Being-M0.7: A Latent World-Action Model for Humanoid Robots

BeingBeyond Team

https://research.beingbeyond.com/being-m07

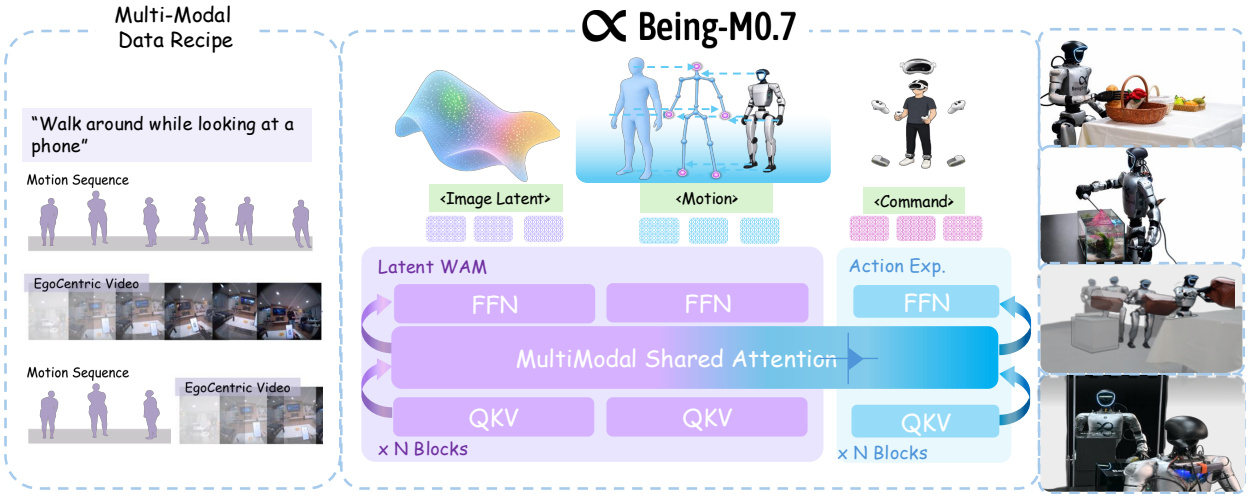

Figure 1: **Being-M0.7 at a glance.** Being-M0.7 is a latent World-Action Model pretrained on egocentric videos and human motions with a Mixture of Transformers (MoT) structure, then grounded into humanoid control through action expert post-training on robot trajectories for diverse loco-manipulation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

Abstract

Humanoid loco-manipulation requires a humanoid robot to coordinate locomotion, posture, and dexterous manipulation while reasoning about future scene evolution and whole-body task progress. However, scalable humanoid learning remains a big challenge because robot demonstrations are costly to collect and pixel-level future prediction is computationally expensive. To this end, we present **Being-M0.7**, a latent world-action model that learns a video-motion prior from a mixed-modality human-centric corpus and adapts it to humanoid robot control. Instead of rolling out future pixels, Being-M0.7 predicts future states in a latent space and couples them with future motion in a compact whole-body representation shared by humans and humanoid robots. During robot post-training, a lightweight action expert grounds the pretrained video-motion prior to executable whole-body commands using limited humanoid demonstrations. Experiments on real-world Unitree G1 tasks show that Being-M0.7 can coordinate state prediction, motion generation, humanoid locomotion, and dexterous manipulation across challenging interaction scenarios.

Date: July 14, 2026

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 [1–6]. Unlike fixed-base manipulation, the robot must jointly decide where to move, how to orient its body, how to 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 not only about the current visual observation, but also about future scene evolution, body motion, and task progress.

Recent robot foundation models and world-action models have made encouraging progress [7–17], but scaling them to humanoid loco-manipulation remains challenging. First, humanoid demonstrations are much harder to scale than fixed-view arm manipulation data. They require synchronized egocentric video, proprioception, and executable whole-body commands, and are constrained by teleoperation difficulty, safety, hardware availability, and environment diversity [1–3]. Second, pixel-level future prediction is expensive and can allocate substantial capacity to appearance details that are only weakly related to control [17, 18]. For humanoid tasks, the predicted visual trajectory may also be noisy due to fast ego-movement, viewpoint changes, and body-induced camera jitter. Third, many action modeling pipelines still emphasize upper-body manipulation or locomotion separately, which can lead to weak coordination between the upper and lower body and limits their suitability for natural whole-body control [1, 2, 5, 19, 20].

Our approach addresses these three challenges in turn. First, to reduce dependence on expensive robot demonstrations, we pre-train on scalable human-centric data: egocentric videos [21, 22], paired egocentric video–motion data [23, 24], and motion sequences [25–29]. Using a unified motion representation aligning with humanoid embodiment, these sources are able to provide diverse interaction contexts, vision-grounded kinematics, and broad whole-body motion coverage for downstream usage. Second, to avoid costly and noisy pixel rollouts, the pretrained model jointly predicts future visual–motion features, focusing its capacity on control-relevant scene and motion dynamics [13, 14, 30]. Third, to couple locomotion with manipulation, we introduce a post-training formulation for whole-body control. A future-conditioned action expert adapts the pretrained prior world-action model to executable robot commands using limited humanoid demonstrations, enabling coordinated whole-body behavior rather than isolated upper-body actions.

Our contributions are summarized as follows:

- We build a large-scale, more than 10,000-hour mixed-modality pre-training corpus for humanoid loco-manipulation, covering human egocentric videos, egocentric video-motion pairs, and human motion sequences. This substantially expands the supervision available to humanoid learning beyond costly teleoperation.
- We present Being-M0.7, which, to our knowledge, is the first latent world-action model for humanoid control and is pre-trained from large-scale vision and motion data jointly. Before being grounded into robot commands, the model learns visual context, future dynamics, and humanoid kinematic structure from scalable human-centric data.
- We introduce a unified motion representation shared by human motion and humanoid robot trajectories, which bridges the morphology gap between humans and humanoid robots, provides richer post-training supervision than executable commands alone, and exposes additional motion-level feedback paths during inference for coordinated whole-body control.
- We demonstrate that Being-M0.7 achieves superior quantitative performance against strong baselines on real-world Unitree G1 tasks, while qualitatively exhibiting coordinated whole-body control across state prediction, motion generation, locomotion, and manipulation.

2 Related Work

Humanoid Loco-manipulation. Humanoid loco-manipulation aims to enable robots to coordinate whole-body locomotion and manipulation in human-centered environments, where interaction often requires simultaneous navigation, balance maintenance and contact reasoning. Compared with fixed-base manipulation

or pure locomotion, humanoid loco-manipulation is substantially more challenging because the policy must reason over high-dimensional whole-body actions, dynamic contacts, object geometry, and long-horizon task progress.

Existing learning-based methods for humanoid loco-manipulation can be broadly grouped into reference-based, reference-free, and emerging hybrid paradigms [3, 31–36]. Reference-based approaches exploit motion priors from retargeted human motions, demonstrations, videos, or optimized trajectories to stabilize whole-body control and produce natural behaviors [37–40]. Representative methods such as ResMimic [39] and HDMI [40] adapt or extract reference motions for interactive policy learning, while recent systems further leverage motion generation, privileged teachers, or large-scale simulation for contact-rich whole-body behaviors [41, 42]. Despite their stability and motion naturalness, these methods often depend on predefined trajectories, making it difficult to adapt to unseen object geometries, changing contacts, or object-specific interaction strategies at scale.

