HEX: Humanoid-Aligned Experts for Cross-Embodiment Whole-Body Manipulation

Anonymous Authors

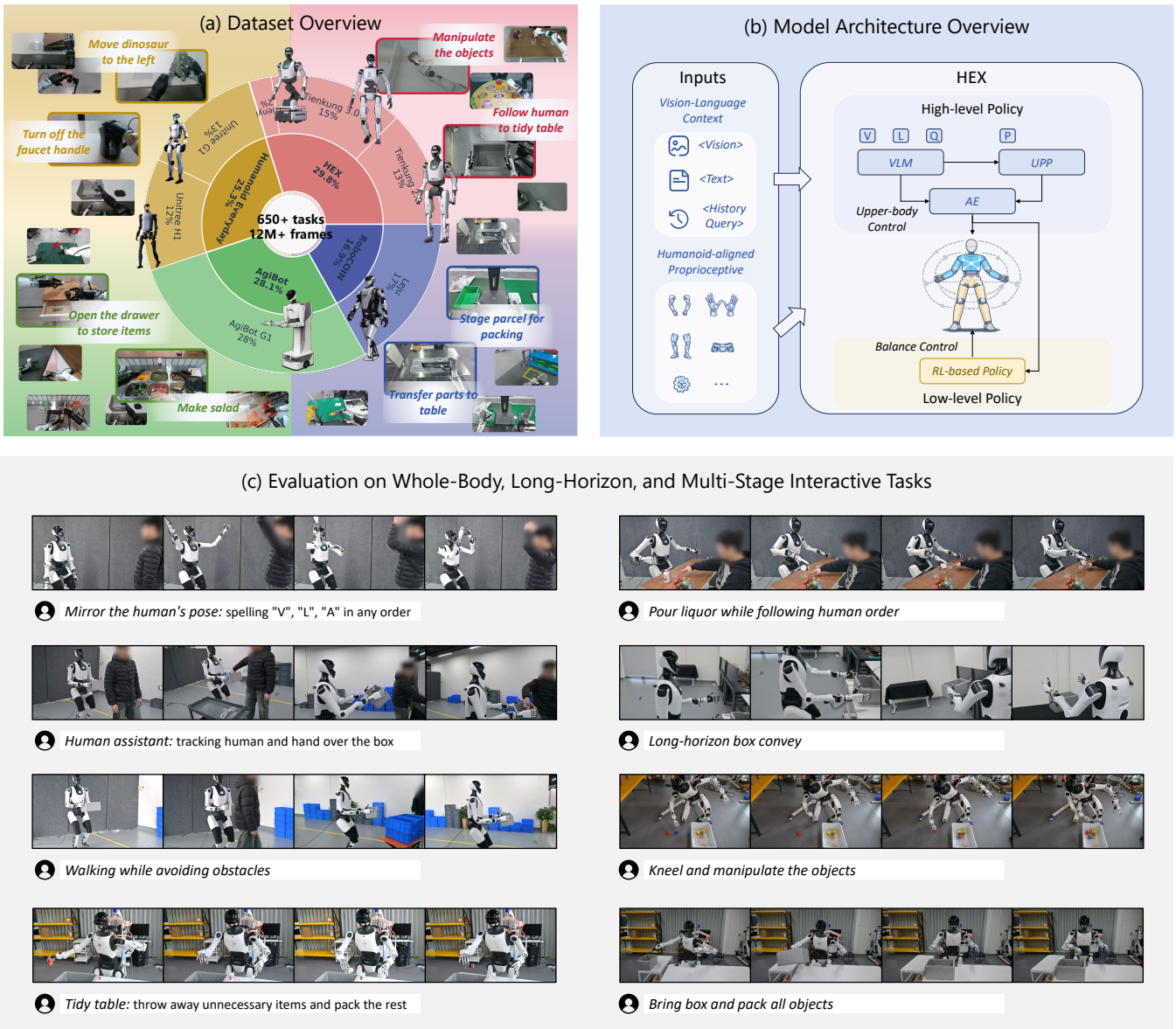


Fig. 1. **Overview of HEX.** (a) HEX is, to the best of our knowledge, the first whole-body VLA framework for full-sized bipedal humanoid robots, pretrained on diverse cross-embodiment humanoid trajectory data. (b) HEX consists of a high-level VLA module and a low-level whole-body controller. The high-level module takes visual-language context and humanoid-aligned proprioceptive state as input, where the VLM encodes visual-language information and the Unified Proprioceptive Predictor (UPP) models structured whole-body state dynamics across body parts. Their features are integrated through dual cross-attention and residual-gated fusion for action generation. The resulting high-level actions directly control the arms and hands, while also serving as intermediate commands for the low-level controller to generate balance-preserving whole-body motions. (c) We validate HEX on Tienkung 2.0 and Tienkung 3.0 across whole-body, long-horizon, and fast-reaction tasks, demonstrating strong performance in diverse interactive manipulation scenarios.

***Abstract*—Humans achieve complex manipulation through coordinated whole-body control, whereas most Vision-Language-Action (VLA) models treat robot body parts largely independently, making high-DoF humanoid control challenging and often unstable. We present HEX, a state-centric framework for coordinated manipulation on full-sized bipedal humanoid robots. HEX introduces a humanoid-aligned universal state representation for scalable learning across heterogeneous embodiments,**

and incorporates a Mixture-of-Experts Unified Proprioceptive Predictor to model whole-body coordination and temporal motion dynamics from large-scale multi-embodiment trajectory data. To efficiently capture temporal visual context, HEX uses lightweight history tokens to summarize past observations, avoiding repeated encoding of historical images during inference. It further employs a residual-gated fusion mechanism with a flow-matching action head to adaptively integrate visual-language cues with proprioceptive dynamics for action generation. Experiments on real-world humanoid manipulation tasks show that HEX achieves state-of-the-art performance in task success rate and generalization, particularly in fast-reaction and long-horizon scenarios.

***Index Terms*—Vision–Language–Action models, humanoid manipulation, mixture-of-experts, cross-embodiment learning**

I. Introduction

Humanoid robots hold the promise of bringing embodied intelligence into complex human environments such as homes and schools. Existing research, however, has largely focused on either locomotion, which enables robots to navigate unstructured environments [1], [2], [3], or hand-centric manipulation, where the lower body remains fixed and control is limited to the arms and hands [4], [5]. In contrast, humans routinely perform tasks that require simultaneous locomotion and manipulation, leveraging coordinated motion of the entire body. Enabling such whole-body manipulation in humanoids remains significantly underexplored. The challenge is fundamental: the robot must maintain dynamic balance while producing high-dimensional, tightly coupled motions across multiple limbs during object interaction.

Existing approaches to humanoid whole-body manipulation mainly follow two paradigms. The first adopts an explicitly decomposed design, where locomotion or navigation and manipulation are controlled by separate policies [6]. While this decomposition simplifies learning and control, it depends on manual task priors and interface design, and becomes increasingly brittle as task complexity grows. Errors can accumulate across modules, making tightly coupled behaviors such as manipulating while moving difficult to achieve. A more recent trend is to adopt a hierarchical design in which the high-level module produces task-relevant commands, such as arm and hand actions [7], [8] or corresponding hand-eye targets [9], while a low-level whole-body controller refines them into high-frequency, balance-preserving motions.

In parallel, recent Vision-Language-Action (VLA) models have introduced stronger visual-semantic understanding and reasoning through large vision-language models, showing promising scalability and generalization [10], [11], [12]. As a result, recent humanoid systems increasingly adopt VLA-style high-level planners together with low-level whole-body controllers for stable execution [7], [8], [13], [14]. Despite these advances, most existing VLA-based approaches remain insufficiently structured for humanoid whole-body manipulation. In many cases, actions are predicted over high-dimensional joints or latent commands without explicitly modeling how body parts interact through shared balance and posture. As a result, the policy may capture task intent semantically, yet still fail to produce coordinated whole-body behavior, especially in fast-reaction and long-horizon scenarios where temporal consistency and whole-body coordination are essential.

To address this limitation, we propose **HEX**, a framework built on the key insight that effective humanoid whole-body manipulation requires both embodiment-aware predictive dynamics and temporally grounded scene understanding, as shown in Figure 1. Specifically, HEX introduces a humanoid-aligned universal state representation that provides a structured basis for modeling whole-body proprioceptive dynamics across different body parts and embodiments. Built upon this representation, HEX captures temporal motion evolution and whole-body coordination in proprioceptive space, enabling scalable predictive modeling for heterogeneous humanoid trajectories. Whole-body manipulation also depends on temporal visual context, especially when object motion, scene evolution, or partial observability makes the current observation insufficient. To this end, HEX summarizes past visual-language context into compact representations while leveraging predictive proprioceptive dynamics to provide state foresight. Together, these designs form a *review-and-forecast* paradigm, where past visual context supports scene understanding and future state prediction supports coordinated whole-body control. The resulting visual-language and predictive state representations are then adaptively fused for action generation, producing smooth and coordinated whole-body behaviors.

