-modality pre-training learns a visual–motion prior with a shared human–humanoid motion representation. Robot mid-training adapts the prior model; gated post-training learns an action expert from its frozen visual context and current robot observations.

During pre-training and mid-training, the prior model represents $p_\theta(Z_\mathcal{Q}, M_\mathcal{Q} \mid Z_\mathcal{C}, M_\mathcal{C}, I)$, with motion providing coarse-grained supervision of whole-body dynamics. Joint video–motion modeling encourages the visual future-token representations to capture both visual evolution and the underlying kinematic changes. During post-training and deployment, the prior model generates visual futures from $(Z_\mathcal{C}, I)$ with explicit motion prediction disabled. The resulting future-token representations carry learned information about future kinematics and serve as policy conditioning for a future-conditioned action expert (Xie et al., 2023), which combines this context with the current robot observation O_q to predict fine-grained actions in a chunk $\mathbf{a}_q = a_{q:q+h-1}$.

3.2 Data Curation

Our raw collection contains more than 10,000 hours before filtering and temporal segmentation. It yields three supervision streams: paired egocentric video–motion data $\mathcal{D}_{VM}$, video-only data $\mathcal{D}_V$, and motion-only data $\mathcal{D}_M$. Paired samples supervise joint prediction; unpaired samples supervise the corresponding modality with shared model parameters. This extends training beyond the paired corpus while retaining cross-modal interaction wherever both modalities are observed. Figure 3 summarizes the data sources and supervision streams; Appendix A.1 details motion processing.

3.3 Model Architecture

Being-M0.7 combines a video–motion prior model with a separate future-conditioned action expert. The expert combines predictive representations learned by the prior model with current robot observations to generate executable commands for downstream robot control tasks.

Video-Motion Mixture of Transformers. We implement the prior model as a Mixture of Transformers (MoT) (Liang et al., 2024). Separate input projections, temporal embeddings, and modality embeddings map visual and motion tokens into a common hidden space. Each block retains modality-specific parameters but combines the streams through shared attention, allowing visual-state evolution and motion dynamics to inform one another. Both streams cross-attend to task instruction features from a frozen DistilBERT encoder (Sanh et al., 2019). Modality-specific heads predict flow velocities in the visual latent space and the motion space, respectively.

A chunk attention mask prevents history tokens from accessing future targets, while noisy future tokens attend to the history and to one another. The prior model therefore denoises a complete future chunk in parallel. Only available modalities contribute supervision. Missing-modality tokens are masked out of interactions with valid tokens; Appendix A.2 specifies the token and attention

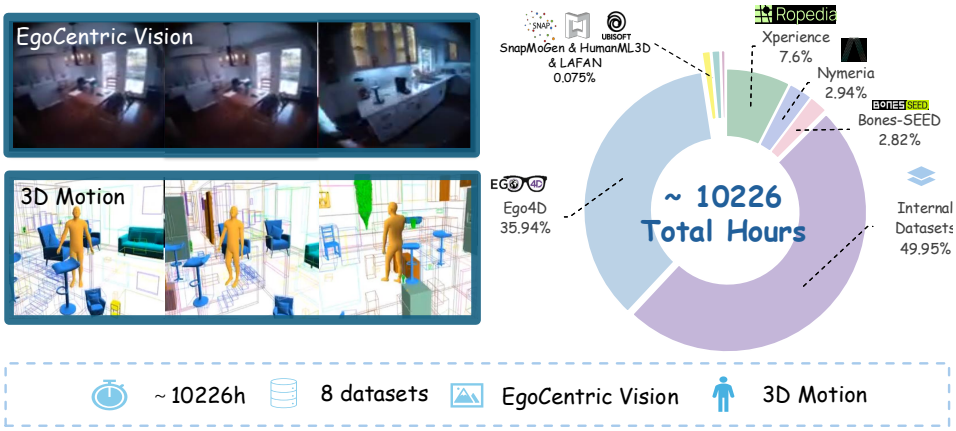

Figure 3: **Being-M0.7 data recipe.** The pre-training corpus is sourced from external public datasets including Ego4D (Grauman et al., 2022), Xperience (Ropedia, 2026), Nymeria (Ma et al., 2024), Bones-SEED (Luo et al., 2025), SnapMoGen (guo et al., 2025), HumanML3D (Guo et al., 2022), and Lafan1 (Harvey et al., 2020), as well as additional data comprising subsets of the training corpora described in Being-H0.5 (Luo et al., 2026a) and Being-M0.5 (Cao et al., 2025; 2026), and commercially acquired datasets. Raw sources are filtered, temporally segmented, and reformatted into three supervision streams–$\mathcal{D}_{VM}$ for video-motion pairs, $\mathcal{D}_V$ for video-only, and $\mathcal{D}_M$ for motion-only–which jointly supervise the visual-motion joint distribution and its marginals, respectively.

handling. Video-only data supervises visual dynamics across diverse scenes, motion-only data supplies kinematic structure, and paired data connects that structure to visual dynamics. The resulting visual representations provide the downstream expert with motion-informed predictive context without requiring explicit motion generation during control. See Appendices A.2 and B.1 for details.

Future-Conditioned Action Expert. The action expert is a separate flow-based transformer. It receives noisy action tokens and the current observation O_q, including the egocentric image and proprioceptive state. The image is encoded using the same frozen DINO encoder, while proprioception is mapped using a learned projection. Local self-attention first combines action and observation tokens. At selected layers, gated cross-attention then injects hidden states $H^\theta(\tau)$ from the frozen prior model, exposing motion-informed predictive context at multiple network depths. Information flows only from the prior model to the expert: action tokens never enter the prior model's computation.

Gated World Context. Let $x_{i,j}$ be action-token i after local observation processing at expert layer j, and $r_{i,j}$ its projected world cross-attention update. We gate the injected residual as

$$\begin{aligned} g_{i,j} &= \sigma\big(w_j^\top [\mathrm{LN}(x_{i,j}); \mathrm{LN}(r_{i,j})] + b_j\big)\,, \\ \Delta x_{i,j} &= g_{i,j}\big(\gamma_j(\tau) \odot r_{i,j}\big), \end{aligned} \tag{1}$$

where $\gamma_j(\tau)$ is the flow-conditioned residual scale. The goal of the token-wise gate is to learn how much predictive context to incorporate given the local state and proposed update, allowing the influence of predictive context to vary across tokens. The gate controls only the contribution of predictive context; action tokens retain access to current observations through local self-attention.

3.4 Training and Inference Recipes

Pre-Training. We train visual and motion prediction with flow matching (Lipman et al., 2023). For clean future tokens $\boldsymbol{x}_0$, sample $\epsilon \sim \mathcal{N}(0, I)$ and $\tau \sim \mathcal{U}(0, 1)$, and set $\boldsymbol{x}_\tau = (1-\tau)\boldsymbol{x}_0 + \tau\epsilon$. For each modality $k \in \{V, M\}$, we train the model to predict the target flow velocity with the loss

$$\mathcal{L}_k = \mathbb{E}_{\boldsymbol{x}_0^k, \epsilon, \tau}\left[\left\| u_\theta^k(\boldsymbol{x}_\tau^k, \tau, c_s) - (\epsilon - \boldsymbol{x}_0^k)\right\|_2^2\right], \tag{2}$$

where c_s contains the clean history and task instruction. Each modality-specific loss is computed only when that modality is observed. The overall objective is

$$\begin{aligned}\mathcal{L}_{\text{pre}} = \mathbb{E}_{\mathcal{D}_{VM}}[\mathcal{L}_V + \mathcal{L}_M + \mathcal{L}_{\text{geom}}] + \mathbb{E}_{\mathcal{D}_V}[\mathcal{L}_V] \\ + \mathbb{E}_{\mathcal{D}_M}[\mathcal{L}_M + \mathcal{L}_{\text{geom}}],\end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{geom}}$ regularizes global trajectories and local end-effector velocities whenever motion is available; Appendix B.3 defines both auxiliary terms.