Reference-free methods instead learn directly from task objectives, proprioception, and perception without replaying fixed reference motions [31, 43, 44]. This formulation is more suitable for generalizable loco-manipulation, but remains challenging due to high-dimensional humanoid actions, sparse rewards, unstable contacts, and long-horizon interactions. To mitigate these difficulties, recent works introduce structured representations or auxiliary priors: VisualMimic [45] uses egocentric perception and motion-statistics-constrained hierarchical control, while LessMimic [31] employs distance-field-based interaction representations to improve geometric generalization and skill composition. Foundation model approaches further explore egocentric demonstrations, latent actions, and large-scale pretraining to expand task coverage [4–6]. Overall, the field is moving from motion-specific controllers toward scalable visuomotor policies that integrate motion priors, perception, and object-level generalization.

Motion Generation. Large-scale datasets such as AMASS [25], HumanML3D [26], and MotionX [27] have enabled text-conditioned models to learn diverse human motions with rich semantic annotations. Building on these resources, T2M-GPT [46] autoregressively generates discrete motion tokens, while diffusion-based methods such as MDM [47] and MotionDiffuse [48] improve diversity and realism through conditional denoising. More general motion models, including MotionGPT [49], MotionGPT-2 [50], and the Large Motion Model [51], further unify multiple motion-language tasks and conditioning modalities. Beyond kinematic plausibility, PhysDiff [52], ReinDiffuse [53], and RLPF [54] incorporate physics-based projection or reinforcement learning to improve physical execution, while RobotMDM [55] adapts generated motions to robotic characters through controller feedback.

Egocentric motion generation complements language-conditioned modeling by grounding body motion in the observations available to an embodied agent. EgoEgo [56] decomposes egocentric body recovery into head-motion estimation and conditional full-body generation, whereas EgoAllo [57] uses SLAM-derived head poses and images to recover body pose, height, and hand motion in a common scene frame. Moving beyond reconstruction, UniEgoMotion [58] unifies egocentric motion reconstruction, forecasting, and generation with a head-centric conditional diffusion model. EgoAgent [59] extends motion prediction to an agent model that jointly represents the current world and predicts future states and actions, while EgoTwin [60] jointly generates egocentric video and human motion to couple viewpoint-consistent visual dynamics with head-centered body dynamics. This progression shows that first-person visual context can turn human motion generation from an isolated kinematic task into observation-grounded predictive modeling of embodied behavior.

Human motion provides scalable supervision for semantic behavior and whole-body coordination, whereas a humanoid policy must additionally account for embodiment differences, physical feasibility, real-time feedback, and executable robot commands. Existing motion-generation methods narrow parts of this gap through physical refinement, robot adaptation, or egocentric prediction, but generally address these components separately rather than learning from human-centric video and motion before adapting the acquired knowledge to humanoid execution. Being-M0.7 follows this trajectory as a prior latent world-action model: it jointly pre-trains latent visual-state prediction and motion generation on scalable human-centric data, uses a unified human–humanoid motion representation to bridge morphology, and delegates embodiment-specific command generation to a lightweight action expert.

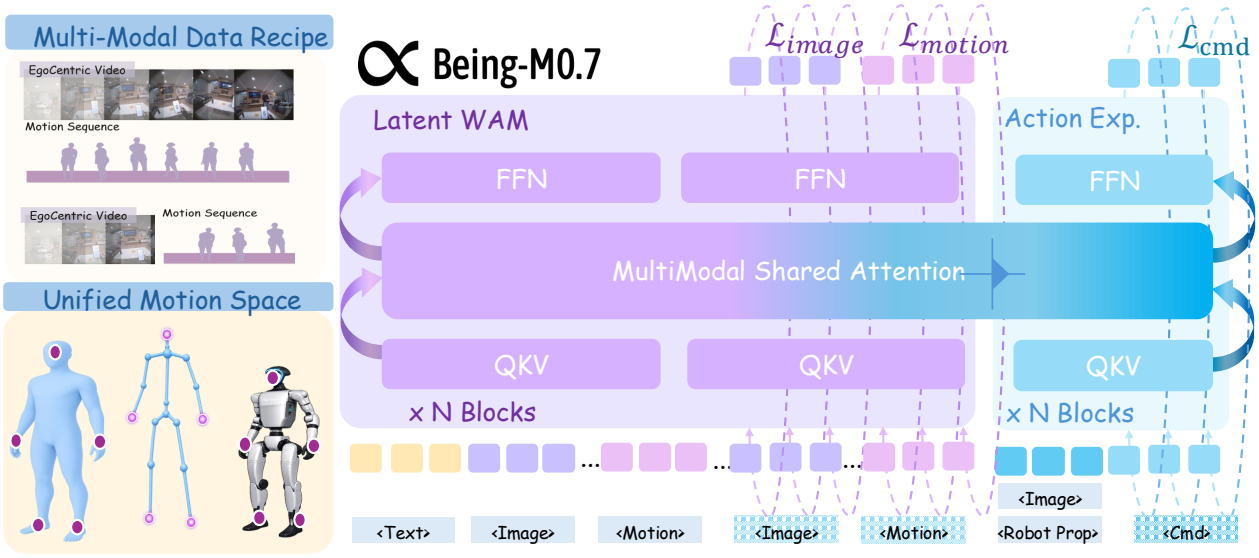

Figure 2: **Overview of Being-M0.7 Training Framework.** Top-left: We construct the pretraining corpus from paired video–motion data, video-only data, and motion-only data. Bottom-left: We develop a unified motion representation shared by humans and humanoid robots, providing richer supervision and feedback for training and inference. Right: The Being-M0.7 model architecture.

Humanoid Foundation Models. In parallel, humanoid robot learning has increasingly moved toward foundation models for whole-body control. Compared with fixed-skill policies, these models seek to absorb heterogeneous supervision from robot trajectories, human videos, language instructions, and simulation, while retaining an action interface that can drive high-DoF humanoid bodies. GR00T N1 [61]represents a generalist humanoid VLA model that couples a vision-language module with a diffusion-transformer action module and trains on heterogeneous robot, human-video, and synthetic data. WholeBodyVLA [5] also follows the VLA direction, but places more emphasis on learning latent actions from low-cost action-free egocentric videos and connecting them to a locomotion-oriented controller for large-space loco-manipulation [5]. Ψ_0 similarly treats egocentric human video as a scalable source of pre-training data, and adopts a staged humanoid VLA recipe that learns transferable visual-action representations before post-training a flow-based action expert on high-quality humanoid robot trajectories [4]. From an embodiment perspective, HEX[62] studies how whole-body manipulation knowledge can transfer across embodiments through humanoid-aligned state representations and expert modules, highlighting that humanoid foundation models must preserve coordination across high-DoF bodies rather than only scaling perception and language modules.

World-action models provide a complementary formulation in which learned video generation models supply structured temporal information for humanoid control. MotionWAM[20] frames real-time humanoid loco-manipulation as a world-action modeling problem: an egocentric video generation model provides intermediate denoising features that condition a motion model, allowing a humanoid to act from a single egocentric camera. This perspective is close to our motivation because it treats human-centric visual experience not merely as demonstrations, but as a source of predictive structure about future scene and motion evolution. However, our objective is different from directly constructing a closed-loop humanoid policy. We instead pre-train a prior latent world-action model, so that the downstream whole-body control policy can make use of temporally organized motion information learned from human-centric data.