We evaluate HEX on a diverse set of real-world humanoid manipulation tasks against strong VLA and imitation learning baselines, including ACT [15], SwitchVLA [16], GR00T N1.5 [13], and $\Pi_{0.5}$ [11]. Across a wide range of task settings, HEX consistently achieves higher task success rates and stronger generalization. The improvements are particularly pronounced in fast-reaction and long-horizon scenarios, where coordinated whole-body dynamics and temporal consistency are critical. In summary, our contributions are fourfold. First, to the best of our knowledge, we present the first whole-body VLA framework for full-sized bipedal humanoid robots. Second, we propose a cross-embodiment humanoid-aligned state representation with predictive proprioceptive modeling for scalable whole-body pretraining. Third, we introduce a review-and-forecast paradigm that combines visual history summarization, future state prediction, and adaptive multimodal fusion for action generation. Finally, extensive experiments on real-world humanoid manipulation benchmarks demonstrate state-of-the-art performance and validate HEX as an effective framework for coordinated whole-body manipulation.

II. Related Work

A. Learning-based Humanoid Whole-body Control

Learning-based humanoid whole-body control has been primarily advanced through reinforcement learning (RL) and imitation learning (IL) [17]. Early RL-based methods such as DeepMimic [2] and AMP [18] established motion-tracking and motion-prior-based policy learning as effective paradigms for acquiring robust and natural humanoid skills. More recent work has extended this line toward real-world, contact-rich, and visually grounded whole-body control, including simulation-pretrained latent action learning for real-world RL [19], force-adaptive loco-manipulation [20], highly

dynamic full-body skill learning [21], large-scale motion-tracking controllers with strong generalization [22], and visual sim-to-real humanoid loco-manipulation via privileged RL and teacher–student policy distillation [23]. In parallel, imitation learning has emerged as an efficient alternative by leveraging human demonstrations, teleoperation trajectories, and motion priors. Recent approaches explore human-to-humanoid imitation from teleoperation or egocentric demonstrations [24], [25], [26], unified motion-tracking and predictive motion priors [27], [28], as well as generative imitation with diffusion-based policies [29], [30]. Despite their success, these methods are primarily designed for skill imitation or task-specific control, and generally provide limited semantic understanding of instructions, goals, and visual context.

More recently, vision–language–action (VLA) models have begun to extend from fixed-base manipulation to humanoid whole-body control. Humanoid-VLA [31] introduces visual integration for humanoid control, while GR00T N1 [13] and Ψ_0 [14] move toward generalist humanoid foundation models trained on large-scale heterogeneous data. To better handle agile whole-body behaviors, LeVERB [32] proposes hierarchical latent vision-language instructions, WholeBodyVLA [8] explores unified latent VLA control for loco-manipulation, and TrajBooster [7] improves downstream adaptation through trajectory-centric retargeting. In contrast to these approaches, which mainly improve semantic grounding and multimodal conditioning, our work explicitly models structured proprioceptive dependencies for humanoid whole-body control, coupling visual-language reasoning with humanoid-aligned state representations and joint past-future temporal modeling to enable coordinated whole-body behavior.

B. Cross-Embodiment Learning for Humanoid Robots

Cross-embodiment learning seeks to transfer knowledge across agents with different morphologies by learning shared behavior representations, aligned control spaces, or generalizable pretrained policies [33], [34], [35]. One line of work focuses on *human-to-humanoid learning*, where human videos or egocentric demonstrations are used as scalable supervision for humanoid control [36], [26], [37], [38], [39], [14]. Representative examples include Mao et al. [36], which leverage large-scale human videos for humanoid pose control, Humanoid Policy∼Human Policy [26], which aligns egocentric human demonstrations with humanoid behaviors in a unified policy space, and Ψ_0 [14], which incorporates human egocentric videos into a staged humanoid foundation-model training recipe. Another line studies *robot-to-humanoid* or *cross-humanoid learning*, where policies or representations are transferred across heterogeneous robotic embodiments [40], [41], [5], [42], [43], [13], [44]. Representative works include H-Zero [40], which enables few-shot transfer to novel humanoids through cross-humanoid pretraining, EAGLE [43], which learns a unified controller across diverse humanoid embodiments, and GR00T N1 [13], which scales humanoid foundation modeling with heterogeneous robot trajectories, human videos, and synthetic data.

Similar in spirit, HEX also targets cross-embodiment humanoid learning, but differs from prior works by introducing a compositional and humanoid-aligned proprioceptive modeling framework. Its Unified Proprioceptive Predictor with Mixture-of-Experts (MoE) operates on canonical body-part abstractions, allowing heterogeneous trajectories from the same embodiment or different humanoids to be encoded in a shared latent space without retraining a monolithic state encoder for every new joint configuration or missing-part setting. By combining reusable part-level encoders with dynamic expert routing, HEX more efficiently exploits both intra- and cross-embodiment data, while capturing structured whole-body and temporal dependencies for coordinated whole-body control.

III. Method

A. Overview

HEX adopts a hierarchical architecture for humanoid whole-body manipulation, consisting of a high-level VLA policy and a low-level RL-based whole-body controller. The high-level policy takes visual-language context together with humanoid-aligned proprioceptive state as input, and produces task-relevant actions for manipulation. These outputs directly govern arm and hand behavior, while also serving as intermediate commands for the low-level controller. The low-level controller operates at a higher control frequency and generates balance-preserving, dynamically feasible whole-body motions for stable execution during locomotion and manipulation.

The high-level VLA policy in HEX consists of three main components. First, a Visual-Language Model (VLM) module encodes current visual-language context together with lightweight temporal review cues. Second, a Unified Proprioceptive Predictor (UPP) models humanoid-aligned state dynamics and captures whole-body interactions through predictive proprioceptive modeling. Third, an action expert integrates visual-language and proprioceptive features through adaptive fusion to generate the final high-level action. Figure 2 provides an overview of the full framework. For the low-level controller, we instantiate skill-specific RL policies trained with motion-guided objectives. In particular, the standing and walking controllers are trained with a DeepMimic-style reference-tracking formulation [2], which is well suited to stable periodic or quasi-static motions with clear target kinematics. In contrast, the half-kneeling controller is trained with an AMP-style objective [18], where an adversarial motion prior encourages natural contact-rich posture transitions without requiring strict frame-wise tracking to a single reference trajectory. In the following, we present the core components of the high-level VLA policy in HEX.

B. VLM with History Query Feature Cache

To incorporate temporal visual-language context without repeatedly encoding long image histories, we introduce a lightweight history query feature cache. At each timestep t, we encode the language instruction, the current visual observation, and a query token using a single vision–language model (VLM). Specifically, we concatenate the language tokens $\mathbf{L}$, visual tokens $\mathbf{V}_t$ extracted from observation $\mathbf{o}_t$, and a query token $\mathbf{Q}_t$, and feed them into the VLM:

$$[\mathbf{L}'_t, \mathbf{V}'_t, \mathbf{Q}'_t] = f_{\mathrm{vlm}}([\mathbf{L}, \mathbf{V}_t, \mathbf{Q}_t]). \tag{1}$$

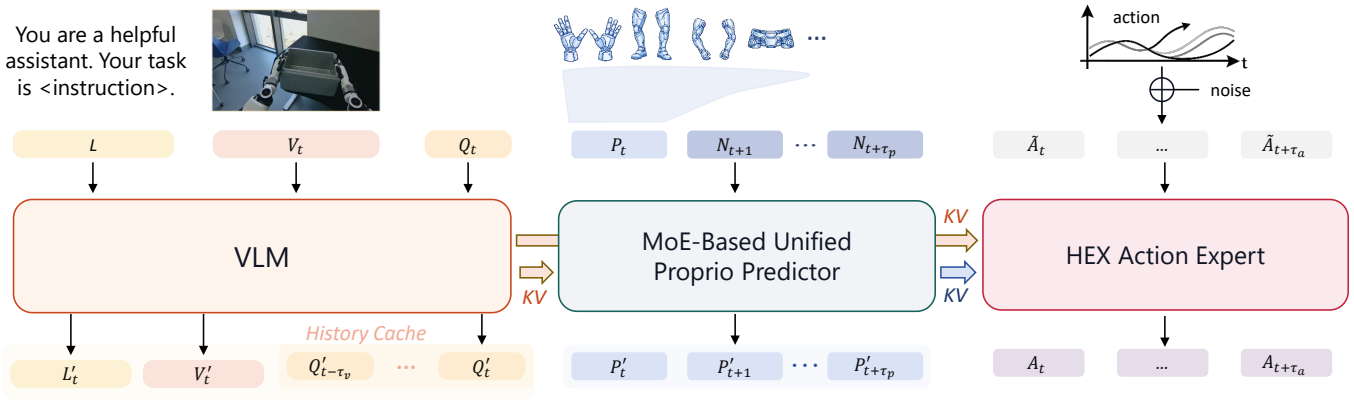


Fig. 2. **Overview of the proposed high-level VLA policy in HEX.** Given a language instruction L, the current visual observation V_t, and a history query token Q_t, the VLM encodes visual-language context together with lightweight temporal review cues summarized in a history cache. In parallel, humanoid-aligned proprioceptive states are organized into structured part-aware tokens and processed by a MoE-based Unified Proprioceptive Predictor, which captures whole-body interactions and forecasts future state dynamics. The resulting visual-language and predictive proprioceptive features are then integrated by the HEX Action Expert through adaptive fusion for action generation, producing task-relevant high-level actions over the prediction horizon.