Mid- and Post-Training. Both stages use the same robot trajectory dataset, first to adapt the prior model and then to train the action expert. During mid-training, we adapt the prior model to robot video, motion, and task text using $\mathcal{L}_{\text{mid}} = \mathcal{L}_V + \mathcal{L}_M$, where $\mathcal{L}_M$ is only applied when motion is available. The shared motion representation transfers the pre-training objectives to robot scenes and body dynamics. We then freeze the adapted prior model for post-training with action supervision.

During post-training, we select an action query q in the future interval of each robot trajectory window. Conditioned on the observed visual history $Z_{\mathcal{C}}$ and task instruction I, the frozen prior model initializes future visual latents from Gaussian noise and integrates its flow ODE to generate a visual future, with explicit motion prediction disabled. Intermediate hidden states along this denoising trajectory provide predictive context $H^\theta(\tau)$ for the action expert, which also receives the current observation O_q. This trains the expert on the prior model's own predictions, reducing the conditioning mismatch with inference. Further details can be found in Appendix B.4.

Following action chunking (Zhao et al., 2023), we corrupt each demonstrated action chunk as $\tilde{\mathbf{a}}_{q,\tau} = (1-\tau)\mathbf{a}_q + \tau\epsilon_a$, where ϵ_a is Gaussian noise, and optimize the action predictor with

$$\mathcal{L}_A = \mathbb{E}\left[\left\| u^a_\phi(\tilde{\mathbf{a}}_{q,\tau}, \tau, H^\theta(\tau), O_q) - (\epsilon_a - \mathbf{a}_q)\right\|^2_2\right]. \tag{4}$$

The frozen latent WAM receives no action-loss gradients in either the visual or motion branch.

Inference. The prior model periodically generates a visual-latent plan from recent images and task text. Its intermediate states are cached as policy-side keys and values and reused across action queries. At each query, the expert receives fresh images and proprioception and denoises an action chunk using the gated cache. The cache is refreshed before its context is fully consumed, separating lower-frequency planning from responsive command generation. Details are provided in Appendix B.5.

4 Experiments

4.1 Experimental Setup

4.1.1 Hardware Setup

Our real-world system is built upon a Unitree G1 humanoid robot for whole-body dexterous teleoperation. The platform integrates dual 7-DoF arms with Linker Hand O6 dexterous hands for upper-body manipulation, and an Intel RealSense D435i RGB camera mounted on the robot head to provide egocentric visual observations. Policy inference runs on an external workstation with a single NVIDIA RTX 4090 GPU. During deployment, a low-level controller communicates with the policy server via ZMQ in a closed loop and executes predicted commands on the robot in real time.

4.1.2 VR-Based Whole-Body Teleoperation Pipeline

We collect real-robot demonstrations using a VR-based whole-body teleoperation system. The operator wears a PICO VR headset and two ankle-mounted PICO trackers and holds two controllers to provide full-body motion input. XRoboToolkit (Zhao et al., 2026) estimates the operator's SMPL pose in real time from the VR device streams, while SONIC (Luo et al., 2025) converts the estimated pose into stable 29-DoF whole-body control commands executed on the Unitree G1 at 50 Hz. During teleoperation, synchronized robot proprioception, egocentric RGB observations, and motion commands (including hand control signals and SONIC motion tokens) are recorded to construct the multi-task post-training dataset. Figure 4 provides an overview of the proposed system.

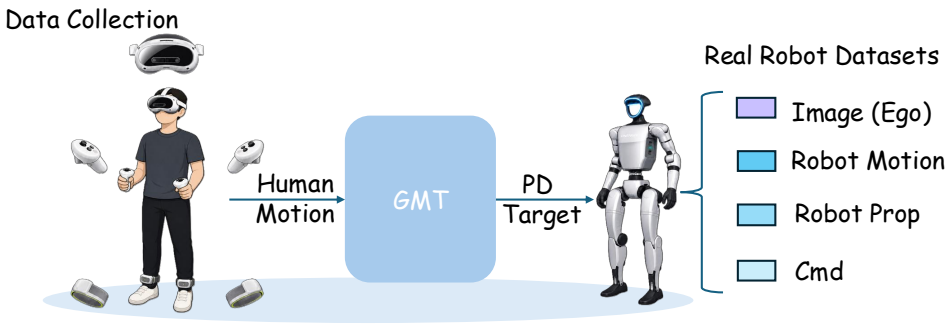

Figure 4: **Overview of the data collection system.** During the data collection stage, an operator uses a VR-based teleoperation system to generate low-level motion commands, and a pretrained whole-body motion tracker controls the humanoid robot. We record synchronized real-world trajectories comprising egocentric images, proprioceptive observations, and motion commands.

4.1.3 Baselines and Multi-Task Training

For simulation and real-world evaluation, we use three vision-language-action baselines: GR00T-N1.6 (Bjorck et al., 2025), Ψ_0 (Wei et al., 2026a), and $\pi_{0.5}$ (Intelligence et al., 2025).

All experiments use a multi-task training setting, including Being-M0.7, all baselines, and all ablation variants. In simulation, each method or variant is trained jointly on demonstrations from all six SIMPLE tasks, with a single model evaluated across the six tasks. For real-world experiments, each method is trained jointly on demonstrations from all three real-world tasks, with a single model evaluated across the three tasks. In both settings, the model is conditioned on the corresponding task instruction during evaluation to specify the task to be performed.

4.2 Simulation Experiments

We evaluate Being-M0.7 on the SIMPLE benchmark (Wei et al., 2026b), which provides a controlled environment for measuring whole-body locomotion and manipulation. We evaluate six tasks at three domain-randomization levels, from easy to hard, with ten trials per task and level: **BendPick**, **Handover**, **Mobile P&P**, **Grasp**, **XMoveBendPick**, and **XMovePick**.

Table 1: **Simulation results on SIMPLE.** Entries report successes out of ten trials at easy / medium / hard domain-randomization levels. Total aggregates six tasks and three levels (180 trials).

Method	Bend Pick	Handover	Mobile P&P	Grasp	XMove BendPick	XMove Pick	Total
GR00T-N1.6	1/7/3	2/3/3	0/0/0	7/8/7	6/6/8	0/0/0	61/180
Ψ_0	10/8/9	6/8/6	0/0/0	10/9/9	7/8/5	9/9/3	116/180
$\pi_{0.5}$	0/0/0	5/7/7	7/6/6	7/8/5	10/9/7	6/2/3	95/180
Being-M0.7 (Ours)	7/7/8	9/9/10	7/4/2	10/10/8	2/4/3	10/10/8	**128/180**

Main Comparison with Baselines. Table 1 reports successes per setting. Being-M0.7 achieves the highest aggregate success rate, with $128/180$ successes (71.1%), compared with $116/180$ (64.4%) for Ψ_0, $95/180$ (52.8%) for $\pi_{0.5}$, and $61/180$ (33.9%) for GR00T-N1.6. Our method achieves $9/9/10$ on Handover and $10/10/8$ on both Grasp and XMovePick. On easy Mobile P&P, it achieves $7/10$, matching $\pi_{0.5}$ and exceeding GR00T-N1.6 and Ψ_0, both at $0/10$. When results are aggregated across difficulty levels, Ψ_0 leads on BendPick, while $\pi_{0.5}$ leads on Mobile P&P and XMoveBendPick.

4.3 Ablation Study

We evaluate three variants on the same six SIMPLE tasks and three difficulty levels (180 trials per variant): **w/o pre-training** omits the prior model's mixed-modality pre-training stage; **w/o mid-**

Table 2: **Ablation results on SIMPLE.** Entries report successes out of ten trials at easy / medium / hard domain-randomization levels (180 trials total). Language-only context replaces prior predictions with task-language features; the expert retains current images and proprioception.

Variant	Bend Pick	Handover	Mobile P&P	Grasp	XMove BendPick	XMove Pick	Total
Being-M0.7 (Full)	7/7/8	9/9/10	7/4/2	10/10/8	2/4/3	10/10/8	**128/180**
w/o pre-training	0/0/0	9/10/9	2/2/3	8/8/5	5/2/4	8/6/7	88/180
w/o mid-training	2/0/0	9/9/8	7/3/4	8/9/7	5/6/4	9/7/10	107/180
Language-only context	4/4/5	8/9/8	4/4/2	9/10/9	7/5/9	4/2/2	105/180

training omits its adaptation to robot trajectories; and **language-only context** replaces the upstream prior model's predictive context supplied to the action expert with language features from the task instruction. In the language-only variant, the expert still receives the current egocentric image and proprioceptive state O_q; only the upstream conditioning context is replaced.