3 Method

In this section, we introduce **Being-M0.7**, a latent world-action model that leverages egocentric video and motion data to learn a prior world-action model, which is then transferred to real-robot humanoid control via a future-conditioned action expert. We first formulate humanoid loco-manipulation as the combination of a prior world-action model and a future-conditioned action expert in Section 3.1. We then describe the

construction of the mixed-modality pre-training dataset and the unified motion representation shared by humans and humanoid robots in Section 3.2. Next, we present the model architecture, including the video-motion Mixture-of-Transformers (MoT) and the future-conditioned action expert, in Section 3.3. Finally, we detail the pre-training, robot post-training, and closed-loop inference recipes. An overview of the Being-M0.7 training framework is shown in Figure 2.

3.1 Problem Formulation

We first consider a prior world-action model for humanoid loco-manipulation from egocentric observations. Let V denote an egocentric video, $Z = \mathcal{E}(V)$ its latent visual representation, M a compact humanoid motion sequence, and I the task instruction. We treat M as a coarse motion-level action plan in the motion space: it describes how a humanoid body should move in the future, but does not specify embodiment-specific executable commands. The prior world-action model therefore predicts future environmental states and coarse future actions conditioned on the short history and the task instruction.

For a temporal window, we split the sequence into a context interval $\mathcal{C} = \{1, \ldots, K\}$ and a future interval $\mathcal{Q} = \{K+1, \ldots, T\}$. The prior world-action model is written as

$$P\left(Z_{\mathcal{Q}}, M_{\mathcal{Q}} \mid Z_{\mathcal{C}}, M_{\mathcal{C}}, I\right).$$

This prior world-action model captures both how the environment evolves and how the humanoid body may move to accomplish the task.

Humanoid loco-manipulation further requires grounding the coarse motion-level plan into embodiment-specific commands. Let O_q denote the robot observation at an action query time q, and let $\mathbf{a}_q$ be the robot action to be generated. We introduce a future-conditioned [63] action expert that refines the predicted future context into executable commands:

$$P\left(\mathbf{a}_q \mid Z_{\mathcal{Q}}, M_{\mathcal{Q}}, O_q, I\right).$$

The complete latent world-action model is obtained by combining the prior world-action model with the action expert:

$$P\left(Z_{\mathcal{Q}}, M_{\mathcal{Q}}, \mathbf{a}_q \mid Z_{\mathcal{C}}, M_{\mathcal{C}}, I, O_q\right) \approx P\left(\mathbf{a}_q \mid Z_{\mathcal{Q}}, M_{\mathcal{Q}}, O_q, I\right) P\left(Z_{\mathcal{Q}}, M_{\mathcal{Q}} \mid Z_{\mathcal{C}}, M_{\mathcal{C}}, I\right). \tag{1}$$

This decomposition separates lower-frequency world modeling from higher-frequency action generation, allowing the future representation to be reused across multiple low-latency control updates.

3.2 Data Curation

Dataset Composition We construct the Being-M0.7 pre-training dataset for humanoid loco-manipulation from a raw collection of more than 10,000 hours of mixed-modality data before filtering and temporal segmentation, the composition and source breakdown are summarized in Figure 3. Instead of relying solely on paired video-motion data, **our pre-training dataset consists of three complementary streams: egocentric video-motion pairs, egocentric videos, and human motion sequences, which expands the available data from thousands to over ten thousand hours.** These streams correspond to paired video-motion, video-only, and motion-only supervision, respectively. We denote the filtered pre-training dataset as

$$\mathcal{D} = \mathcal{D}_{VM} \cup \mathcal{D}_{V} \cup \mathcal{D}_{M},$$

where $\mathcal{D}_{VM}$ supervises the joint distribution of vision and motion, while $\mathcal{D}_{V}$ and $\mathcal{D}_{M}$ constrain the marginal distributions of vision and motion, respectively.

Latent Visual Representation For the visual modality, all frames are encoded by a DINO [64] encoder: $Z = \mathcal{E}(V)$. These visual latents are used both as the inputs and as the generation targets. Consequently, the model performs visual-latent-level generation rather than pixel-level video generation. This reduces the dimensionality of the generation space and biases the visual objective toward semantic state, object layout, and scene evolution instead of low-level pixel reconstruction, which is the reason we refer to our model as a latent world-action model.

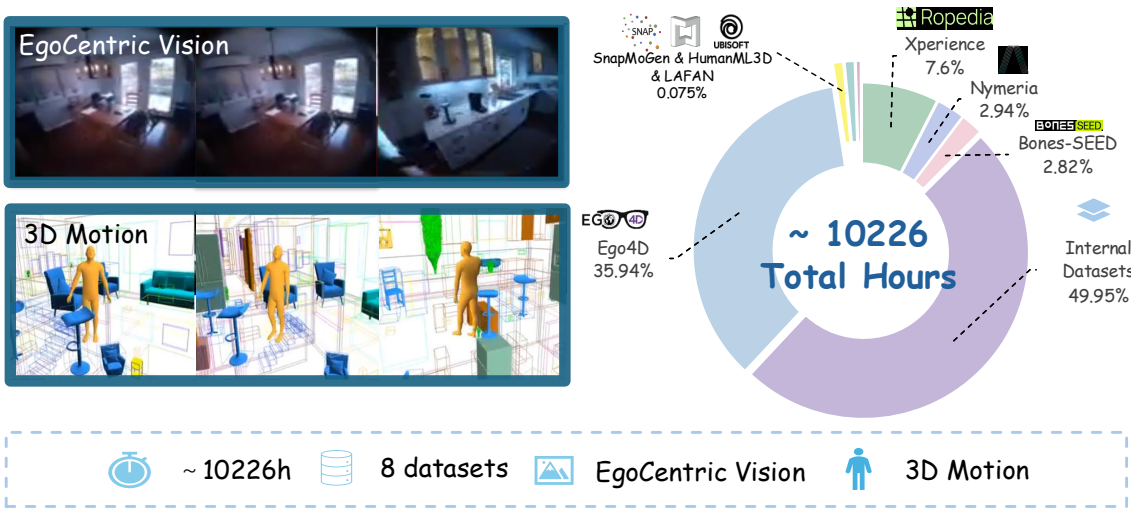

Figure 3: **Being-M0.7 data recipe.** The pre-training corpus is sourced from external public datasets including Ego4D [21], Xperience [65], Nymeria [24], Bones-SEED [19], SnapMoGen [28], HumanML3D [26], and Lafan1 [66], as well as internal datasets comprising part of the Being-H0.5 [16] training data, part of the Being-M0.5 [29, 67] training data, 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 their marginals, respectively.

Unified Motion Representation For motion modality, we convert heterogeneous human motion sources into a unified head-root representation. Unlike conventional body-centric representations that use the pelvis as the root, we use the head as the root joint, which better aligns the motion representation with the egocentric visual stream. During conversion, each sequence is canonicalized, for example by removing the initial heading and aligning the body with a floor-centered coordinate frame. This standardization reduces dataset-specific nuisance factors and makes motion statistics more consistent across sources.

We further introduce a unified compact motion representation that keeps only the head, two hands, and two feet. This representation does not aim to preserve full-body kinematic detail; instead, it retains the key interaction and contact cues needed for downstream robot control. Because the same representation can be constructed from both human motion and humanoid robot trajectories, it bridges the morphology gap between humans and robots. **In robot post-training, it provides richer supervision than executable command labels alone, and during inference it exposes additional motion-level feedback about robot kinematics, supporting coordinated whole-body control.**

3.3 Model Architecture

Being-M0.7 instantiates the factorization in Equation (1) with a prior world-action model and a lightweight action expert. The prior world-action model captures $P(Z_{\mathcal{Q}}, M_{\mathcal{Q}} \mid Z_{\mathcal{C}}, M_{\mathcal{C}}, I)$, while the action expert models $P(\mathbf{a}_q \mid Z_{\mathcal{Q}}, M_{\mathcal{Q}}, O_q, I)$. The two components interact through a one-way connection: the action expert receives intermediate representations from the pretrained prior world-action model, but no action information is passed back to the prior model. In this way, action prediction builds on the prior model's learned representations without altering how it processes its inputs.