The resulting query feature $\mathbf{Q}'_t$ serves as a compact summary of the current visual-language context. Rather than propagating the query token itself across time, we generate a fresh query feature at every timestep and store it in a fixed-length cache:

$$\mathcal{M}_t^{\mathrm{vl}} = \{\mathbf{Q}'_{t-\tau_v}, \ldots, \mathbf{Q}'_{t-1}, \mathbf{Q}'_t\}, \tag{2}$$

where τ_v denotes the visual history window, set to 2 in all experiments. This cache stores only compact query features, rather than the original images or full VLM activations. Together with the current-step visual-language features $\mathbf{L}'_t$ and $\mathbf{V}'_t$, $\mathcal{M}_t^{\mathrm{vl}}$ provides recent semantic context for the subsequent proprioceptive modeling and action generation modules.

This design provides an efficient form of visual review: temporal scene information is preserved through a compact feature memory, while the VLM itself remains a single-step feed-forward encoder applied only to the current frame. As a result, HEX can exploit temporal visual context without incurring the substantial cost of repeatedly encoding long visual histories.

C. UPP with Morphology-based MoE

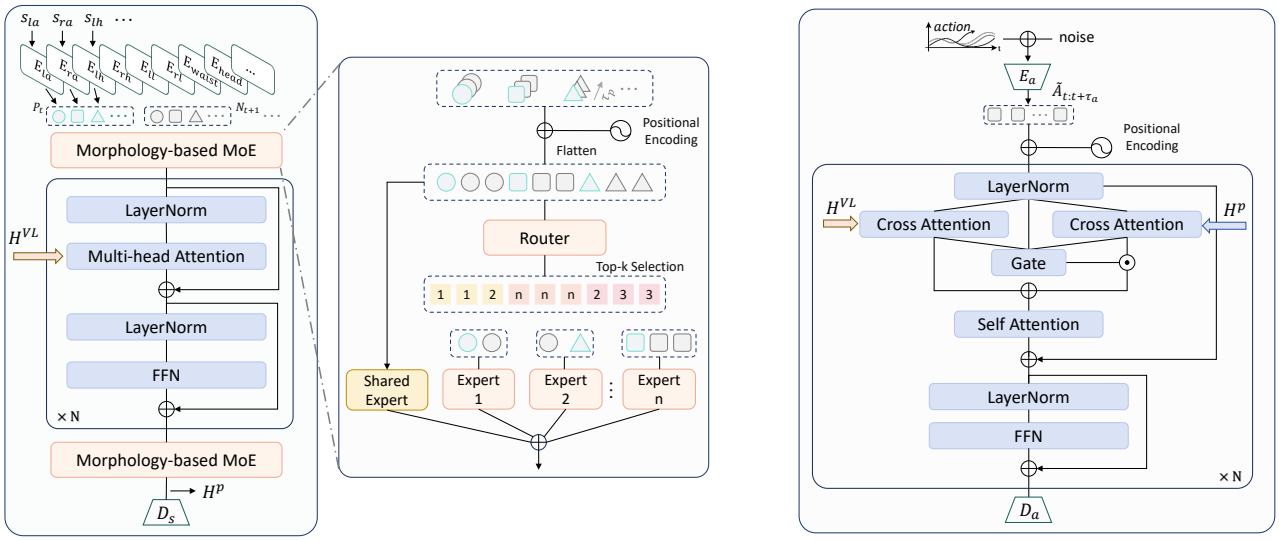
Humanoid proprioceptive state can take many forms and may include a rich combination of signals, such as whole-body joint positions, velocities, accelerations, and hand tactile feedback. As sensor suites become more capable, the amount and diversity of proprioceptive information available to the policy continue to grow. We argue that, in such settings, simply encoding the current state is insufficient for coordinated whole-body control, as the policy must model not only heterogeneous proprioceptive signals, but also the structured interactions among different body parts.

To enable efficient cross-embodiment learning over such heterogeneous proprioceptive observations, we organize the input state using a fixed set of canonical body-part slots, including left/right arms, left/right hands, left/right legs, head, waist, and an auxiliary `others` slot for remaining signals. Although we use this set of canonical slots in the current work, the formulation is readily extensible to richer or more fine-grained body-part decompositions. For an embodiment e, the raw proprioceptive state $\mathbf{s}_t^{(e)}$ may vary in dimensionality and composition. We therefore map each available part into a shared latent space and insert a learned missing-part token when a part is absent, yielding structured part latents:

$$\mathbf{P}_t^{(e)} \in \mathbb{R}^{P\times d}, \tag{3}$$

where P denotes the number of canonical part slots and d is the latent dimension. In this way, prediction is performed in a shared latent space rather than in the raw embodiment-specific state space. However, structured part representations alone are not sufficient for whole-body control. Beyond organizing heterogeneous proprioceptive signals into a common body-part space, the policy must still model how different body parts interact and evolve jointly over time. To this end, HEX employs a Unified Proprioceptive Predictor (UPP), illustrated in Figure 3 (a), which operates on part-aligned latent tokens to capture cross-part dependencies and short-term embodied dynamics. Starting from the structured part latents, we form a spatio-temporal token sequence by concatenating the current part tokens with a set of learnable future query tokens $\mathbf{N}_{t+1:t+\tau_p}$, where τ_p denotes the future prediction horizon, and then adding both temporal and part positional embeddings. This yields a shared tokenized representation over body parts and short-horizon time slots.

To better accommodate embodiment- and token-specific variations while preserving a shared predictive backbone, UPP incorporates lightweight morphology-aware MoE modules at the input and output boundaries of the predictor. As shown in Figure 3 (a), after flattening the part-time token grid into a sequence, each spatio-temporal token is routed by a learned top-k gate to a small set of experts, while a shared expert branch provides a common transformation across all tokens. This token-wise routing allows different body parts and temporal slots to adapt to different experts according to their local dynamics and embodiment-specific statistics. The routed expert outputs are aggregated using the corresponding routing weights, and combined with the shared expert output



(a) Unified Proprio Predictor (b) Action Expert

Fig. 3. **Left and middle: Unified Proprioceptive Predictor (UPP).** Morphology-based proprioceptive states are first mapped into canonical body-part tokens and augmented with learnable future query tokens. These spatio-temporal tokens are processed by a shared transformer backbone sandwiched by morphology-aware MoE adaptation modules, yielding future proprioceptive latents $\mathbf{H}^p$. The middle panel details the morphology-aware MoE, where flattened part-time tokens are routed by a learned top-k gate to a set of routed experts, while a shared expert provides a common transformation across all tokens. This design enables token-wise specialization for embodiment- and part-dependent variations while preserving reusable dynamics across embodiments. **Right: Action Expert.** Noisy action tokens are encoded and conditioned on both visual-language features $\mathbf{H}^{VL}$ and predicted proprioceptive features $\mathbf{H}^p$ through dual cross-attention. A learned gate adaptively injects the state branch on top of the visual-language branch, followed by self-attention and feed-forward refinement. The resulting denoised features are decoded into high-level actions for arm and hand control and for downstream whole-body execution.

to maintain a stable common transformation. In this way, the routed experts capture embodiment- and part-specific variations, whereas the shared expert preserves reusable dynamics across embodiments. As a result, the MoE modules act as lightweight adaptation layers around the shared transformer backbone, enabling token-level specialization without sacrificing a unified latent dynamics model.

Between the two MoE adaptation modules, a shared transformer backbone models embodiment-agnostic temporal dynamics in the latent space. The backbone operates over the full part-time token sequence and uses interleaved self-attention and visual-language-conditioned attention to model both intra-state dependencies and task-relevant contextual dynamics. Conditioned on the current proprioceptive latent $\mathbf{P}_t$, the current language and visual features $\mathbf{L}'_t$ and $\mathbf{V}'_t$, and the visual-language history cache $\mathcal{M}_t^{\mathrm{vl}}$, UPP predicts future proprioceptive latents over a horizon τ_p:

$$\mathbf{P}'_{t+1:t+\tau_p} = f_{\mathrm{upp}}\left(\mathbf{P}_t, \mathbf{L}'_t, \mathbf{V}'_t, \mathcal{M}_t^{\mathrm{vl}}\right). \tag{4}$$

The predicted future latent states capture short-horizon evolution of the whole-body state, including coordinated changes across body parts, and provide future-oriented proprioceptive cues for downstream action generation.

D. Action Expert with Adaptive Fusion

As shown in Figure 3 (b), the action expert generates high-level actions by iteratively denoising action tokens with noises under dual conditioning from visual-language and proprioceptive features. Unlike direct fusion between the two modalities, our action expert uses the evolving action representation itself as the query, and conditions it on both the VLM outputs and the UPP outputs through two parallel cross-attention branches. This design allows action generation to be guided jointly by motion-level vision-semantic context and future-oriented proprioceptive dynamics.