Contribution of Training Stages. As shown in Table 2, removing pre-training reduces aggregate success from $128/180$ (71.1%) to $88/180$ (48.9%), a drop of 22.2 percentage points. Removing robot mid-training yields $107/180$ (59.4%), a drop of 11.7 points. The largest task-level degradation in both variants occurs on BendPick, where successes fall from $22/30$ to $0/30$ and $2/30$, respectively. These aggregate results support the contributions of both video–motion prior learning during pre-training and subsequent robot-domain adaptation during mid-training.

Utility of Upstream Predictive Context. Replacing the prior model's output context with language-only features reduces aggregate success to $105/180$ (58.3%), a drop of 12.8 percentage points from the full model. The full model improves over this variant on four of the six tasks, with the largest gain on XMovePick ($28/30$ versus $8/30$). Language-only context performs better on XMoveBendPick ($21/30$ versus $9/30$), while both variants obtain $28/30$ on Grasp. Thus, the benefit varies across tasks, but the aggregate improvement indicates that the upstream prior model provides useful predictive information for downstream action generation beyond task language and current robot observations. We discuss possible interference among visually similar tasks in Appendix C.3.

4.4 Real-World Experiments

We quantitatively evaluate Being-M0.7 on three real-world loco-manipulation tasks: Walk to Fish **(W2F)**, Walk to Mirror **(W2M)**, and Walk to Doll **(W2D)**. All three tasks require the robot to approach the target before manipulating it, testing coordinated locomotion and manipulation alongside tool use, reflection-based reasoning, or direct grasping. We compare with GR00T-N1.6 (Bjorck et al., 2025), Ψ_0 (Wei et al., 2026a), and $\pi_{0.5}$ (Intelligence et al., 2025). Figure 5 illustrates these tasks.

Table 3: **Quantitative real-world evaluation.** Each task is evaluated over five trials. Entries report successful trials / total trials; bold numbers indicate the best result, including ties.

Method	W2F	W2M	W2D	Overall
GR00T-N1.6	**3/5**	**5/5**	**5/5**	**13/15**
Ψ_0	0/5	1/5	2/5	3/15
$\pi_{0.5}$	2/5	2/5	3/5	7/15
Being-M0.7 (Ours)	**3/5**	**5/5**	**5/5**	**13/15**

Quantitative Evaluation. We evaluate each method over five trials per task, totaling fifteen trials. Physical markers fix the robot starting positions and table positions across methods, while target objects are randomly initialized across the left, center, and right regions of the workspace. Each trial has a 30-second limit and is judged by a human evaluator: success requires grasping the doll in W2M and W2D, or scooping up the toy fish in W2F. As shown in Table 3, Being-M0.7 achieves $13/15$ successful trials (86.7%), matching GR00T-N1.6 and exceeding $\pi_{0.5}$ ($7/15$) and Ψ_0 ($3/15$).

- **Walk to Fish.** The robot approaches the water tank and uses a handheld net to scoop up a toy fish. This evaluates the complete locomotion-and-tool-use sequence. Being-M0.7 and GR00T-N1.6 each succeed in $3/5$ trials, compared with $2/5$ for $\pi_{0.5}$ and $0/5$ for Ψ_0.
- **Walk to Mirror.** The robot approaches a table, infers the location of a hidden toy from its mirror reflection, and grasps it. This task combines visual reasoning under partial observability with whole-body positioning and manipulation. Being-M0.7 and GR00T-N1.6 each succeed in $5/5$ trials, compared with $2/5$ for $\pi_{0.5}$ and $1/5$ for Ψ_0.
- **Walk to Doll.** The robot approaches a table and directly grasps a visible doll. This task evaluates coordinated locomotion and target-directed grasping during the transition from approach to manipulation, without the mirror-based observation requirement. Being-M0.7 and GR00T-N1.6 each succeed in $5/5$ trials, compared with $3/5$ for $\pi_{0.5}$ and $2/5$ for Ψ_0.

Qualitative Case Studies. Figure 5 shows execution sequences for the three evaluated tasks. From top to bottom, W2M shows approaching the table and grasping a hidden doll using mirror observations; W2F shows approaching the water tank and scooping a toy fish with a net; and W2D shows approaching the table and grasping a directly visible doll. These sequences illustrate the transition from locomotion to manipulation in the corresponding task settings. Detailed task descriptions and additional snapshots are provided in Appendix D.1.

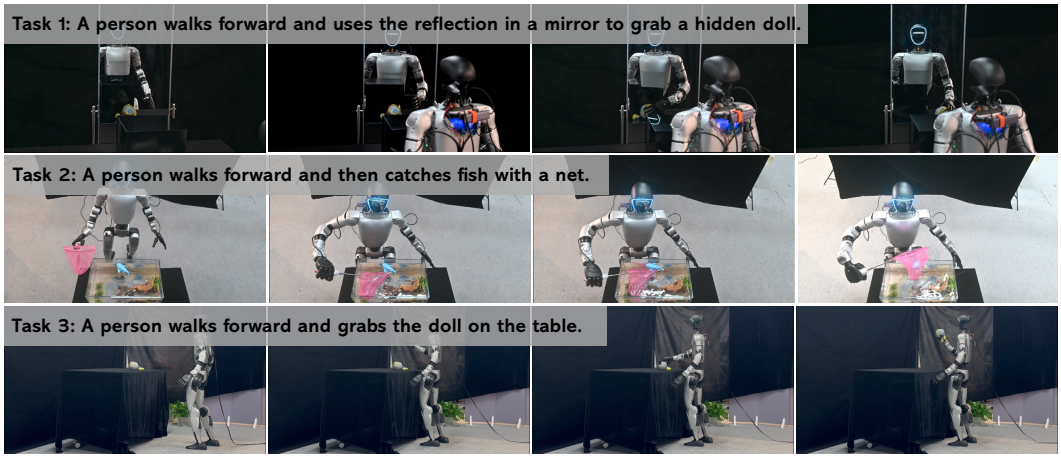

Figure 5: **Representative real-world execution sequences.** Rows show the three evaluated tasks, from top to bottom: **Walk to Mirror (W2M)**, approaching the table and grasping a hidden doll using mirror observations; **Walk to Fish (W2F)**, approaching the tank and scooping a toy fish with a handheld net; and **Walk to Doll (W2D)**, approaching the table and grasping a directly visible doll.

5 Conclusion

We presented Being-M0.7, a latent world-action model that transfers visual–motion priors learned from mixed-modality human data to humanoid loco-manipulation. Joint latent visual and motion prediction learns a predictive prior, robot mid-training adapts it to robot trajectories, and a gated action expert combines the frozen prior's predictive context with current observations to generate executable whole-body commands for body control. Experiments on SIMPLE and real-world Unitree G1 tasks demonstrate the effectiveness of this framework, while ablations support the contributions of pre-training, robot mid-training, and predictive context to downstream control.

Limitations and Future Work. Our evaluation covers a limited range of task horizons, and performance on longer-horizon loco-manipulation tasks remains to be established. In our current setup, the SONIC whole-body controller also has limited end-effector tracking accuracy, which makes robot demonstration collection challenging. Future work will extend the framework to longer-horizon tasks and refine SONIC to improve tracking precision and facilitate demonstration collection.