Crucially, the modality-specific branches of MoT allow the same architecture to learn from paired video–motion data as well as video-only and motion-only data. **This mixed-modality design expands the avaliable pre-training supervision from the thousand-hour scale of paired data to a corpus of more than 10,000 hours**, while retaining explicit cross-modal interaction whenever both modalities are present. The motion stream serves as a coarse-grained action supervision: it captures embodiment-agnostic whole-body intent and kinematic structure, whereas the action expert grounds the prior world-action model

into embodiment-specific robot commands.

Multimodal Input Embedding. Using the visual latents z_t and compact motion vectors m_t defined in Section 3.2, we apply separate learned input projections to map the two modalities into the common hidden dimension of the MoT backbone. 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 [68] and supplied as language-conditioning features.

Video-Motion Mixture of Transformers. The prior world-action model is implemented as a Mixture-of-Transformers (MoT) [69] 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 [70].

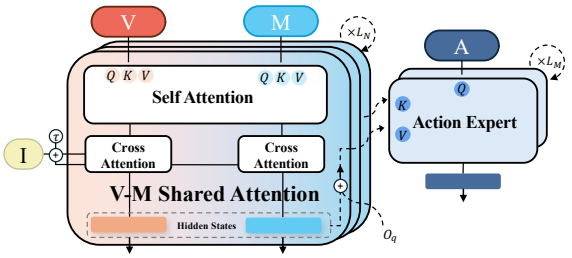

Figure 4: The details of the shared multimodal attention: *V, M, I,* and *A* denote the hidden states of vision, motion, instruction, and action, respectively. τ is the flow matching timestep. The hidden states are injected into the action expert in the form of cross attention along with robotic observation O_q.

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.

Future-Conditioned Action Expert. During post-training, we append a separate flow-based transformer that realizes the action expert in Section 3.1. When the prior world-action model denoises $(Z_{\mathcal{Q}}^{\tau}, M_{\mathcal{Q}}^{\tau})$ conditioned on $(Z_{\mathcal{C}}, M_{\mathcal{C}}, I)$, we denote its hidden video-motion sequence before layer l by $H_l^{\theta}(\tau)$. The action expert embeds the noisy action chunk $\tilde{\mathbf{a}}_{q,\tau}$ into an initial token sequence A_q^0 and separately embeds the robot observation O_q. Here, O_q consists of the current egocentric image, proprioceptive state and normalized execution progress. The image is encoded by the same frozen DINO encoder as the visual stream, while the proprioceptive state is mapped through a learned projection. At expert layer j, action tokens attend to the concatenation of (i) the hidden video–motion sequence from a corresponding prior layer, (ii) the current action token states, and (iii) the current robot-observation tokens:

$$A_q^{j+1} = \mathcal{B}_j\left(A_q^j \,\middle|\, \left[H_{\ell_j}^{\theta}(\tau), A_q^j, O_q\right], \tau\right).$$

Here $\mathcal{B}_j$ denotes the j-th expert block, ℓ_j associates expert layer j with a selected layer of the prior world-action model, and information flows only from the prior world-action model to the expert. Conditioning on representations from multiple network depths provides the policy with both fine-grained visual and motion features and high-level future-dynamics representations, while O_q retains closed-loop responsiveness at the robot control rate. Detailed architecture hyperparameters are provided in Appendix A.1.

3.4 Training and Inference Recipes

Pre-Training. We pre-train the prior latent world-action model on a mixture of paired video-motion, video-only, and motion-only data within a unified objective. Paired samples supervise the conditional joint generation of future visual latents and motion trajectories, while video-only and motion-only samples provide marginal supervision for their respective modalities. This design allows heterogeneous data streams to jointly constrain a single video-motion model without introducing separate unimodal stages.

In particular, motion-only training teaches whole-body kinematic structure and dynamics even in the absence of visual observations, while paired training aligns this coarse-grained action representation with scene evolution. This recipe allows robot-free data to contribute both **visual context** and **kinematic structure** before robot command grounding.

The model is trained via a flow matching [71] objective. For each observable future token x_0, where x_0 may be either a visual latent token or a motion token, we sample Gaussian noise ϵ and a flow matching timestep $\tau \sim \mathcal{U}(0,1)$, and supervise the velocity field along the linear probability path:

$$x_\tau = (1-\tau)x_0 + \tau\epsilon, \qquad u^\star = \epsilon - x_0.$$

The model predicts a task-conditioned velocity $u_\theta(x_\tau, \tau, I, Z_\mathcal{C}, M_\mathcal{C})$ from the noisy future token x_τ, task instruction I, and the context frames $(Z_\mathcal{C}, M_\mathcal{C})$. Supervision is applied only to valid target tokens whose modality is available in the corresponding training sample. We denote the resulting visual and motion flow matching losses by $\mathcal{L}_V$ and $\mathcal{L}_M$, respectively:

$$\mathcal{L}_V = \frac{1}{|\Omega_V|}\sum_{t\in\Omega_V} \left\| u_\theta^V(z_{\tau,t}, \tau, I, Z_\mathcal{C}) - u_t^{V\star} \right\|_2^2, \qquad \mathcal{L}_M = \frac{1}{|\Omega_M|}\sum_{t\in\Omega_M} \left\| u_\theta^M(m_{\tau,t}, \tau, I, M_\mathcal{C}) - u_t^{M\star} \right\|_2^2.$$

Each sample activates only the terms associated with its observed modalities: paired samples use both visual $\mathcal{L}_V$ and motion losses $\mathcal{L}_M$, video-only samples use only $\mathcal{L}_V$, and motion-only samples use only $\mathcal{L}_M$. Whenever motion targets are available, we additionally apply geometry-aware regularization

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

which encourages physically consistent global motion and local end-effector dynamics. The definitions of these auxiliary terms are provided in Appendix A.3.

The overall pre-training objective is

$$\min_\theta \ \mathbb{E}_{s\sim\mathcal{D}_{VM}}\left[\mathcal{L}_V + \mathcal{L}_M + \mathcal{L}_{\text{geom}}\right] + \mathbb{E}_{s\sim\mathcal{D}_V}\left[\mathcal{L}_V\right] + \mathbb{E}_{s\sim\mathcal{D}_M}\left[\mathcal{L}_M + \mathcal{L}_{\text{geom}}\right],$$

where $\mathcal{D}_{VM}$, $\mathcal{D}_V$, and $\mathcal{D}_M$ denote paired video-motion, video-only, and motion-only data streams.

Post-Training. The pretrained prior world-action model with parameters θ learns a video-motion generative model from egocentric data, where predicted motion provides a coarse-grained action space for future behavior. To adapt this prior world-action model to real-robot control, we attach a lightweight action expert with parameters ϕ and post-train the combined model on robot trajectories.

We construct the robot motion sequence M from robot odometry and joint states using forward kinematics, in the same representation used for human motion. Consequently, each robot trajectory supplies not only executable-action supervision, but also vision and motion targets. **These additional training signals preserve and ground the pretrained video-motion dynamics** while the action expert learns embodiment-specific commands.