Let $\mathbf{X}_t$ denote the current action hidden states at diffusion step t, let $\mathbf{H}_t^{vl} = [\mathbf{L}'_t, \mathbf{V}'_t, \mathcal{M}_t]$ denote the full set of VLM features from the last layer, and let $\mathbf{H}_t^p = \mathbf{P}_{t:t+\tau_p}$ denote the proprioceptive features produced by UPP. AE first normalizes the current action states,

$$\widehat{\mathbf{X}}_t = \mathrm{Norm}(\mathbf{X}_t), \tag{5}$$

and then applies two cross-attention operations in parallel:

$$\begin{aligned} \mathbf{H}_t^{vl} &= \mathrm{Attn}_{vl}\left(\widehat{\mathbf{X}}_t, \mathbf{H}^{vl}\right), \\ \mathbf{H}_t^{p} &= \mathrm{Attn}_{p}\left(\widehat{\mathbf{X}}_t, \mathbf{H}^{p}\right). \end{aligned} \tag{6}$$

where the action states serve as queries, while the visual-language and proprioceptive features serve as two conditioning memories. To combine the two conditioning branches, AE uses a gated fusion mechanism conditioned on both cross-attended features and the current normalized action states:

$$\mathbf{g}_t = \sigma\left(W_g\left[\mathbf{H}_t^{vl}; \mathbf{H}_t^{p}; \widehat{\mathbf{X}}_t\right]\right), \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid function. The fused conditioning signal is then computed as

$$\mathbf{F}_t = \mathbf{H}_t^{v} + \mathbf{g}_t \odot \mathbf{H}_t^{p}, \tag{8}$$

and injected into the action states through a residual update:

$$\mathbf{X}'_t = \mathbf{X}_t + \mathbf{F}_t. \tag{9}$$

After this dual-conditioning stage, AE further applies self-attention and a feed-forward block to refine the action representation. In this way, the visual-language branch provides semantic and motion-relevant guidance for task execution, while the proprioceptive branch injects future-oriented whole-body dynamics and whole-body coordination cues. The gate adaptively modulates the contribution of the state branch during denoising, enabling the model to generate high-level actions that directly control the arms and hands while remaining consistent with downstream whole-body execution.

E. Cross-Embodiment Training

We pretrain HEX on trajectory datasets collected from multiple humanoid embodiments, covering diverse kinematics, dynamics, and embodiment-specific state-action spaces. Training jointly optimizes an action-generation objective for the action expert and an auxiliary future-state prediction objective for the Unified Proprioceptive Predictor.

For action generation, we adopt a flow-matching objective. Given a clean future action trajectory $\mathbf{A}_{t:t+\tau_a}$ and Gaussian noise $\mathbf{N}$ of the same shape, we construct a noisy action trajectory

$$\widetilde{\mathbf{A}}_{t:t+\tau_a} = (1-\lambda)\mathbf{N} + \lambda \mathbf{A}_{t:t+\tau_a}, \qquad \lambda \sim \mathcal{U}(0,1), \quad (10)$$

and define the corresponding velocity target as

$$\mathbf{Vel}_{t:t+\tau_a} = \mathbf{A}_{t:t+\tau_a} - \mathbf{N}. \quad (11)$$

Let $\mathbf{Z}_t^a$ denote the final hidden representation produced by the Action Expert, and let $\mathbf{P}_{t+1:t+\tau_p}$ denote the future proprioceptive latents predicted by UPP. The action decoder $D_a(\cdot)$ maps $\mathbf{Z}_t^a$ to velocity predictions, while the state decoder $D_s(\cdot)$ maps $\mathbf{P}_{t+1:t+\tau_p}$ to future proprioceptive states. Let $\mathbf{s}_{t+1:t+\tau_p}$ denote the corresponding ground-truth future proprioceptive trajectory. We then define the training objectives as

$$\begin{aligned} \mathcal{L}_a &= \|D_a(\mathbf{Z}_t^a) - \mathbf{Vel}_{t:t+\tau_a}\|_2^2, \\ \mathcal{L}_s &= \left\| D_s(\mathbf{P}'_{t+1:t+\tau_p}) - \mathbf{s}_{t+1:t+\tau_p} \right\|_2^2, \\ \mathcal{L} &= \mathcal{L}_a + \alpha \mathcal{L}_s, \end{aligned} \quad (12)$$

where $\mathcal{L}_a$ supervises high-level action denoising under dual conditioning from visual-language and proprioceptive features, while $\mathcal{L}_s$ encourages UPP to model short-horizon proprioceptive evolution in the shared latent space. Because both objectives are defined over the shared body-part-aligned representation, the same training formulation naturally extends across heterogeneous humanoid embodiments. In practice, we optionally adopt a staged schedule for optimization stability: we first warm up UPP using $\mathcal{L}_s$, and then jointly optimize UPP and the Action Expert under the combined objective.

IV. Experiments

We conduct extensive experiments on real-world humanoid whole-body manipulation to evaluate the effectiveness, generalization ability, and practical behavior of HEX. In particular, we aim to answer the following four questions:

- **RQ1:** To what extent does HEX improve performance over strong state-of-the-art baselines on real-world humanoid whole-body manipulation, particularly in seen and long-horizon settings? (Section IV-B)
- **RQ2:** How effectively does HEX generalize under unseen scene variations? (Section IV-C)
- **RQ3:** What is the effect of each major component in HEX on the overall performance? (Section IV-D)
- **RQ4:** What insights can be drawn from HEX regarding expert routing behavior, inference efficiency, and failure modes? (Section IV-E)

A. Experiment Setup

Hardware and data collection. As shown in Figure 4, we adopt a modular teleoperation pipeline for data collection. Head motion is controlled by computer-issued commands that regulate upward and downward pitch under a fixed data collection protocol. Arm and hand motions are teleoperated through an isomorphic arm–hand interface [45], while waist and leg motions are controlled using a handheld joystick. To evaluate cross-embodiment capability, we use two humanoid platforms, Tienkung 2.0 and Tienkung 3.0. Our data collection procedure follows RoboMIND [46], [47].

Baselines. To ensure a fair comparison of high-level VLA policies, we use the same RL-based low-level controller for balance control across all methods, thereby isolating the contribution of the high-level policy. All models are provided with the same input information, while the use of state inputs follows each model's original setting. We compare HEX with the following IL and VLA baselines. Unless otherwise specified, all remaining hyperparameters follow the original implementations.

ACT [15] is a small-scale vision–action model that combines a Transformer encoder–decoder with action chunking. We set the action horizon to 200 and train it for 35k steps on 8 NVIDIA A100 GPUs with a per-GPU batch size of 64.

SwitchVLA [16] is a medium-scale VLA framework for execution-aware task switching under changing instructions. We set the action horizon to 200 and train it for 100 epochs on 8 NVIDIA A100 GPUs with a per-GPU batch size of 200.

GR00T N1.5 [13] is a large-scale humanoid VLA model trained on both real-world teleoperation and large-scale simulated data. It predicts action sequences from the final-layer VLM features. We set the action horizon to 100 and train it for 50k steps on a single NVIDIA A100 GPU with a batch size of 64.

$\pi_{0.5}$ [11] is a large-scale general-purpose robot foundation model in which the VLM and action expert share the same attention backbone. We set the action horizon to 100 and train it for 50k steps on 8 NVIDIA A100 GPUs with a per-GPU batch size of 8.

Pretraining Datasets. As shown in Figure 1 (a), our training corpus comprises over 12M frames collected from seven humanoid embodiments across four data sources. First, our in-house HEX dataset contains approximately 4M frames from three embodiments: the legged humanoids Tienkung 2.0 and Tienkung 3.0, and the wheeled humanoid Tienyi.

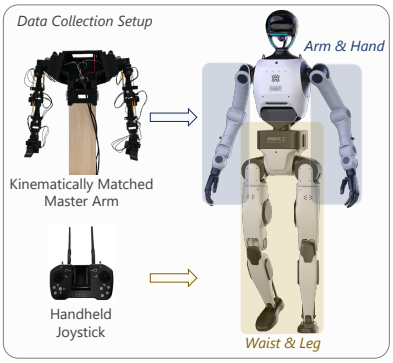


Fig. 4. Real-Robot teleoperation data collection Setup.

Owing to differences in data collection protocols and system versions, the state and action definitions are not fully consistent across embodiments. For example, the state of Tienkung 2.0 includes upper- and lower-body proprioceptive signals, while Tienkung 3.0 may additionally incorporate IMU measurements and hand tactile signals. Second, the Humanoid Everyday dataset [48] provides approximately 3.4M frames from the legged humanoids Unitree G1 and H1. Its state representation includes both upper- and lower-body information, whereas the action space contains only upper-body actions. Third, AgiBot World Colosseo [49] contributes 3.8M frames from a wheeled AgiBot humanoid platform. We use its G1-retargeted version [7], in which the original actions are transformed into a format executable by legged humanoids. Finally, we include 2.3M frames from the Leju legged humanoid subset of RoboCOIN [50]. Although these datasets differ substantially in embodiment, state composition, and action parameterization, they can all be leveraged for pretraining within our cross-embodiment architecture.