AI USE STATEMENT

Generative AI tools were used to polish the language, check for grammatical errors, and generate visual elements for Figure 1. The authors take responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

The method describes the data representation, model architecture, and training objectives. Appendix B.1 and Appendix B.2 report architecture and training hyperparameters; Appendix B.3 defines the auxiliary losses. Appendix B.5 describes the inference procedure, and Appendix D.1 provides real-world task descriptions.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Qingwei Ben, Feiyu Jia, Jia Zeng, Junting Dong, Dahua Lin, and Jiangmiao Pang. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. *arXiv preprint arXiv:2502.13013*, 2025.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi 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.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Bin Cao, Sipeng Zheng, Ye Wang, Lujie Xia, Qianshan Wei, Qin Jin, Jing Liu, and Zongqing Lu. Being-m0. 5: A real-time controllable vision-language-motion model. *arXiv preprint arXiv:2508.07863*, 2025.

Bin Cao, Sipeng Zheng, Hao Luo, Boyuan Li, Jing Liu, and Zongqing Lu. Opent2m: No-frill motion generation with open-source, large-scale, high-quality data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30640–30649, 2026.

Lu Chen, Yizhou Wang, Shixiang Tang, Qianhong Ma, Tong He, Wanli Ouyang, Xiaowei Zhou, Hujun Bao, and Sida Peng. Acquisition through my eyes and steps: A joint predictive agent model in egocentric worlds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2502.05857.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023.

Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. In *Conference on Robot Learning*, pp. 2828–2844. PMLR, 2025.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5152–5161, 2022.

chuan guo, Inwoo Hwang, Jian Wang, and Bing Zhou. Snapmogen: Human motion generation from expressive texts. In *Advances in Neural Information Processing Systems*, volume 38, pp. 99939–99955, 2025.

Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020.

Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris M Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *Conference on Robot Learning*, pp. 1516–1540. PMLR, 2025.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023.

Haoran Jiang, Jin Chen, Qingwen Bu, Li Chen, Modi Shi, Yanjie Zhang, Delong Li, Chuanzhe Suo, wang chuang, zhihui peng, and Hongyang Li. WholebodyVLA: Towards unified latent VLA for whole-body loco-manipulation control. In *The Fourteenth International Conference on Learning Representations*, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning*, pp. 2679–2713. PMLR, 2025.

Jiaman Li, C. Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2212.04636.

Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024. URL https://arxiv.org/abs/2411.04996.

Yutang Lin, Jieming Cui, Yixuan Li, Baoxiong Jia, Yixin Zhu, and Siyuan Huang. Lessmimic: Long-horizon humanoid interaction with unified distance field representations. *arXiv preprint arXiv:2602.21723*, 2026.

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.

Hao Luo, Ye Wang, Wanpeng Zhang, Sipeng Zheng, Ziheng Xi, Chaoyi Xu, Haiweng Xu, Haoqi Yuan, Chi Zhang, Yiqing Wang, et al. Being-h0. 5: Scaling human-centric robot learning for cross-embodiment generalization. *arXiv preprint arXiv:2601.12993*, 2026a.

Hao Luo, Wanpeng Zhang, Yicheng Feng, Sipeng Zheng, Haiweng Xu, Chaoyi Xu, Ziheng Xi, Yuhui Fu, and Zongqing Lu. Being-h0. 7: A latent world-action model from egocentric videos. *arXiv preprint arXiv:2605.00078*, 2026b.

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *arXiv preprint arXiv:2511.07820*, 2025.

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, 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.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

Chaitanya Patel, Hiroki Nakamura, Yuta Kyuragi, Kazuki Kozuka, Juan Carlos Niebles, and Ehsan Adeli. Uniegomotion: A unified model for egocentric motion reconstruction, forecasting, and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2508.01126.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2212.09748.

Ropedia. Xperience-10M: A large-scale egocentric multimodal dataset with structured 3D/4D annotations. Hugging Face dataset, 2026. https://huggingface.co/datasets/ropedia-ai/xperience-10m, accessed July 11, 2026.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. URL https://arxiv.org/abs/1910.01108.

Modi Shi, Shijia Peng, Jin Chen, Haoran Jiang, Tianyu Li, Di Huang, Ping Luo, Hongyang Li, and Li Chen. Egohumanoid: Unlocking in-the-wild loco-manipulation with robot-free egocentric demonstration. *arXiv preprint arXiv:2602.10106*, 2026.

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.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

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.

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.

Haoyang Weng, Yitang Li, Nikhil Sobanbabu, Zihan Wang, Zhengyi Luo, Tairan He, Deva Ramanan, and Guanya Shi. Hdmi: Learning interactive humanoid whole-body control from human videos. *arXiv preprint arXiv:2509.16757*, 2025.

Zhihui Xie, Zichuan Lin, Deheng Ye, Qiang Fu, Yang Wei, and Shuai Li. Future-conditioned unsupervised pretraining for decision transformer. In *International Conference on Machine Learning*, pp. 38187–38203. PMLR, 2023.

Jingqiao Xiu, Fangzhou Hong, Yicong Li, Mengze Li, Wentao Wang, Sirui Han, Liang Pan, and Ziwei Liu. Egotwin: Dreaming body and view in first person. *arXiv preprint arXiv:2508.13013*, 2025. URL https://arxiv.org/abs/2508.13013.

Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, Yu-Xiong Wang, and Liang-Yan Gui. InterAct: Advancing large-scale versatile 3d human-object interaction generation. In *CVPR*, 2025.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2410.03665.

Shaofeng Yin, Yanjie Ze, Hong-Xing Yu, C Karen Liu, and Jiajun Wu. Visualmimic: Visual humanoid loco-manipulation via motion tracking and generation. *arXiv preprint arXiv:2509.20322*, 2025.

Siheng Zhao, Yanjie Ze, Yue Wang, C Karen Liu, Pieter Abbeel, Guanya Shi, and Rocky Duan. Resmimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning. *arXiv preprint arXiv:2510.05070*, 2025.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.016.

Zhigen Zhao, Liuchuan Yu, Ke Jing, and Ning Yang. Xrobotoolkit: A cross-platform framework for robot teleoperation. In *2026 IEEE/SICE International Symposium on System Integration (SII)*, pp. 15–20. IEEE, 2026.

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.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

A Additional Method Details

A.1 Data Sources and Motion Processing

Figure 3 summarizes the raw pre-training corpus and its three supervision streams. Source hours refer to the collection before filtering and temporal segmentation.

Head-root motion vector. The shared 22D representation uses the head as the root for both human and humanoid motion. In storage order, the vector contains root translation (dimensions 1–3), root rotation in a continuous 6D representation (4–9), initial head height h_0 (10), and the local 3D positions of the left hand (11–13), right hand (14–16), left foot (17–19), and right foot (20–22). The 6D rotation encodes two axes of the head orientation; the decoder orthonormalizes them and obtains the third axis by a cross product. Positions and heights are measured in meters, and rotation coordinates are dimensionless.

Let p_t and R_t be the global head position and orientation, and p_t^e the global position of end-effector e. The canonical frame is floor-aligned with y as the vertical axis. Its origin lies below the initial head position, so $p_0 = (0, h_0, 0)$. Human motion is expressed in this head-root frame by retaining the head pose and transforming the two hand and two foot positions as

$$\ell_t^e = R_t^\top (p_t^e - p_t), \qquad p_t^e = p_t + R_t \ell_t^e.$$

Robot motion uses the same component order and coordinate convention; the robot data loader reads the preprocessed 22D trajectories stored alongside each recorded episode. The shared representation therefore describes head motion and end-effector positions without requiring identical human and robot joint-angle parameterizations.

Translation encoding and resampling. The stored translation channels contain frame-to-frame head displacements $\delta p_t = (\delta x_t, \delta y_t, \delta z_t)$ in the canonical global frame. The initial displacement is reset to zero when constructing a robot window. With `use_root_pos=true`, these displacements are accumulated before subsampling to 5 Hz, so the model input contains relative head positions d_t rather than per-frame displacements:

$$d_t = \sum_{s \le t} \delta p_s, \qquad p_t = d_t + (0, h_0, 0).$$

The rotation, initial-height, and local-position channels retain their meanings during resampling. Thus, the first ten model-input dimensions are $[d_t; \mathrm{rot6D}(R_t); h_0]$, followed by the twelve local-position dimensions.