During post-training, we sample a video-motion window from a robot trajectory and select an action query time q within the prediction interval. The video frames are encoded into visual latents, and the prior world-action model is trained with the same visual and motion flow matching losses used during pre-training. From its

forward pass, we extract hidden states $H^\theta(\tau)$ from selected layers, which summarize the task-conditioned video-motion context and are provided to the action expert through a one-way connection.

We adopt an action-chunking strategy [72], where the policy predicts a sequence of consecutive actions at each query step q. Given the ground-truth action chunk $\mathbf{a}_q = a_{q:q+h-1}$, we sample $\epsilon_a \sim \mathcal{N}(0, I)$ and construct $\tilde{\mathbf{a}}_\tau = (1-\tau)\mathbf{a}_q + \tau\epsilon_a$. The action expert predicts the corresponding flow matching velocity conditioned on the noisy action chunk, prior hidden context, robot observation O_q:

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

The final post-training objective is

$$\min_{\theta,\phi}\ \mathcal{L}_V(\theta) + \mathcal{L}_M(\theta) + \lambda_a \mathcal{L}_A(\theta, \phi),$$

where λ_a controls the weight of action supervision. Additional post-training details are provided in Appendix A.4.

Inference. At inference time, the prior world-action model is run periodically to generate a future video-motion plan from recent observations and task text. Its intermediate hidden states are converted into policy-side caches and reused across multiple control windows.

In addition to feedback through updated images and proprioception, **the unified motion representation exposes a motion-level feedback path**: recent robot kinematics can be compared with and incorporated into the predicted coarse whole-body plan before the cache is refreshed. The action expert can therefore respond to deviations in body and end-effector motion without relying on visual feedback alone.

At denoising step n, the action expert predicts the velocity field and updates the action sample as

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

where H_n^θ denotes the cached prior at step n, and O_{cur} denotes the current robot observation, respectively.

The action expert generates action chunks at a high frequency by reusing the current KV cache of the prior model, which is refreshed at a lower frequency with updated robot feedback before its cached context is fully consumed. This design separates lower-frequency world planning from low-latency action generation, while keeping the policy grounded in both the current plan produced by the prior world-action model and real-time robot observations. The detailed inference formulation is provided in Appendix A.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 equipped with Linker Hand O6 dexterous hands for upper-body manipulation, together with an Intel RealSense D435i RGB camera mounted on the robot head to provide egocentric visual observations. Policy inference is performed 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 the 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, two ankle-mounted PICO trackers, and two handheld controllers to provide full-body motion. XRoboToolkit [73] estimates the operator's SMPL pose in real time from the VR device streams, while SONIC [19] converts the estimated pose into stable 29-DoF whole-body control commands executed

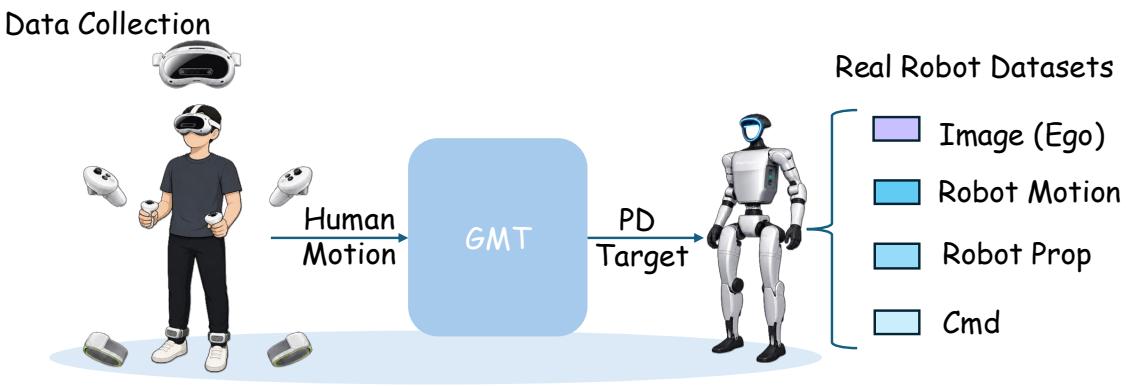

Figure 5: **Overview of the data collection system.** During the data collection stage, an operator uses a VR-based teleoperation system to generate lower-level motion commands, while a pretrained general whole-body motion tracker controls the humanoid robot to execute the corresponding motions and collect synchronized real-world trajectories, including egocentric images, proprioceptive observations, and motion commands.

on the Unitree G1 at 50 Hz. During teleoperation, synchronized robot proprioception, egocentric RGB observations, motion commands (including hand control signals, and SONIC motion tokens) are recorded to construct the task-specific post-training dataset. The overview of the proposed system is as shown in Figure 5.

4.2 Real-World Experiments

We evaluate Being-M0.7 on four real-world tasks: **Mirror Toy Grasping**, **Water-Tank Fish Scooping**, **Tabletop Organization**, and **Obstacle-Avoidance Basket Carrying**. We report quantitative comparisons with the baselines on Mirror Toy Grasping and Water-Tank Fish Scooping, and present qualitative case studies for all four tasks in Figure 6. Together, these experiments cover visual reasoning under partial observability, dexterous tool use, long-horizon manipulation, and coordinated locomotion and manipulation. We compare our method with Ψ_0 [4] and GR00T-N1.6 [61].

Table 1: Quantitative real-world evaluation on Water-Tank Fish Scooping and Mirror Toy Grasping.

Method	Mirror (Near)	Mirror (Far)	Fish	Overall
GR00T-N1.6	1/5	0/5	1/5	2/15
Ψ_0	0/5	1/5	2/5	3/15
Being-M0.7	**3/5**	**1/5**	**3/5**	**7/15**

Quantitative Evaluation. To make the comparison as fair as possible, we place physical markers at the predefined robot starting positions and initialize all methods from the same marked positions.

- **Mirror Toy Grasping.** As shown in Table 1, the robot starts either 0.5 m (Near) or 1 m (Far) from the workspace. It must approach the table, infer the location of a hidden toy from its mirror reflection, and grasp the toy. This task evaluates visual reasoning under partial observability together with whole-body locomotion and manipulation. Being-M0.7 succeeds in 4/10 trials overall, compared with 1/10 for both GR00T-N1.6 and Ψ_0.
- **Water-Tank Fish Scooping.** As shown in Table 1, we evaluate only the manipulation stage: the robot is initialized standing at the designated position in front of the water tank and uses a handheld net to acquire a toy fish. This setting does not include the preceding locomotion phase and focuses on visual feedback, tool use, and coordinated arm control. Being-M0.7 succeeds in 3/5 trials, outperforming GR00T-N1.6 (1/5) and Ψ_0 (2/5).