Implementation Details. HEX is built on the vision-language model Qwen3-VL-2B-Instruct [51]. The UPP is a 4-layer transformer with hidden size 768, forecasting a 50-step future state horizon. To model embodiment-specific dynamics, we employ a MoE module [52] with 16 routed experts and 2 shared experts, using top-1 softmax routing and an auxiliary load-balancing loss with weight 0.01. The action head is a 16-layer DiT-B with hidden size 1024, which predicts 100-step action chunks conditioned on visual–language features and predicted future states. During pretraining, HEX is trained for 200k steps with a per-device batch size of 16 and an action chunk size of 100, requiring approximately 1K A100 GPU hours. Optimization is performed using AdamW, with learning rates of 1.0×10^{-5} for the VLM and 2.0×10^{-5} for both the UPP and action modules. We adopt a cosine learning rate schedule with a minimum learning rate of 10^{-6}, using 5k warmup steps for the main model and 2k warmup steps for the UPP model. For fine-tuning, each task is trained for 20k steps using AdamW. The learning rate is set to 1.0×10^{-5} for the Qwen-VL interface and 4.0×10^{-5} for both the UPP and action modules. The warmup steps for the UPP model are reduced to 1k. During inference, we further apply linear interpolation only to the predicted arm and hand actions to improve motion smoothness.

B. Comparing with SoTA

1) Seen Scenarios: Seen scenarios refer to evaluation settings where the test environments closely match those in the training data. This setting primarily assesses the ability of VLA models to reproduce demonstrated trajectories from observations. All methods are evaluated over 12 trials.

Post-training Datasets. We collect seven real-robot tasks in total, including four on Tienkung 2.0 and three on Tienkung 3.0. These tasks cover whole-body control involving the arms and hands, waist, and legs, as well as multiple scenarios requiring timely human–robot interaction, enabling evaluation of both task success rate and response speed across different VLA models. The seven tasks are as follows.

Task 1: Mirror the human's pose. The robot observes a person standing in front of it and imitates the posed gestures, including "V," "L," and "A," in real time. We collect 108 trajectories for training.

Task 2: Pour liquor while following human order. A liquor bottle and three cups are placed on the table in front of the robot. The robot pours liquid into the cup indicated by the human's pointing. We collect 100 trajectories for training.

Task 3: Human assistant. The robot carries a box and follows a human collaborator to assist with organizing objects across two tables. We collect 98 trajectories for training.

Task 4: Walking while avoiding obstacles. As the robot walks forward, it must stop promptly when a person or cart passes through its path, and then resume walking once the path is clear. We collect 100 trajectories for training.

Task 5: Kneel and manipulate the objects. The robot kneels down to pick up blocks and place them into a box. We collect 300 trajectories for training.

Task 6: Tidy Table. The robot clears the tabletop by sorting scattered objects into a box and disposing of paper scraps into a trash bin. We collect 100 trajectories for training.

Task 7: Bring box and pack all objects. The robot first retrieves a box and then packs all target objects into it. We collect 100 trajectories for training.

Results. As shown in Table I, in in-distribution settings, despite their much smaller parameter scales, ACT and SwitchVLA remain competitive with several-billion-parameter models, suggesting that small and medium-sized models are already sufficient to fit seen trajectories effectively. In particular, ACT exhibits especially strong trajectory-fitting ability and produces remarkably smooth hand motions, especially on Tasks 1, 2, 6, and 7, with almost no observable latency. Among the large-scale models, $\pi_{0.5}$ shows slightly better motion smoothness and higher success rates than GR00T N1.5, while HEX achieves the best overall performance. Compared with ACT and SwitchVLA, however, these larger models tend to produce less smooth motions in highly reactive in-distribution tasks, such as Tasks 1 and 2, and often exhibit mild lag or stuttering during execution. Within this in-distribution setting, the advantage of HEX lies in its stronger balance between task success and motion quality among large-scale models. We attribute this improvement to the explicit future-state conditioning in HEX, which provides additional dynamic cues beyond current visual observations.

TABLE I
TASK SUCCESS RATES (%) OF DIFFERENT METHODS ON TIENKUNG 3.0 AND TIENKUNG 2.0 TASKS. THE ICONS DENOTE THE MAIN BODY PARTS INVOLVED IN CONTROL FOR EACH TASK: ARM (🦾), HAND (🖐), WAIST (🦵), AND LEG (🚶).

Method	Para.	Avg (%)	Tienkung 3.0 (🦾 🖐 🦵)	Tienkung 3.0 (🦾 🖐 🦵)	Tienkung 3.0 (🦾 🦵 🚶)
			Kneel and manipulate the objects	Tidy Table	Bring box and pack all objects
ACT [15]	80M	57.1	83.3	8.3	8.3
SwitchVLA [16]	0.3B	40.5	0.0	8.3	0.0
GR00T N1.5 [13]	3B	70.2	100.0	41.7	33.3
$\pi_{0.5}$ [11]	3.3B	71.8	100.0	35.7	25.0
HEX	2.4B	**79.8**	100.0	41.7	41.7
Method	Para.	Tienkung 2.0 (🦾 🦵)	Tienkung 2.0 (🦾 🖐 🦵)	Tienkung 2.0 (🦾 🦵 🚶)	Tienkung 2.0 (🦾 🦵 🚶)
		Mirror the human's pose	Pour liquor while following the human order	Human assistant	Walking while avoiding obstacles
ACT [15]	80M	83.3	83.3	66.7	66.7
SwitchVLA [16]	0.3B	100.0	41.7	58.3	75.0
GR00T N1.5 [13]	3B	83.3	66.7	66.7	100.0
$\pi_{0.5}$ [11]	3.3B	83.3	91.7	75.0	91.7
HEX	2.4B	100.0	91.7	83.3	100.0

TABLE II
TASK SUCCESS RATES (%) ON THE **LONG-HORIZON BOX CONVEY TASK** IN TIENKUNG 2.0.

Method	Tienkung 2.0 (🦾 🖐 🦵 🚶)			
	Grasp Box	Turn Around	Walk to Table	Place Box
ACT [15]	80.0	80.0	46.7	26.7
SwitchVLA [16]	73.3	60.0	33.3	13.3
GR00T N1.5 [13]	73.3	66.7	40.0	20.0
$\pi_{0.5}$ [11]	100.0	100.0	73.3	40.0
HEX	100.0	100.0	73.3	53.3

2) Long-Horizon Scenarios: Long-horizon tasks are composed of multiple subtasks, where different stages require different body parts, such as the waist and hands in some stages and the legs in others. This heterogeneous composition increases task complexity and gives rise to more pronounced cascading errors across stages.

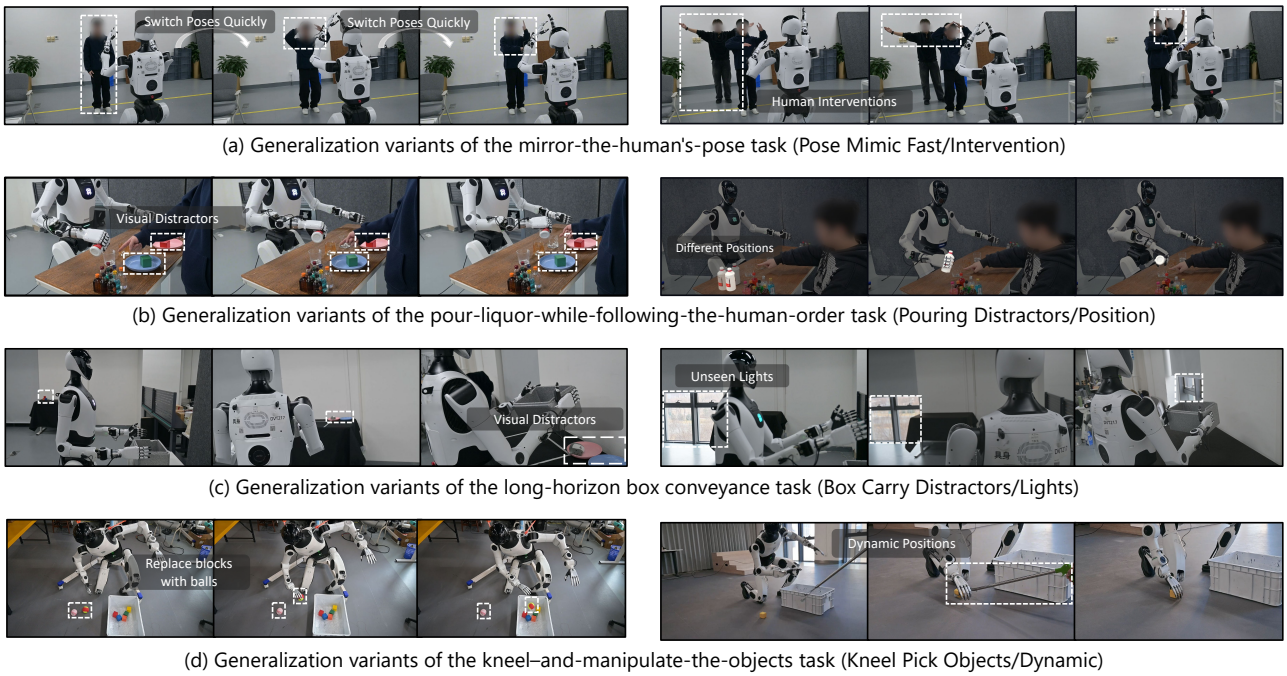
Post-training Datasets. We collect a long-horizon box conveyance task consisting of four stages: squating down to grasp the box, turning toward the table, moving to the table and stopping in front of it, and squating down again to place the box. In total, we collect 56 trajectories for training. Each task is evaluated over 15 trials.