Normalization. Human-motion loaders apply per-channel standardization $(m - \mu)/\sigma$ when the dataset's stored mean and standard deviation are available; standard deviations below 10^{-5} are replaced by one. If these statistics are absent, the loader retains the unstandardized motion values. The robot mid-training loader uses unstandardized 22D motion vectors. This motion preprocessing is separate from post-training normalization of proprioception and executable commands, whose statistics are computed from the robot training split.

A.2 Architecture Details

Input Embeddings. We map per-frame visual latents z_t and motion vectors m_t from Section 3.1 into a common MoT hidden space using separate learned projections. Appendix A.1 details the motion representation. Temporal positional embeddings encode their aligned time indices, while modality-type embeddings distinguish visual and motion inputs. For video-only or motion-only samples, supervision is applied only to the available modality. The task instruction I is encoded separately by the frozen DistilBERT (Sanh et al., 2019) and supplied as language-conditioning features.

Missing Modalities. The model keeps a fixed interleaved layout of visual and motion token slots. For a video-only sample, projected motion features are replaced by a learned missing-motion embedding; for a motion-only sample, projected visual features are replaced by a learned missing-image embedding. Modality-presence flags are combined with temporal-validity masks to define

the valid history and future tokens. Valid queries cannot attend to absent-modality keys or values. An invalid query is allowed to attend only to itself in this shared attention operation, preventing an all-masked attention row; its output is excluded from supervision and from the context supplied to the action expert. Hence, placeholder tokens do not exchange information with valid visual or motion tokens, while paired samples retain cross-modal attention.

During action post-training and deployment, `use_robot_motion=false` sets all motion-presence flags to false. The motion state is initialized to zero and receives no flow updates; only visual future states are integrated and cached for the expert. The implementation still evaluates the shared MoT forward pass, including the masked motion slots and motion output projection, whose predictions are discarded. Disabling motion prediction therefore removes motion conditioning, motion-state generation, and motion supervision, without pruning all motion-branch computation.

MoT Blocks. The prior model is implemented as a Mixture-of-Transformers (MoT) (Liang et al., 2024) with modality-specific computation and shared multimodal attention. In each block, latent visual and motion tokens use separate layer normalizations, query/key/value and output projections, and feed-forward networks. Their projected queries, keys, and values are then combined in a shared attention operation, allowing every valid latent visual token to exchange information with motion tokens across time while retaining modality-specific parameterization. Both streams additionally cross-attend to the instruction features. A sinusoidal embedding of the flow timestep τ modulates each modality's attention, language cross-attention, and feed-forward computation through modality-specific adaptive layer normalization (Peebles & Xie, 2023).

The first K video-motion frames are clean context tokens, while tokens in the future interval are corrupted at the sampled flow timestep. We use a chunk attention mask under which context tokens attend within the observed history but cannot access the future, whereas all noisy future tokens jointly attend to the full history and to one another. This non-causal target interaction lets the model denoise the entire future video-motion chunk in parallel. Finally, modality-specific adaptive normalization and output projections produce velocity fields in the visual latent space and the continuous motion space, respectively.

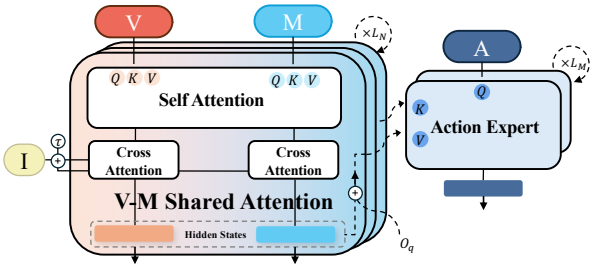

Figure 6: **Attention interfaces.** V, M, I, and A denote vision, motion, instruction, and action hidden states; τ is the flow timestep. The prior model learns from visual and motion streams during pre-training and mid-training. During action post-training and inference, only its visual hidden states are supplied to the expert, alongside current robot observations O_q. The gated residual is defined in Eq. 1.

Layer-Wise Action Interface. At an expert layer connected to the prior model, local self-attention first processes the action and current-observation tokens. The subsequent world cross-attention uses only the noisy action tokens as queries; proprioceptive tokens do not query the prior model or receive this cross-attention residual. Proprioception still conditions action generation through local self-attention. Keys and values are constructed from the generated future visual-token hidden states before the corresponding layer of the prior model, excluding the prior model's observed-history input tokens. The projected cross-attention residual is applied only to the action tokens and gated as in Eq. 1. This one-way connection allows the expert to access multiple representation depths without sending action information into the frozen prior model. Selected prior model layers and model dimensions are reported in Appendix B.1.

B Implementation Details

B.1 Architecture Hyperparameters

Table 4 reports the model sizes used in our implementation. Layer indices of the prior model are one-based.

Table 4: Architecture hyperparameters of the video–motion prior model and action expert.

Hyperparameter	Value
Video–Motion Prior Model	
Transformer layers	30
Hidden width	768
Attention heads	24
Visual FFN expansion ratio	6
Motion FFN expansion ratio	2
Dropout	0.1
Action Expert	
Transformer layers	30
Hidden width	1024
Attention heads	32
FFN expansion ratio	4
Connected prior model layers	$3, 6, 9, \ldots, 30$
Action chunk length	30
Dropout	0.1

B.2 Hyperparameters

Table 5 reports the temporal and optimization settings. Effective batch sizes include all eight GPUs, with one optimizer update per batch and no gradient accumulation. The robot-stage settings follow the real-world mid-training and post-training configurations.

Training data and sampling. Pre-training combines paired video–motion, video-only, and motion-only samples. The mixed-data loader shuffles the concatenated datasets without explicit source reweighting (`dataset_weights=null`); therefore, the stream proportions are determined by the numbers of valid examples after filtering, rather than a fixed equal-modality mixture. Both robot stages use the same training and held-out episode lists. Mid-training reads video, text, and 22D motion, whereas post-training reads video, text, current proprioception, and demonstrated action chunks.

Optimization and initialization. The DINO and DistilBERT encoders remain frozen in all stages. Pre-training and mid-training update the prior model; mid-training starts from the last saved pre-training checkpoint with a fresh optimizer and learning-rate scheduler. The post-training configuration initializes the prior from a mid-training checkpoint at 50,000 optimizer updates, freezes it in evaluation mode, and trains only the action expert and its observation/context projections. All stages use linear learning-rate warmup followed by cosine decay. Pre-training checkpointing monitors validation local end-effector position error and retains the two best checkpoints and the latest checkpoint. Robot-stage checkpointing records validation loss and retains all saved checkpoints; the post-training metric is the command flow-matching loss.

Visual and language features. For the action expert's current image, we instead use the patch tokens from the final hidden layer, excluding the classification and register tokens. Images are resized to 224×224 and normalized with ImageNet channel statistics. The frozen `distilbert-base-uncased` encoder supplies final-layer token features, with instructions truncated to at most 64 tokens and padding masked in language cross-attention.

Table 5: Temporal and training settings. The action-query interval is determined by executing a complete 30-command chunk at 50 Hz.

Temporal setting	**Value**
Sequence length T / observed frames K	25 / 1
Generated future frames / horizon	24 / 4.8 s
Input resolution	224×224
Prior prediction frame rate	5 Hz
Action chunk length H_a	30
Action-query interval / rate	0.6 s / ≈ 1.67 Hz
Command-execution interval / rate	0.02 s / 50 Hz
Cache refresh	Before the prediction window is exhausted
Prior sampling steps during post-training	3
Prior / action sampling steps at inference	10 / 10

Optimization setting	**Pre-training**	**Mid-training**	**Post-training**
GPUs	8	8	8
Batch size per GPU	256	64	64
Effective batch size	2048	512	512
Learning rate	2×10^{-4}	10^{-4}	10^{-4}
Optimizer	AdamW	AdamW	AdamW
Weight decay	0.05	0.05	0.05
Gradient norm clipping	1.0	1.0	1.0
Precision	FP16 mixed	FP16 mixed	FP16 mixed
Trainable module	Prior	Prior	Action expert
$\lambda_{\text{local-vel}}$	1.0	0	—
World-context dropout	—	—	0.1

B.3 Geometry-aware Auxiliary Losses

During pre-training, the motion branch uses two geometry-aware auxiliary losses in addition to the flow matching loss:

$$\mathcal{L}_{\text{geom}} = \lambda_{\text{traj}}\mathcal{L}_{\text{traj}} + \lambda_{\text{local-vel}}\mathcal{L}_{\text{local-vel}}.$$

The clean-motion estimate can be recovered from the predicted velocity as

$$\hat{m}_{0,t} = m_{\tau,t} - \tau u_\theta^M(m_{\tau,t}, \tau, c, \mathcal{C}).$$

Let $g(m_t) \in \mathbb{R}^3$ denote the decoded global head/root position, and let $\ell(m_t)$ denote the local end-effector coordinates. Both are decoded after undoing any motion normalization. Let b_t indicate a valid, unpadded frame and r^M indicate that motion is present, and define $\Omega_M = \{t \in \mathcal{Q} \mid b_t = 1,\ r^M = 1\}$.