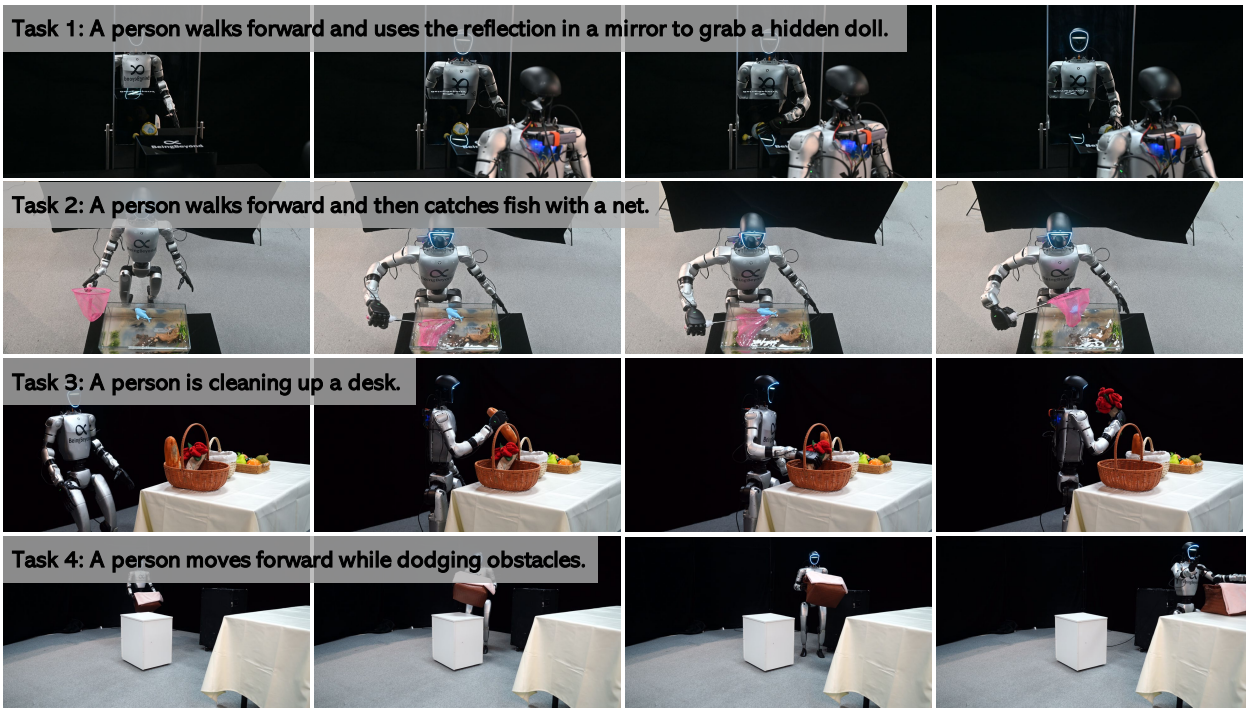

Figure 6: **Representative real-world case studies.** We evaluate Being-M0.7 on four challenging whole-body manipulation tasks that require complementary embodied capabilities. From left to right and top to bottom: (a) **Mirror Toy Grasping:** reasoning under partial observability through mirror reflections, (b) **Water-Tank Fish Scooping:** tool use under uncertain object dynamics, (c) **Tabletop Organization:** long-horizon multi-stage manipulation, and (d) **Obstacle-Avoidance Basket Carrying:** coordinated locomotion and load-aware manipulation with obstacle avoidance. Together, these tasks demonstrate robust closed-loop coordination between perception, reasoning, locomotion, and dexterous manipulation.

Qualitative Case Studies. In the qualitative setting of **Water-Tank Fish Scooping**, we demonstrate the complete whole-body behavior, including approaching the tank, positioning the body for interaction, and subsequently scooping a toy fish with the net. This setting highlights the integration of locomotion and tool-based manipulation over an extended trajectory, complementing the quantitative evaluation of the scooping stage alone. The remaining case studies assess complementary capabilities: **Mirror Toy Grasping** examines reflection-based reasoning and dexterous grasping; **Tabletop Organization** evaluates long-horizon, multi-stage manipulation; and **Obstacle-Avoidance Basket Carrying** tests obstacle-aware locomotion and stable object transportation. Detailed task descriptions are provided in Appendix B.1.

5 Conclusion

We presented Being-M0.7, a latent world-action model for humanoid loco-manipulation. The key idea is to learn a video-motion world prior from egocentric data and adapt it to executable humanoid control through action-level post-training. By modeling future visual states in a latent space and coupling them with compact motion dynamics, Being-M0.7 provides a structured planning context for downstream robot action generation without relying on pixel-level world prediction.

Experiments in real world demonstrate the potential of latent world-action modeling for whole-body loco-manipulation tasks that require perception, motion prediction, and coordinated control. While the current system shows promising capabilities, it also points to several directions for future work, including more sophisticated post-training process, scaling robot post-training data, improving robustness under distribution shifts, and extending the framework to broader tasks, environments, and humanoid embodiments.

References

[1] 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*, pages 1516–1540. PMLR, 2025.

[2] 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.

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

[4] 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*, 2026.

[5] 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.

[6] 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.

[7] 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.

[8] 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*, pages 2165–2183. PMLR, 2023.

[9] 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)*, pages 6892–6903. IEEE, 2024.

[10] 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.

[11] 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.

[12] 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*, pages 2679–2713. PMLR, 2025.

[13] 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.

[14] 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.

[15] 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.

[16] 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*, 2026.

[17] 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*, 2026.

[18] Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: A fine-grained world model for robot manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9834–9844, 2025.

[19] 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.

[20] 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.

[21] 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*, pages 18995–19012, 2022.

[22] Xiaofeng Wang, Kang Zhao, Feng Liu, Jiayu Wang, Guosheng Zhao, Xiaoyi Bao, Zheng Zhu, and Yingya Zhang. Egovid-5m: A large-scale video-action dataset for egocentric videos generation. *Advances in Neural Information Processing Systems*, 38, 2026.

[23] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.

[24] 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*, pages 445–465. Springer, 2024.

[25] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019.

[26] 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*, pages 5152–5161, 2022.

[27] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36:25268–25280, 2023.

[28] 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, pages 99939–99955, 2025.

[29] 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.

[30] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

[31] 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.

[32] Zepeng Wang, Jiangxing Wang, Shiqing Yao, Yu Zhang, Ziluo Ding, Ming Yang, Yuxuan Wang, Haobin Jiang, Chao Ma, Xiaochuan Shi, et al. General humanoid whole-body control via pretraining and fast adaptation. *arXiv preprint arXiv:2602.11929*, 2026.

[33] Yuhui Fu, Feiyang Xie, Chaoyi Xu, Jing Xiong, Haoqi Yuan, and Zongqing Lu. Demohlm: From one demonstration to generalizable humanoid loco-manipulation. *IEEE Robotics and Automation Letters*, 2026.

[34] Bohan Zhou, Haoqi Yuan, Yuhui Fu, and Zongqing Lu. Learning diverse bimanual dexterous manipulation skills from human demonstrations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 28919–28927, 2026.

[35] Yitang Li, Yuanhang Zhang, Wenli Xiao, Chaoyi Pan, Haoyang Weng, Guanqi He, Tairan He, and Guanya Shi. Hold my beer: Learning gentle humanoid locomotion and end-effector stabilization control. In *Conference on Robot Learning*, pages 4506–4523. PMLR, 2025.

[36] Qingzhou Lu, Yao Feng, Baiyu Shi, Michael Piseno, Zhenan Bao, and C Karen Liu. Gentlehumanoid: Learning upper-body compliance for contact-rich human and object interaction. *arXiv preprint arXiv:2511.04679*, 2025.

[37] Yixuan Li, Yutang Lin, Jieming Cui, Tengyu Liu, Wei Liang, Yixin Zhu, and Siyuan Huang. Clone: Closed-loop whole-body humanoid teleoperation for long-horizon tasks. In *9th Annual Conference on Robot Learning*, 2025.

[38] Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C Karen Liu, Rocky Duan, and Guanya Shi. Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. *arXiv preprint arXiv:2509.26633*, 2025.

[39] 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.

[40] 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.

[41] Tairan He, Zi Wang, Haoru Xue, Qingwei Ben, Zhengyi Luo, Wenli Xiao, Ye Yuan, Xingye Da, Fernando Castañeda, Shankar Sastry, et al. Viral: Visual sim-to-real at scale for humanoid loco-manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13430–13441, 2026.