Results. As shown in Table II, HEX achieves the best performance across all stages of the long-horizon box convey task, outperforming the baselines by a clear margin. Notably, on the final *Place Box* stage, HEX surpasses the strongest baseline by around 15%, indicating its superior ability to sustain stable execution and reduce cascading errors over long-horizon whole-body manipulation.

C. Generalization Study

Evalation Tasks. As shown in Figure 5, we evaluate generalization on four tasks from the seen-scenario setting, including three standard tasks (Tasks 1, 2, and 5) and one long-horizon task. For *Pose Mimic*, we consider Pose Mimic Fast, which increases the speed of human pose switching, and Pose Mimic Intervention, where an additional person in the background continuously performs distracting poses. A total of 18 trials are conducted, including 5 trials each for the V-, L-, and A-shaped poses, and 3 trials for the return-hand pose. For *Pouring*, we evaluate Pouring Distractors by adding irrelevant objects, and Pouring Position by changing the bottle location. In the distractor setting, three cups are tested with 5 trials each. In the position-variation setting, the bottle is placed at 9 different locations, with one trial for each cup at each location. For *Kneel Pick*, we evaluate Kneel Pick Dynamic, where object positions are changed during execution, and Kneel Pick Objects, where the original blocks are replaced with unseen balls. Each variant is evaluated with 15 grasp-and-place trials. For *Box Carry*, we evaluate Box Carry Distractors by introducing additional surrounding objects, and Box Carry Lights by changing the lighting condition. Each variant is evaluated over 15 trials.

Results. Figure 6 summarizes the results on eight generalization variants across four seen tasks. HEX achieves the best overall average success rate of 61.8%, substantially outperforming $\pi_{0.5}$ (44.3%), GR00T N1.5 (41.0%), and SwitchVLA (22.4%). Overall, HEX performs best on nearly all variants. In *Pose Mimic*, HEX matches the best result under fast switching (100%) and achieves the highest success rate under human intervention (85.7%). Notably, under human intervention, all methods except HEX fail to maintain the "L" hand gesture after being distracted by a background person. In addition, GR00T N1.5 and $\pi_{0.5}$ often produce inaccurate `L`-shaped poses. In contrast, HEX remains substantially more robust to such interference. In *Pouring*, HEX shows the clearest advantage, improving from 0% for all baselines to 53.3% under visual distractors, and reaching 55.5% under bottle-position changes. After distractor objects are introduced, all baselines tend to start pouring before receiving a human instruction and then remain fixed at one location. We conjecture that these models mistakenly treat the red plate as the human



(a) Generalization variants of the mirror-the-human's-pose task (Pose Mimic Fast/Intervention)

(b) Generalization variants of the pour-liquor-while-following-the-human-order task (Pouring Distractors/Position)

(c) Generalization variants of the long-horizon box conveyance task (Box Carry Distractors/Lights)

(d) Generalization variants of the kneel–and-manipulate-the-objects task (Kneel Pick Objects/Dynamic)

Fig. 5. **Generalization tasks.** Two distribution-shift variants for each of four seen tasks: *Pose Mimic*, *Pouring*, *Box Carry*, and *Kneel Pick*.

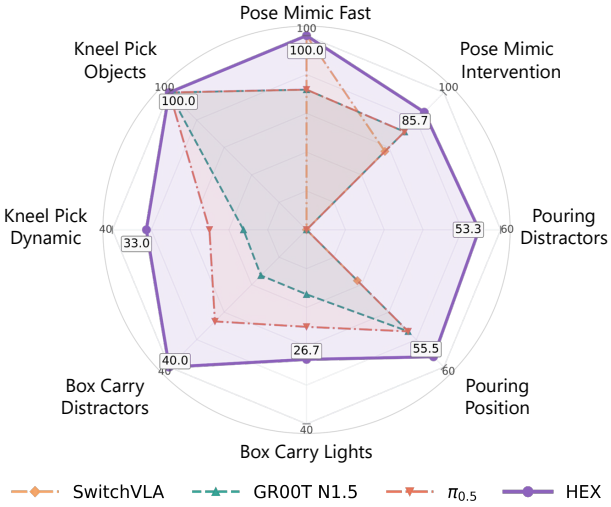


Fig. 6. Generalization results across unseen task variants.

hand, whereas HEX does not exhibit this failure mode. In *Box Carry*, HEX also achieves the best results under both unseen lighting (26.7%) and unseen surrounding objects (40.0%). For *Kneel Pick*, HEX attains 33.0% under dynamic position changes and 100% under unseen objects. These results indicate that HEX generalizes more robustly under diverse distribution shifts, including faster human motion, human interference, visual distractors, object-position changes, lighting variation, and dynamic scene changes.

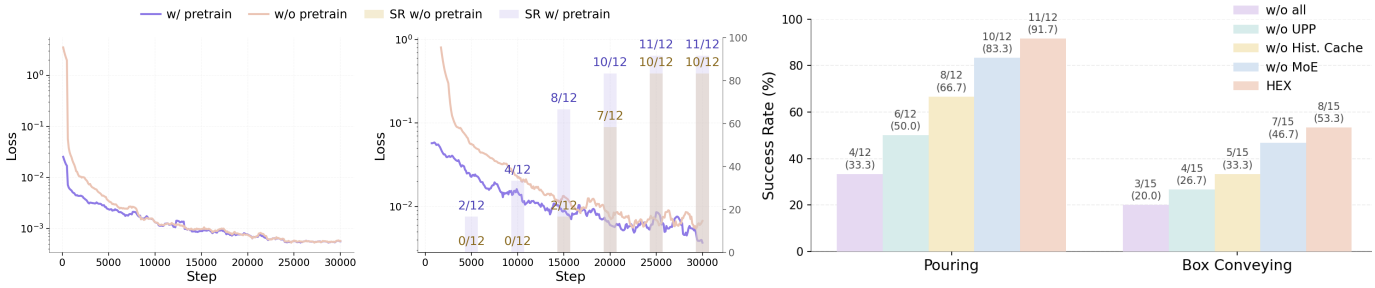
D. Ablation Study

Ablation on Pretraining. Figure 7 (a) shows that pretraining mainly improves optimization efficiency rather than the final converged performance in our single-task setting. Specifically, the pretrained model exhibits clearly lower state and action losses in the early stage of training, indicating better initialization and faster fitting. As training proceeds, the gap in state loss becomes marginal after around 10k steps, while the pretrained model still maintains a generally lower action loss overall. This optimization advantage is also reflected in early-stage task success: at 5k/10k/15k/20k steps, the pretrained model achieves 2/12, 4/12, 8/12, and 10/12 success, compared with 0/12, 0/12, 2/12, and 7/12 without pretraining. However, the difference becomes small at later stages, with both models reaching similar final success rates (11/12 vs. 10/12). These results suggest that, under the single-task setting, the primary benefit of pretraining is faster fitting and improved sample efficiency, rather than a substantial gain in final performance.

Ablation on Model Components. Figure 7 (b) evaluates the contributions of the VLM history cache, the UPP, and the MoE design within UPP. Performance improves consistently as these components are progressively introduced. On *Pouring*, success increases from 4/12 without all components to 6/12 without UPP, 8/12 without the history cache, 10/12 without MoE, and 11/12 for the full HEX. A similar trend is observed on *Box Conveying*, where performance improves from 3/15 to 4/15, 5/15, 7/15, and finally 8/15. Among the evaluated components, the UPP has the strongest effect, as its removal results in the largest performance degradation on both tasks.

E. Other Anayses

Failure Analysis. Figure 8 shows that different methods fail not only at different rates, but also in different stages. In the two seen-scenario tasks, failures are relatively concentrated in a small number of key sub-stages, mainly related to object grasping, placement, and multi-object handling. In contrast, the long-horizon box conveyance task exhibits more distributed failures across grasping, turning, locomotion, and



(a) Effect of pretraining (left: state loss; right: action loss and success rate)

(b) Effect of key model components

Fig. 7. **Ablation study.** (a) Effect of pretraining. Left: state loss. Right: action loss together with success-rate comparisons at different training stages. Pretraining improves optimization. (b) Effect of key model components. Performance improves consistently as the history cache, UPP, and MoE design are progressively introduced, and the full HEX achieves the best success rates on both *Pouring* and *Box Conveying*.

(a) Failure mode of table tidy task

(b) Failure mode of bring-box-and-pack-all-objects task

(c) Failure mode of long-horizon box conveyance task

Fig. 8. Failure analysis across different methods and tasks. Each Sankey diagram shows how failed trials are distributed across task stages and fine-grained error types, with flow width proportional to the number of failures.

final placement, indicating that longer action chains amplify error accumulation and cross-stage dependency. Across tasks, HEX generally yields fewer failures and a more concentrated failure distribution than the baselines. This suggests that its advantage is not only higher overall success, but also improved robustness to error propagation across sequential stages, especially in long-horizon execution.