The global trajectory auxiliary loss is

$$\mathcal{L}_{\text{traj}} = \frac{1}{|\Omega_M|} \sum_{t \in \Omega_M} \left\| g(\hat{m}_{0,t}) - g(m_t) \right\|_2 .$$

The local velocity auxiliary loss constrains frame-to-frame local end-effector motion. Define

$$\Omega_M^\Delta = \{t \mid t \in \mathcal{Q},\ b_t = 1,\ b_{t-1} = 1,\ r^M = 1\}.$$

Then

$$\mathcal{L}_{\text{local-vel}} = \frac{1}{|\Omega_M^\Delta| d_\ell} \sum_{t \in \Omega_M^\Delta} \left\| \Delta\ell(\hat{m}_{0,t}) - \Delta\ell(m_t) \right\|_2^2 ,$$

where $\Delta\ell(m_t) = \ell(m_t) - \ell(m_{t-1})$ and $d_\ell = 12$ for the two hands and two feet in the head/root frame. An auxiliary term is zero when its valid index set is empty. We use $(\lambda_{\text{traj}}, \lambda_{\text{local-vel}}) = (0.1, 1.0)$ in pre-training.

B.4 POST-TRAINING DETAILS

Sampling from the Frozen Prior Model. During post-training, the observed visual history is clamped and explicit motion prediction is disabled. The frozen prior model initializes future visual latents from Gaussian noise and integrates its flow ODE using $N_{\text{train}} = 3$ sampling steps to generate a visual future. The expert's conditioning states are extracted along this sampled trajectory, rather than obtained by adding noise to ground-truth future latents. At each sampled world denoising timestep, we retain hidden states from the selected prior model layers. For an action flow timestep τ, $H^{\theta}(\tau)$ uses the nearest sampled world denoising timestep. This procedure trains the expert on the prior model's own predictions, matching the source of predictive context used at inference. The three-step setting applies to sampling from the prior model during post-training; inference uses the ten-step schedule in Appendix B.5. Gaussian corruption of demonstrated action chunks is still used for the expert's flow-matching objective.

Predictive Context Selection. Hidden states from all 24 generated future frames, at indices $\{K+1, \ldots, T\}$ with $K = 1$ and $T = 25$, are exposed to the action expert as world cross-attention keys and values. The observed input frame conditions the prior model internally but is excluded from the expert's cache. The expert receives its current robot observation separately through O_q. The same future-token selection is used during post-training and inference.

World-Context Dropout. For real-world experiments, we apply sample-level world-context dropout with probability 0.1 during post-training, with the same dropout decision shared across all expert layers. This encourages the action expert to use the current observation O_q and limits over-reliance on the prior model's predictive context. World-context dropout is not used in simulation experiments.

Action Space. For real-world whole-body control, we use SONIC (Luo et al., 2025). Each action is a 66D vector: a 64D SONIC universal control token followed by two scalar open/close commands, ordered as left hand and right hand. Recorded hand commands are binary values in $[0, 1]$. Proprioceptive inputs and action targets are standardized per channel using the robot training-split statistics; predicted commands are transformed back to controller units before execution.

Robot Observation. The robot observation O_q consists of the current egocentric image and a 63D proprioceptive vector containing the recorded roll and pitch, angular velocity, joint positions, and joint velocities. The image is encoded into patch tokens using the same frozen DINO encoder used for the visual stream. The post-training configuration disables the optional normalized execution-progress token (`policy_use_progress_token=false`); the runtime command counter tracks progress within the prediction window but is not supplied as a separate policy token.

Prior Model Layer Selection. The prior model and action expert each have 30 layers, with corresponding layers paired by depth. World cross-attention is applied at every third expert layer, at one-based indices $3, 6, 9, \ldots, 30$, using hidden states before the matching prior layer. The other expert layers process local observation and action tokens without injecting world context.

Temporal Configuration. Each prior window contains one observed frame ($K = 1$) and 24 future frames ($T = 25$). At a prediction frame rate of 5 Hz, the future spans $24/5 = 4.8$ seconds. All 24 future-token hidden states are retained as predictive context. At every action query, the expert receives the latest single image and proprioceptive state from the robot observation stream.

The three temporal rates have distinct roles. The prior's 5 Hz frame rate specifies the spacing of predicted visual states (0.2 seconds). The low-level controller executes commands at 50 Hz (0.02 seconds per command). Each expert query predicts an action chunk of length $H_a = 30$, covering $30/50 = 0.6$ seconds, and a new query is made after that chunk is executed. The nominal query rate is therefore $1/0.6 \approx 1.67$ Hz. The prior context is refreshed before the prediction window is exhausted.

B.5 Inference Details

Both the prior model and the action expert use $N = 10$ Flow Matching steps with denoising timesteps

$$\{\tau_n\}_{n=1}^{N}.$$

During denoising by the prior model, we extract the intermediate representations

$$\left\{H_l^\theta(\tau_n)\right\}_l$$

from the selected prior model layers, retaining all 24 generated future positions specified in Appendix B.4. These representations are converted into policy-side key-value caches; observed-history input tokens are not cached for the expert. Following the notation in the main text, we collectively denote the cached prior model context associated with denoising step n as

$$H_n^\theta.$$

Because the prior model and the action expert use the same denoising schedule, H_n^θ is reused at the corresponding action denoising timestep τ_n.

The action expert initializes the action chunk from Gaussian noise,

$$\mathbf{a}^0 \sim \mathcal{N}(0, I),$$

and performs N denoising steps. At step n, it predicts

$$\hat{u}^n = u_\phi^a\left(\mathbf{a}^{n-1}, \tau_n, H_n^\theta, O_{\text{cur}}\right),$$

where O_{cur} denotes the current robot observation. The action sample is updated as

$$\mathbf{a}^n = \mathbf{a}^{n-1} - \hat{u}^n \Delta\tau_n, \qquad n = 1, \ldots, N.$$

The final output is the executable action chunk

$$\mathbf{a}^N = (a_t, \ldots, a_{t+H_a-1}), \qquad H_a = 30.$$

At the start of a window, we acquire the conditioning image, generate the 24-frame future, and cache its hidden states. During execution, a command counter records the position within the current prediction window. Each action query uses the latest O_{cur} and the same cached future to produce 30 commands, which the low-level controller executes at 50 Hz. The next query occurs after 0.6 seconds of command execution. Before the current prediction window is exhausted, a new image is acquired, the prior regenerates the future, the cache is replaced, and the counter is reset. Between cache refreshes, every action query remains conditioned on the latest robot observation. These rates describe the nominal execution schedule; denoising compute time is distinct from the duration represented by a command chunk.

C SIMPLE Benchmark and Simulation Task Details

C.1 Benchmark and Evaluation Protocol

SIMPLE (SIMulation-based Policy Learning and Evaluation) is a simulation framework for humanoid loco-manipulation (Wei et al., 2026b). It combines MuJoCo physics with Isaac Sim rendering and supports demonstration collection through motion planning and VR teleoperation. We evaluate six Unitree G1 tasks spanning tabletop grasping, bimanual transfer, and mobile manipulation. The `MP` and `Teleop` suffixes in the environment names identify the demonstration-generation pipelines.