[42] Ilyass Taouil, Michal Ciebelski, Shafeef Omar, Haizhou Zhao, Angela Dai, Aaron M Johnson, and Majid Khadiv. Motiondisco: Motion discovery for extreme humanoid loco-manipulation. *arXiv preprint arXiv:2606.06139*, 2026.

[43] Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. InterMimic: Towards universal whole-body control for physics-based human-object interactions. In *CVPR*, 2025.

[44] 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.

[45] 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.

[46] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14730–14740, 2023.

[47] 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.

[48] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4115–4128, 2024.

[49] 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.

[50] Yuan Wang, Di Huang, Yaqi Zhang, Wanli Ouyang, Jile Jiao, Xuetao Feng, Yan Zhou, Pengfei Wan, Shixiang Tang, and Dan Xu. Motiongpt-2: A general-purpose motion-language model for motion generation and understanding. *arXiv preprint arXiv:2410.21747*, 2024.

[51] Mingyuan Zhang, Daisheng Jin, Chenyang Gu, Fangzhou Hong, Zhongang Cai, Jingfang Huang, Chongzhi Zhang, Xinying Guo, Lei Yang, Ying He, et al. Large motion model for unified multi-modal motion generation. In *European Conference on Computer Vision*, pages 397–421. Springer, 2024.

[52] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16010–16021, 2023.

[53] Gaoge Han, Mingjiang Liang, Jinglei Tang, Yongkang Cheng, Wei Liu, and Shaoli Huang. Reindiffuse: Crafting physically plausible motions with reinforced diffusion model. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2218–2227, 2025.

[54] Junpeng Yue, Zepeng Wang, Yuxuan Wang, Weishuai Zeng, Jiangxing Wang, Xinrun Xu, Yu Zhang, Sipeng Zheng, Ziluo Ding, and Zongqing Lu. Rl from physical feedback: Aligning large motion models with humanoid control. *arXiv preprint arXiv:2506.12769*, 2025.

[55] Agon Serifi, Ruben Grandia, Espen Knoop, Markus Gross, and Moritz Bächer. Robot motion diffusion model: Motion generation for robotic characters. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–9, 2024.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] Shuanghao Bai, Meng Li, Xinyuan Lv, Jiawei Wang, Xinhua Wang, Fei Liao, Chengkai Hou, Langzhe Gu, Wanqi Zhou, Kun Wu, et al. Hex: Humanoid-aligned experts for cross-embodiment whole-body manipulation. *arXiv preprint arXiv:2604.07993*, 2026.

[63] 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*, pages 38187–38203. PMLR, 2023.

[64] 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.

[65] 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.

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

[67] 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*, pages 30640–30649, 2026.

[68] 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.

[69] 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.

[70] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[71] 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.

[72] 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.

[73] 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)*, pages 15–20. IEEE, 2026.

Appendix

Author List

Core Contributors: Junpeng Yue*, Boyuan Li*, Yuxuan Wang*, Zepeng Wang, Yuhui Fu, Feiyang Xie, Yu Zhang, Jing Zhang, Jiangxing Wang, Zongqing Lu† [1]

Contributors: Weibo Li, Xiaofei Zheng, Yuming Fang

A Implementation Details

A.1 Architecture Hyperparameters

Table 2 reports the model sizes used in our implementation. Prior-layer indices are one-based.

Table 2: Architecture hyperparameters of the video–motion prior and action expert.

Hyperparameter	Value
Video–Motion Prior	
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	6
Hidden width	384
Attention heads	8
FFN expansion ratio	4
Connected prior layers	1, 7, 13, 18, 24, 30
Action chunk length	10
Dropout	0.1

A.2 Hyperparameters

Table 3 summarizes the hyperparameters and settings used in our experiments.

A.3 Geometry-aware Auxiliary Losses

In addition to the flow matching loss, the motion branch uses two geometry-aware auxiliary losses. 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.

The global trajectory auxiliary loss is

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

[1] *Equal Contribution †Corresponding Author

Table 3: Hyperparameters and training settings.

Hyperparameter	Value
General Settings	
Sequence length T	25
Context length K	5
Input resolution	224×224
Prior frame rate	5Hz
Action expert frame rate	50Hz
Optimizer	AdamW
Weight decay	0.05
Pre-training	
Batch size	256 (per GPU)
Learning rate	0.0002
Warmup epochs	10
Gradient clipping	1.0
Post-training	
Batch size	128 (per GPU)
Learning rate	0.0001
Warmup steps	1000
Gradient clipping	1.0

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

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

where $\Delta\ell(m_t) = \ell(m_t) - \ell(m_{t-1})$ and d_ℓ is the dimensionality of the local end-effector coordinates. For the 22D motion representation, $\ell(m_t)$ corresponds to the local positions of the two hands and two feet in the head/root frame.

A.4 Post-training Details

Action Space. In theory, the action expert is compatible with any kind of low-level robot controllers. Currently, we only focus on the WBC setting. For this setting, we use SONIC [19]. The action expert predicts the universal control token required by SONIC along with the dexterous-hand open/close commands.

Robot Observation. The robot observation O_q consists of the current egocentric image, proprioceptive state and normalized execution progress. The proprioceptive state includes joint positions, joint velocities, and the gravity direction projected into the robot frame, while the egocentric image is encoded using the same frozen DINO encoder as the visual stream.

Prior-Layer Selection. Because the prior world-action model is deeper than the action expert, the prior layers connected to the expert are selected approximately uniformly across the prior depth. Each expert layer is paired with one selected prior layer.

Temporal Configuration. Video and motion sequences are represented at 5 Hz, while robot commands are executed at 50 Hz. The action expert predicts chunks of length

$$H_a = 20,$$

corresponding to a 0.4-second action horizon.

A.5 Inference Details

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

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

During prior denoising, we extract the intermediate representations

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

from the selected prior layers. These representations are converted into policy-side key-value caches. Following the notation in the main text, we collectively denote the cached prior context associated with denoising step n as

$$H_n^\theta.$$

Because the prior world-action 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}denote 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

$$\mathbf{a}^N = (a_t, \ldots, a_{t+H_a-1}), \qquad H_a = 20.$$

The cached prior context is refreshed every 3.5 seconds, while the action expert is queried every 0.4 seconds. Thus, one prior rollout is reused for approximately nine action-expert queries. At every query, the action expert receives the latest O_{cur}, allowing it to perform closed-loop correction while reusing the lower-frequency prior context. A single action-expert query requires less than 0.01 seconds, and the resulting commands are executed by the low-level controller at 50 Hz.

B Real World Experiment Details

B.1 Task description

We present detailed descriptions of the four real-world tasks and visualize representative qualitative executions in Figures 7–10. For the quantitative evaluations on Mirror Toy Grasping and Water-Tank Fish Scooping, we place physical markers at the predefined robot starting positions and initialize all methods from the same marked positions to make the comparison as fair as possible.

Mirror Toy Grasping. The robot walks toward a table and approaches a partially enclosed box containing a toy. Since the toy is initially invisible from the robot's viewpoint, it must infer the object's location from its reflection in a mirror before reaching into the box for grasping. In the quantitative evaluation, the robot starts either 0.5 m or 1 m from the workspace. This task requires reasoning under partial observability together with accurate whole-body positioning and dexterous manipulation.

Water-Tank Fish Scooping. We consider two settings for this task. In the quantitative setting, the robot is initialized standing at the designated position in front of the water tank, and only the manipulation stage of scooping a toy fish with a handheld net is evaluated. In the qualitative setting, the robot performs the complete whole-body sequence, including approaching the tank, positioning its body for interaction, and scooping a toy fish. Compared with direct grasping, this task requires coordinated tool use while handling water-induced visual distortion and changing object locations. The two settings respectively assess the scooping stage and its integration with locomotion in a longer-horizon behavior.