MoE Routing Pattern. Figure 9 reveals a clear difference between the two routing locations. Before the transformer blocks, expert assignments are largely stable over time and vary little across subtask transitions, suggesting that the routing mainly encodes persistent body-part-specific specialization. After the transformer blocks, the routing becomes markedly more phase-dependent, with major switches aligning well with semantic subtask boundaries. This effect is particularly evident in the leg channels: lower-index experts dominate during static support phases, whereas higher-index experts are selected during turning and forward locomotion. These results suggest that placing the MoE after the transformer blocks enables expert selection to better reflect the evolving control demands of long-horizon whole-body manipulation.

Latency. Figure 10 compares the latency–accuracy trade-off of different methods on an RTX 4090. ACT achieves the lowest latency, but with a substantial drop in success rate. Among the large-scale baselines, GR00T N1.5 attains a relatively favorable latency–accuracy trade-off, achieving competitive performance at lower latency than both $\pi_{0.5}$ and HEX. HEX nevertheless achieves the highest success rate overall (79.8%) with 73.34 ms latency, outperforming all baselines in task success while remaining faster than $\pi_{0.5}$. Overall, these results show that HEX provides the strongest effectiveness under a practical inference budget.

V. Conclusion

We presented **HEX**, a framework for humanoid whole-body manipulation that addresses a key limitation of existing VLA-style approaches: they often do not explicitly model how different body parts interact under shared balance and posture. HEX tackles this problem through a humanoid-aligned universal state representation, predictive modeling of

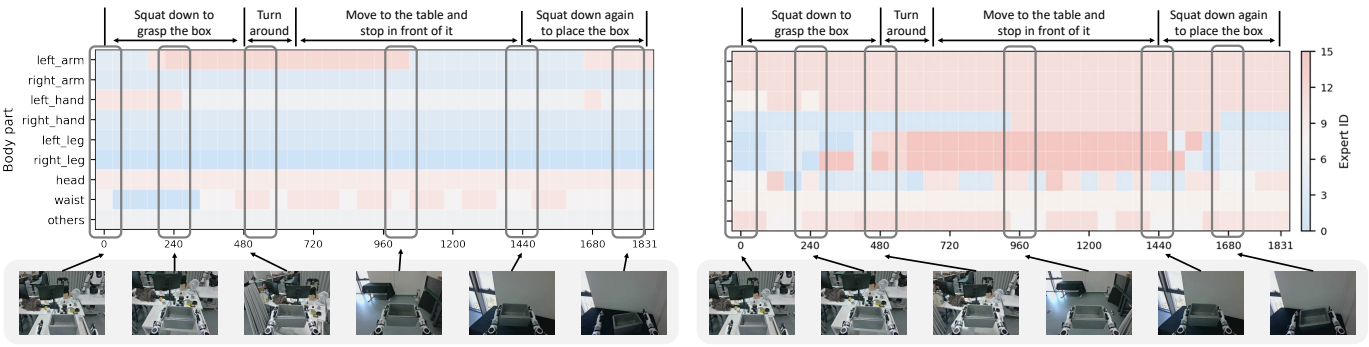


Fig. 9. Comparison of MoE routing patterns before and after the transformer blocks during a long-horizon box conveyance task. Left: MoE before the transformer blocks. Right: MoE after the transformer blocks. The heatmaps show the selected expert ID for each body part over time, together with representative frames and annotated subtask boundaries. Compared with the largely static routing before the transformer blocks, the routing after the transformer blocks shows clearer phase-dependent switching, suggesting stronger state-dependent control specialization.

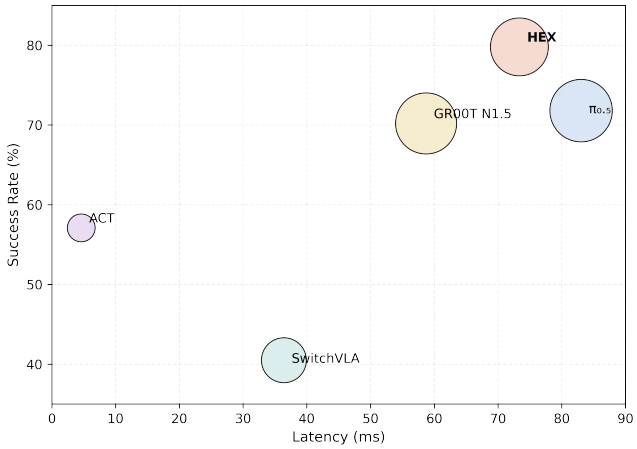


Fig. 10. Latency–accuracy comparison on a single NVIDIA RTX 4090 GPU, where bubble size indicates the number of model parameters.

whole-body proprioceptive dynamics, and adaptive fusion of visual-language context with future state evolution. This leads to more coherent and stable whole-body action generation. Extensive experiments on real-world humanoid manipulation tasks show that HEX achieves superior performance over strong baselines, particularly in fast-reaction and long-horizon settings where coordinated whole-body behavior is essential. Overall, our results highlight the importance of explicitly modeling structured body-part interaction for general and scalable humanoid manipulation.

REFERENCES

[1] J. Nakanishi, J. Morimoto, G. Endo, G. Cheng, S. Schaal, and M. Kawato, "Learning from demonstration and adaptation of biped locomotion," *Robotics and autonomous systems*, vol. 47, no. 2-3, pp. 79–91, 2004.

[2] X. B. Peng, P. Abbeel, S. Levine, and M. Van de Panne, "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills," *ACM Transactions On Graphics (TOG)*, vol. 37, no. 4, pp. 1–14, 2018.

[3] I. Radosavovic, T. Xiao, B. Zhang, T. Darrell, J. Malik, and K. Sreenath, "Real-world humanoid locomotion with reinforcement learning," *Science Robotics*, vol. 9, no. 89, p. eadi9579, 2024.

[4] J. Li, Y. Zhu, Y. Xie, Z. Jiang, M. Seo, G. Pavlakos, and Y. Zhu, "Okami: Teaching humanoid robots manipulation skills through single video imitation," in *Conference on Robot Learning*. PMLR, 2025, pp. 299–317.

[5] R. Yang, Q. Yu, Y. Wu, R. Yan, B. Li, A.-C. Cheng, X. Zou, Y. Fang, X. Cheng, R.-Z. Qiu *et al.*, "Egovla: Learning vision-language-action models from egocentric human videos," *arXiv preprint arXiv:2507.12440*, 2025.

[6] Z. Xie, J. Tseng, S. Starke, M. Van De Panne, and C. K. Liu, "Hierarchical planning and control for box loco-manipulation," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 6, no. 3, pp. 1–18, 2023.

[7] J. Liu, P. Ding, Q. Zhou, Y. Wu, D. Huang, Z. Peng, W. Xiao, W. Zhang, L. Yang, C. Lu *et al.*, "Trajbooster: Boosting humanoid whole-body manipulation via trajectory-centric learning," in *2026 IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

[8] H. Jiang, J. Chen, Q. Bu, L. Chen, M. Shi, Y. Zhang, D. Li, C. Suo, C. Wang, Z. Peng *et al.*, "Wholebodyvla: Towards unified latent vla for whole-body loco-manipulation control," in *The Fourteenth International Conference on Learning Representations*, 2026.

[9] S. Chen, Y. Ye, Z.-a. Cao, P. Xu, J. Lew, and K. Liu, "Hand-eye autonomous delivery: Learning humanoid navigation, locomotion and reaching," in *Conference on Robot Learning*. PMLR, 2025, pp. 4058–4073.

[10] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong *et al.*, "Openvla: An open-source vision-language-action model," in *Conference on Robot Learning*. PMLR, 2025, pp. 2679–2713.

[11] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai *et al.*, "$\pi_{0.5}$: A vision-language-action model with open-world generalization," in *Conference on Robot Learning*, 2025.

[12] C. Cui, P. Ding, W. Song, S. Bai, X. Tong, Z. Ge, R. Suo, W. Zhou, Y. Liu, B. Jia *et al.*, "Openhelix: A short survey, empirical analysis, and open-source dual-system vla model for robotic manipulation," *arXiv preprint arXiv:2505.03912*, 2025.

[13] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang *et al.*, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[14] S. Wei, H. Jing, B. Li, Z. Zhao, J. Mao, Z. Ni, S. He, J. Liu, X. Liu, K. Kang, S. Zang, W. Yuan, M. Pavone, D. Huang, and Y. Wang, "ψ_0: An open foundation model towards universal humanoid loco-manipulation," 2026. [Online]. Available: https://arxiv.org/abs/2603.12263

[15] T. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *Robotics: Science and Systems XIX*, 2023.

[16] M. Li, Z. Zhao, Z. Che, F. Liao, K. Wu, Z. Xu, P. Ren, Z. Jin, N. Liu, and J. Tang, "Switchvla: Execution-aware task switching for vision-language-action models," *arXiv preprint arXiv:2506.03574*, 2025.