We follow the officially released six-task evaluation protocol, retaining the official task implementations, initialization and domain-randomization configurations, and task-specific success evaluators. We use the default random-seed settings provided by the official implementation without overriding them. Each method is evaluated on ten trials for every task and randomization level, giving $6 \times 3 \times 10 = 180$ trials per method. The easy, medium, and hard settings correspond to Levels 0, 1, and 2, respectively. Table 6 summarizes the cumulative perturbations. Higher levels add variation; individual task success counts need not decrease monotonically across levels.

All simulation baseline results in Table 1 are obtained from our own training and evaluation runs using the respective official implementations. All methods are evaluated on the same machine. The ablation variants in Table 2 follow the same task configurations, default seed settings, success criteria, and trial counts as the main comparison. We report successful trials for each task and level, and compute the aggregate success rate over all 180 trials.

Table 6: **SIMPLE evaluation levels.** Each level retains the variations of the preceding level. The definitions follow the official evaluation protocol.

Level	Table notation	Randomization
0	Easy	Table materials; distractor identities and initial positions.
1	Medium	Level 0 variations, plus substantial lighting changes.
2	Hard	Level 1 variations, plus target-object initial-position perturbations.

C.2 The Six Evaluated Tasks

Table 7 maps the task names in the main results to their simulation environments and summarizes their objectives (Wei et al., 2026b). Together, the tasks vary the required reaching height, use of two hands, and integration of locomotion with object interaction.

Table 7: **Task definitions and environment mapping.** Append each suffix below to `G1Wholebody` and prepend `simple/` to obtain the full environment identifier.

Task / environment suffix	Objective and coordination requirement
BendPick `BendPickMP-v0`	Bend toward a low table and grasp the target. The benchmark requires lifting it more than 5 cm above the table, coupling body posture with reaching.
Handover `HandoverTeleop-v0`	Transfer an object between the two hands and place it in a container, requiring coordinated grasping, handover, and release.
Mobile P&P `LocomotionPick BetweenTablesTeleop-v0`	Pick up an object and transport it to a container on another table, maintaining the grasp during locomotion and completing placement.
Grasp `TabletopGraspMP-v0`	Grasp a tabletop target and lift it more than 5 cm above the table. This setting emphasizes local hand–object alignment and lifting.
XMoveBendPick `XMoveBendPickTeleop-v0`	Move laterally, bend toward a low target, and lift it, combining body repositioning with a change in reaching height.
XMovePick `XMovePickTeleop-v0`	Move laterally and lift a target object, testing the transition from locomotion to grasping.

C.3 Task Demonstration Snapshots

Figure 7 illustrates the six tasks using one demonstration sequence per task from the SIMPLE dataset (Wei et al., 2026b). Each row contains four selected egocentric observations showing the progression of the task, including reaching, object interaction, and, where applicable, transport or handover. Labels give the original frame indices within each sequence.

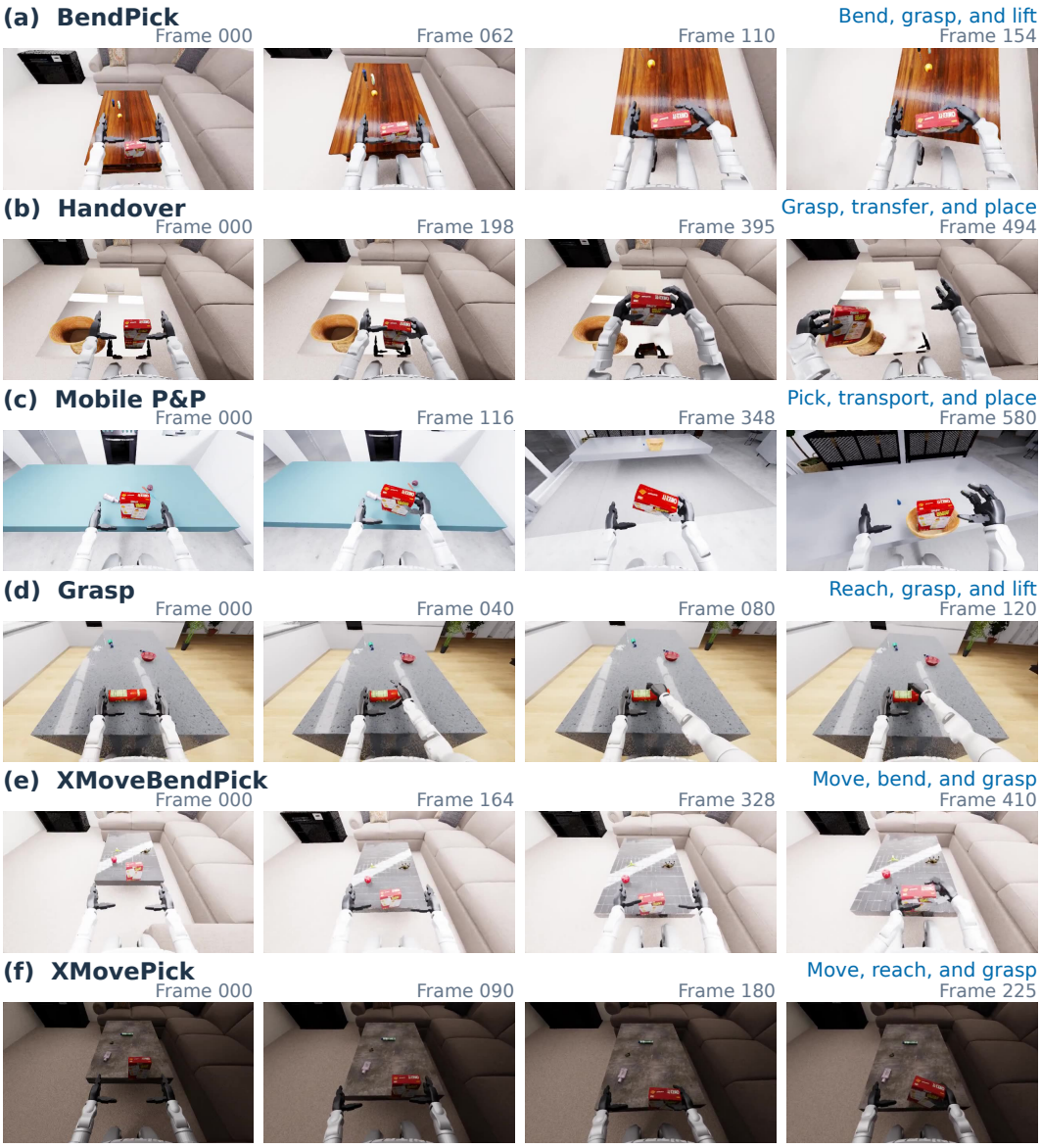

Figure 7: **Egocentric demonstrations of the six SIMPLE tasks.** Each row shows one dataset trajectory, with time advancing from left to right. BendPick and Grasp illustrate reaching and lifting; Handover shows bimanual transfer and placement; Mobile P&P shows object transport between tables; and XMoveBendPick and XMovePick combine body repositioning with grasping. Frames retain their original field of view and colors. Frame indices refer to the source image sequences; the selected intervals may differ within and across tasks.

Discussion of Task Interference. Compared with language-only context, the full model improves substantially on XMovePick and BendPick but underperforms on XMoveBendPick. These tasks share similar visual observations (Figure 7), which may introduce task interference under joint multi-task training. Increasing model capacity may help alleviate this interference.