Tabletop Organization. The robot walks toward a table, transfers a baguette between two baskets, then grasps a rose before leaving the workspace. The task consists of multiple sequential manipulation stages and evaluates long-horizon planning, scene understanding, and coordinated locomotion-manipulation behaviors.

Obstacle-Avoidance Basket Carrying. The robot carries a basket through an environment containing multiple obstacles. During navigation, it performs forward walking and lateral stepping while maintaining balance and avoiding collisions. This task evaluates coordinated locomotion, obstacle avoidance, and stable object transportation.

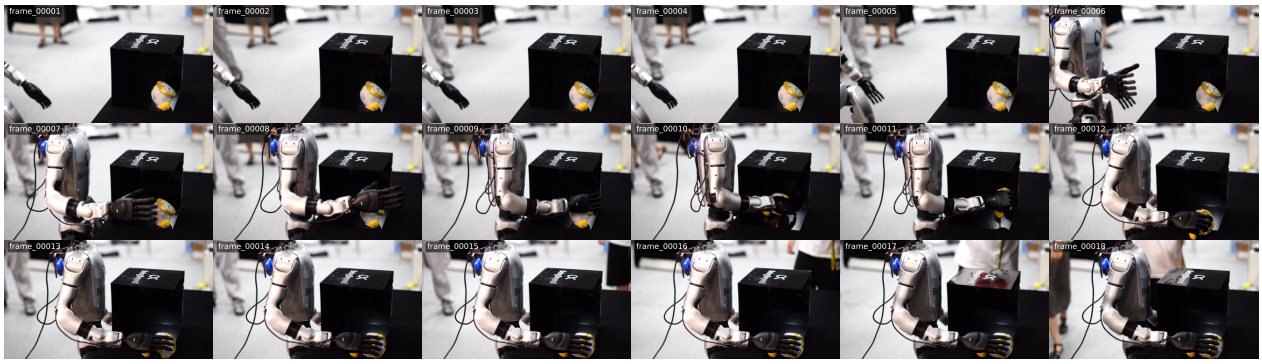

Figure 7: Complete snapshot of Mirror Toy Grasping task.

Figure 8: Complete snapshot of Water-Tank Fish Scooping task.

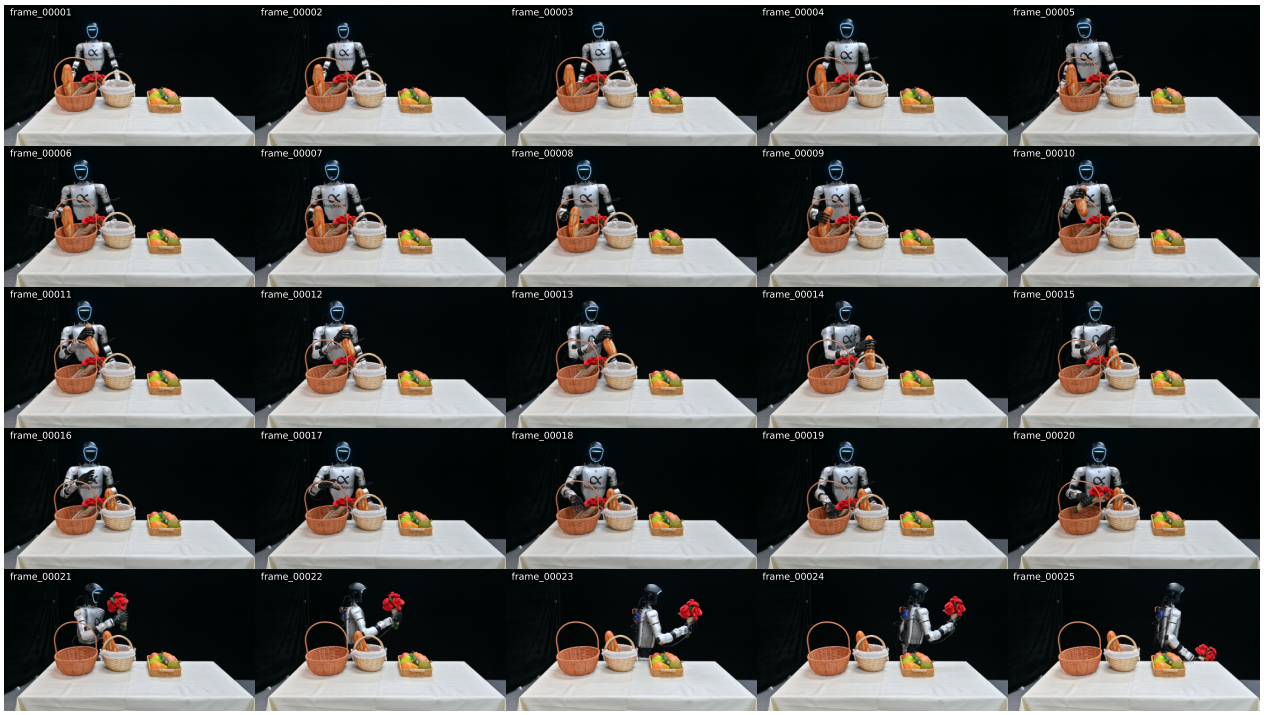

Figure 9: Complete snapshot of Obstacle-Avoidance Basket Carrying.

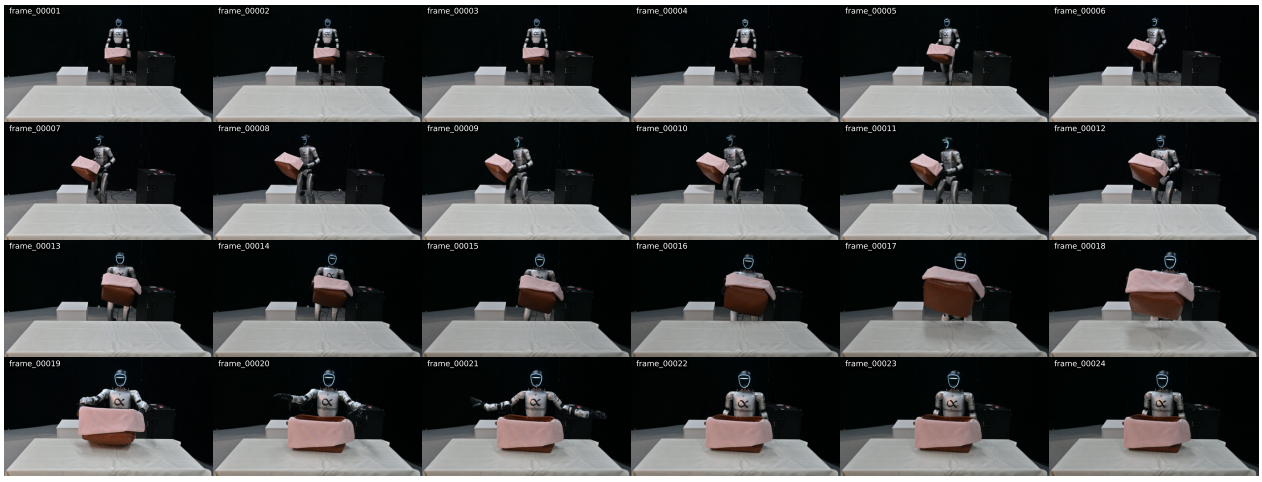

Figure 10: Complete snapshot of Tabletop Organization.