[17] Y. Ze, Z. Chen, J. P. Araujo, Z.-a. Cao, X. B. Peng, J. Wu, and K. Liu, "Twist: Teleoperated whole-body imitation system," in *Conference on Robot Learning*. PMLR, 2025, pp. 2143–2154.

[18] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa, "Amp: Adversarial motion priors for stylized physics-based character control," *ACM Transactions on Graphics (ToG)*, vol. 40, no. 4, pp. 1–20, 2021.

[19] J. Hu, P. Stone, and R. Martín-Martín, "Slac: Simulation-pretrained latent action space for whole-body real-world rl," in *Conference on Robot Learning*. PMLR, 2025, pp. 2966–2982.

[20] Y. Zhang, Y. Yuan, P. Gurunath, I. Gupta, S. Omidshafiei, A.-a. Aghamohammadi, M. Vazquez-Chanlatte, L. Pedersen, T. He, and G. Shi, "Falcon: Learning force-adaptive humanoid loco-manipulation," *8th Annual Learning for Dynamics\& Control Conference*, 2026.

[21] W. Xie, J. Han, J. Zheng, H. Li, X. Liu, J. Shi, W. Zhang, C. Bai, and X. Li, "Kungfubot: Physics-based humanoid whole-body control for learning highly-dynamic skills," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[22] Z. Luo, Y. Yuan, T. Wang, C. Li, S. Chen, F. Castaneda, Z.-A. Cao, J. Li, D. Minor, Q. Ben *et al.*, "Sonic: Supersizing motion tracking for natural humanoid whole-body control," *arXiv preprint arXiv:2511.07820*, 2025.

[23] T. He, Z. Wang, H. Xue, Q. Ben, Z. Luo, W. Xiao, Y. Yuan, X. Da, F. Castañeda, S. Sastry *et al.*, "Viral: Visual sim-to-real at scale for humanoid loco-manipulation," *arXiv preprint arXiv:2511.15200*, 2025.

[24] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," in *Conference on Robot Learning*. PMLR, 2025, pp. 2828–2844.

[25] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. M. Kitani, C. Liu, and G. Shi, "Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," in *Conference on Robot Learning*. PMLR, 2025, pp. 1516–1540.

[26] R.-Z. Qiu, S. Yang, X. Cheng, C. Chawla, J. Li, T. He, G. Yan, D. J. Yoon, R. Hoque, L. Paulsen *et al.*, "Humanoid policy human policy," in *Conference on Robot Learning*. PMLR, 2025, pp. 2888–2906.

[27] Z. Chen, M. Ji, X. Cheng, X. Peng, X. B. Peng, and X. Wang, "Gmt: General motion tracking for humanoid whole-body control," *arXiv preprint arXiv:2506.14770*, 2025.

[28] C. Lu, X. Cheng, J. Li, S. Yang, M. Ji, C. Yuan, G. Yang, S. Yi, and X. Wang, "Mobile-television: Predictive motion priors for humanoid whole-body control," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 5364–5371.

[29] Q. Liao, T. E. Truong, X. Huang, Y. Gao, G. Tevet, K. Sreenath, and C. K. Liu, "Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion," *arXiv preprint arXiv:2508.08241*, 2025.

[30] Y. Ze, Z. Chen, W. Wang, T. Chen, X. He, Y. Yuan, X. B. Peng, and J. Wu, "Generalizable humanoid manipulation with 3d diffusion policies," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 2873–2880.

[31] P. Ding, J. Ma, X. Tong, B. Zou, X. Luo, Y. Fan, T. Wang, H. Lu, P. Mo, J. Liu *et al.*, "Humanoid-vla: Towards universal humanoid control with visual integration," *arXiv preprint arXiv:2502.14795*, 2025.

[32] H. Xue, X. Huang, D. Niu, Q. Liao, T. Kragerud, J. T. Gravdahl, X. B. Peng, G. Shi, T. Darrell, K. Sreenath *et al.*, "Leverb: Humanoid whole-body control with latent vision-language instruction," *arXiv preprint arXiv:2506.13751*, 2025.

[33] L. Wang, X. Chen, J. Zhao, and K. He, "Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers," in *Advances in neural information processing systems*, vol. 37, 2024, pp. 124 420–124 450.

[34] J. Yang, C. Glossop, A. Bhorkar, D. Shah, Q. Vuong, C. Finn, D. Sadigh, and S. Levine, "Pushing the limits of cross-embodiment learning for manipulation and navigation," in *Robotics: Science and Systems*, 2024.

[35] R. Doshi, H. R. Walke, O. Mees, S. Dasari, and S. Levine, "Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation," in *Conference on Robot Learning*. PMLR, 2025, pp. 496–512.

[36] J. Mao, S. Zhao, S. Song, T. Shi, J. Ye, M. Zhang, H. Geng, J. Malik, V. Guizilini, and Y. Wang, "Learning from massive human videos for universal humanoid pose control," in *International Conference on Humanoid Robots*, 2025.

[37] H. Weng, Y. Li, N. Sobanbabu, Z. Wang, Z. Luo, T. He, D. Ramanan, and G. Shi, "Hdmi: Learning interactive humanoid whole-body control from human videos," *arXiv preprint arXiv:2509.16757*, 2025.

[38] M. Shi, S. Peng, J. Chen, H. Jiang, Y. Li, D. Huang, P. Luo, H. Li, and L. Chen, "Egohumanoid: Unlocking in-the-wild loco-manipulation with robot-free egocentric demonstration," *arXiv preprint arXiv:2602.10106*, 2026.

[39] H. Yang, J. Bao, Y. Xin, H. Song, Y. Tian, B. Zhao, D. Wang, and X. Li, "Zerowbc: Learning natural visuomotor humanoid control directly from human egocentric video," *arXiv preprint arXiv:2603.09170*, 2026.

[40] Y. Lin, M. Liu, Y. Xue, M. Zhou, Y. Yu, J. Pang, and W. Zhang, "H-zero: Cross-humanoid locomotion pretraining enables few-shot novel embodiment transfer," *arXiv preprint arXiv:2512.00971*, 2025.

[41] R. Punamiya, D. Patel, P. Aphiwetsa, P. Kuppili, L. Y. Zhu, S. Kareer, J. Hoffman, and D. Xu, "Egobridge: Domain adaptation for generalizable imitation from egocentric human data," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[42] Y. Xue, Y. Lin, W. Dong, Y. Tang, J. Wang, J. Pang, M. Zhou, M. Liu, and W. Zhang, "Scalable and general whole-body control for cross-humanoid locomotion," *arXiv preprint arXiv:2602.05791*, 2026.

[43] Q. Peng, Y. Lin, Y. Xue, J. Pang, and W. Zhang, "Embodiment-aware generalist specialist distillation for unified humanoid whole-body control," *arXiv preprint arXiv:2602.02960*, 2026.

[44] H. Luo, Y. Wang, W. Zhang, S. Zheng, Z. Xi, C. Xu, H. Xu, H. Yuan, C. Zhang, Y. Wang *et al.*, "Being-h0. 5: Scaling human-centric robot learning for cross-embodiment generalization," *arXiv preprint arXiv:2601.12993*, 2026.

[45] Z. Xu, Y. Zhao, K. Wu, N. Liu, J. Ji, Z. Che, C. H. Liu, and J. Tang, "Hacts: a human-as-copilot teleoperation system for robot learning," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 15 475–15 481.

[46] K. Wu, C. Hou, J. Liu, Z. Che, X. Ju, Z. Yang, M. Li, Y. Zhao, Z. Xu, G. Yang *et al.*, "Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation," in *Robotics: Science and Systems (RSS)*, 2025.

[47] C. Hou, K. Wu, J. Liu, Z. Che, D. Wu, F. Liao, G. Li, J. He, Q. Feng, Z. Jin *et al.*, "Robomind 2.0: A multimodal, bimanual mobile manipulation dataset for generalizable embodied intelligence," *arXiv preprint arXiv:2512.24653*, 2025.

[48] Z. Zhao, H. Jing, X. Liu, J. Mao, A. Jha, H. Yang, R. Xue, S. Zakharor, V. Guizilini, and Y. Wang, "Humanoid everyday: A comprehensive robotic dataset for open-world humanoid manipulation," *arXiv preprint arXiv:2510.08807*, 2025.

[49] Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, S. Gao, X. He, X. Hu, X. Huang *et al.*, "Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025, pp. 3549–3556.

[50] S. Wu, X. Liu, S. Xie, P. Wang, X. Li, B. Yang, Z. Li, K. Zhu, H. Wu, Y. Liu *et al.*, "Robocoin: An open-sourced bimanual robotic data collection for integrated manipulation," *arXiv preprint arXiv:2511.17441*, 2025.

[51] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge *et al.*, "Qwen3-vl technical report," *arXiv preprint arXiv:2511.21631*, 2025.

[52] Z. Du, B. Liu, Y. Liang, Y. Shen, H. Cao, X. Zheng, Z. Feng, Z. Wu, J. Yang, and Y.-G. Jiang, "Himoe-vla: Hierarchical mixture-of-experts for generalist vision-language-action policies," *arXiv preprint arXiv:2512.05693*, 2025.