C.4 Visual Variation Across Evaluation Levels

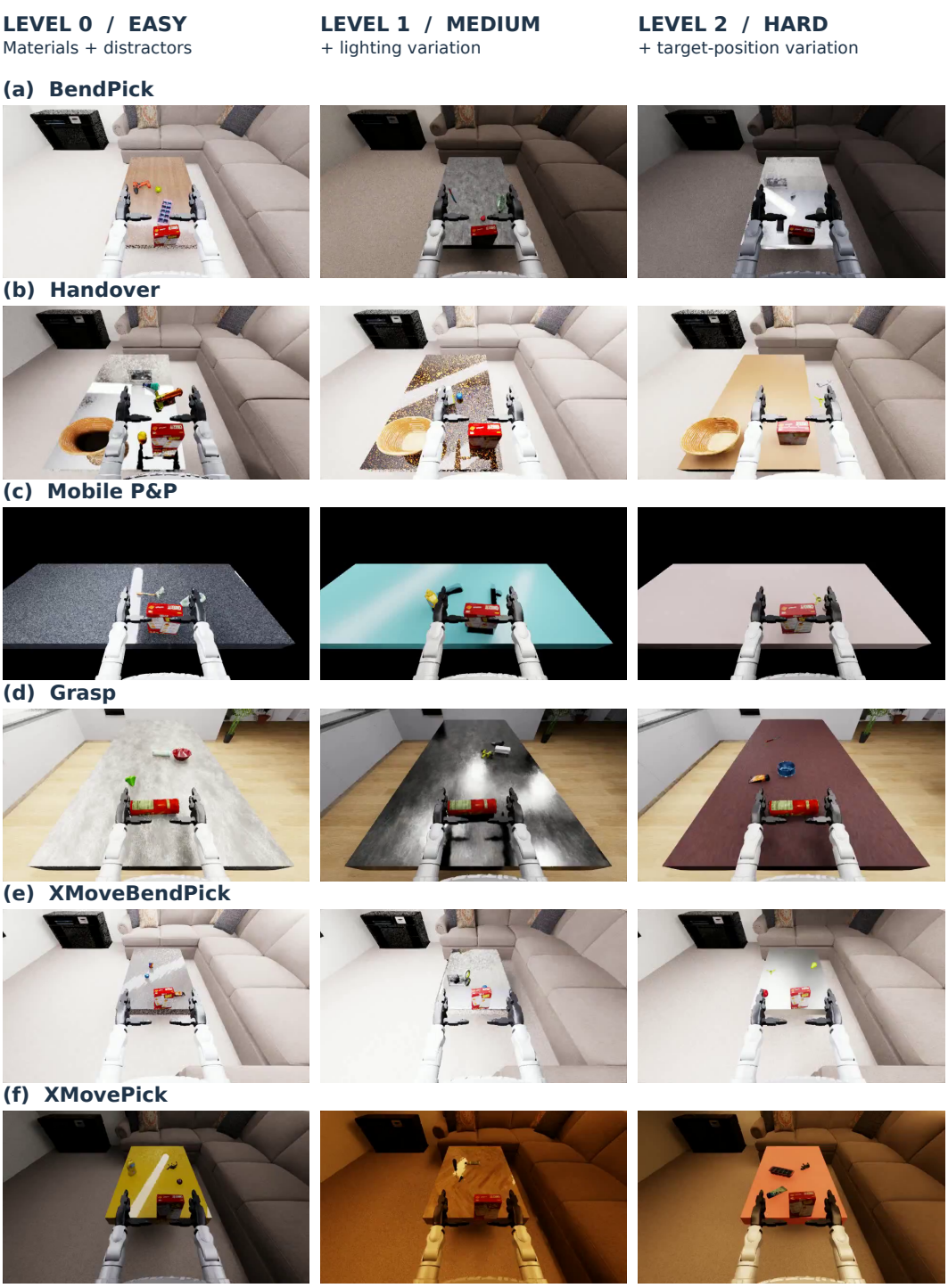

Figure 8: **Six tasks under three cumulative randomization levels.** Each cell is the first frame of episode 0 in the corresponding GR00T-N1.6 evaluation folder, viewed through the left head camera. Columns follow the level definitions in Table 6. These are separate evaluation scenes; the shared episode index does not establish matched random seeds across levels. Images retain the recorded lighting, materials, and object arrangements without color adjustment or cropping.

D Real-World Experimental Details

D.1 Task Descriptions

We quantitatively evaluate three real-world loco-manipulation tasks: Walk to Fish (W2F), Walk to Mirror (W2M), and Walk to Doll (W2D). All three tasks include approaching the target and performing the subsequent manipulation. We evaluate Being-M0.7, GR00T-N1.6, Ψ_0, and $\pi_{0.5}$ over five trials per task, totaling fifteen trials per method, as reported in Table 3. Physical markers fix the robot starting positions and the table positions across methods. Target objects are randomly initialized at locations spanning the left, center, and right regions of the task workspace. Each trial allows at most 30 seconds for the complete approach-and-manipulation sequence. A human evaluator judges success: the robot must successfully grasp the target doll in W2M and W2D, or scoop up the toy fish with the net in W2F. Trials that do not meet the corresponding criterion within 30 seconds are counted as failures.

Walk to Fish (W2F). The robot walks toward a water tank, positions its body for interaction, and uses a handheld net to scoop up a toy fish. The quantitative evaluation covers this complete approach-and-scooping sequence. The task requires coordinated locomotion, body positioning, and tool use guided by visual feedback. The Water-Tank Fish Scooping case study in Figure 9 illustrates this sequence.



Figure 9: Snapshots of the complete Walk to Fish task sequence.

Walk to Mirror (W2M). The robot walks toward a table and approaches a partially enclosed box containing a toy. Since the toy is initially hidden from direct view, the robot must infer its location from a mirror reflection before reaching into the box to grasp it. The task combines indirect visual observations with whole-body positioning and manipulation. The Mirror Toy Grasping case study in Figure 10 illustrates this behavior.

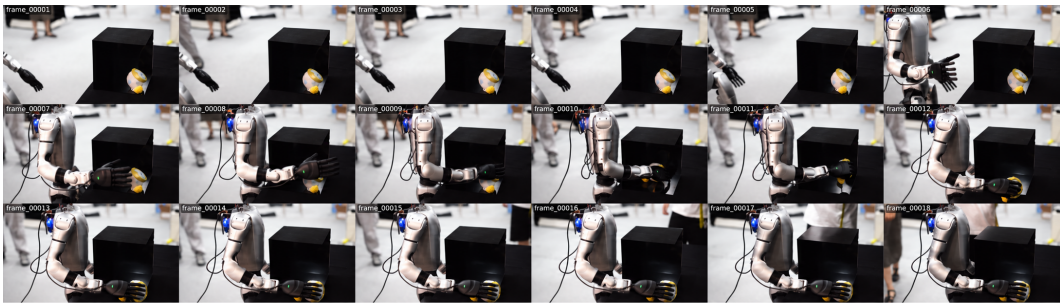

Figure 10: Snapshots of the complete Walk to Mirror task sequence.

Walk to Doll (W2D). The robot walks toward a table and grasps a directly visible doll. This task evaluates the transition from locomotion to target-directed grasping, requiring the robot to coordinate its approach, body positioning, and arm motion. Together, W2M and W2D assess approach-and-grasp sequences with mirror-based and direct target observations, respectively.

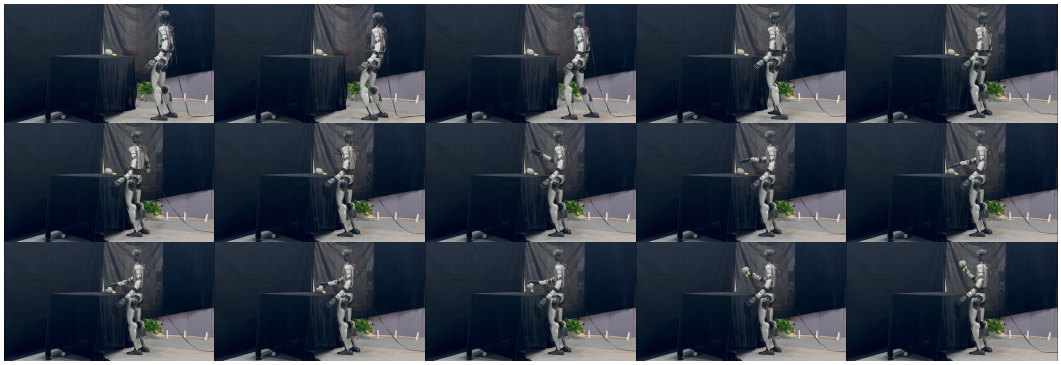

Figure 11: Snapshots of the complete Walk to Doll task sequence.