Learning Humanoid Loco-Manipulation with Responsibility-Induced Specialized Experts for Vision-Language-Action Models

Hogun Kee*, Wooseok Oh*, Jooyoung Kim, Jaeyeon Jeong, Hyewoo Jung, Hyeondal Son, Hosung Lee, Minjae Kang, and Songhwai Oh

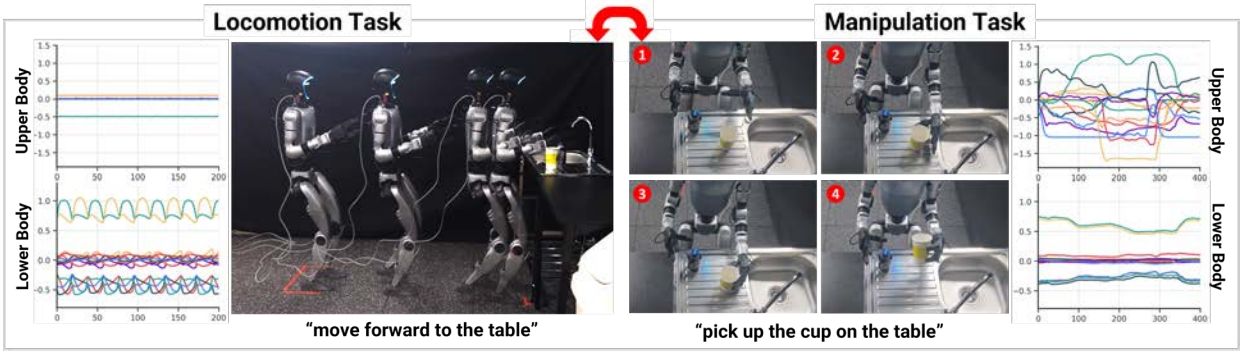

Fig. 1. Humanoid loco-manipulation demonstrations executed by RISE on Unitree G1. We illustrate representative locomotion and manipulation-dominant behaviors, together with the corresponding upper and lower-body state trajectories over time. The plots highlight largely separated whole-body behavior patterns and the need for reliable coordination between them.

***Abstract*— In this paper, we present Responsibility-Induced Specialized Experts (RISE), a responsibility-driven multi-expert Vision–Language–Action (VLA) framework that learns long-horizon humanoid loco-manipulation from segmented locomotion and manipulation-oriented demonstrations. Humanoid loco-manipulation remains challenging due to (i) heterogeneous whole-body behaviors with substantially different action distributions and (ii) the difficulty of collecting reliable long-horizon demonstrations that include behavior transitions. RISE addresses these challenges by inducing behavior-specialized experts via responsibility-driven learning, where each expert's responsibility reflects how well it reconstructs the target action. This mechanism encourages experts to specialize in different behavioral patterns and enables the policy to model heterogeneous action distributions. The resulting expert weighting dynamics allow RISE to coordinate experts over time and infer behavior transitions during execution, enabling long-horizon loco-manipulation despite being trained only on segmented data. We evaluate RISE in simulation and on a real humanoid robot, demonstrating robust loco-manipulation with frequent locomotion–manipulation transitions and improved performance over baselines trained on segmented demonstrations.**

I. Introduction

Humanoid robots have the potential to perform complex tasks in human environments by combining locomotion and manipulation capabilities. However, enabling such loco-manipulation behaviors remains challenging due to the need to coordinate diverse whole-body motions under high-dimensional control. Compared to fixed-base manipulators, humanoids introduce substantially larger action spaces due to whole-body articulation and dexterous hands, increasing the complexity of whole-body control.

Recent advances in Vision–Language–Action (VLA) models have shown promise in learning general robot policies that integrate perception, language understanding, and control [1]–[3]. These VLA-based robot policies have recently been extended from fixed-base manipulators to high-DoF humanoid platforms [4]. However, applying VLA policies to humanoid loco-manipulation remains challenging, in part because collecting large-scale long-horizon humanoid demonstrations is difficult in practice.

One practical challenge lies in collecting long-horizon demonstrations for humanoid robots. Due to floating-base dynamics and high degrees of freedom [5], long trajectories are prone to failure, often invalidating entire demonstrations. A practical alternative is to collect shorter behavioral segments, such as locomotion-oriented and manipulation-oriented behaviors obtained from separate subtasks. While such segmented datasets simplify data collection, they introduce new challenges when learning policies for long-horizon tasks that require switching between behaviors. In this work, we study how to learn long-horizon humanoid behaviors from segmented datasets.

However, learning long-horizon humanoid behaviors from segmented datasets remains challenging for two main reasons. First, humanoid tasks involve heterogeneous behavioral patterns with substantially different action distributions [6]. For example, locomotion and manipulation require distinct coordination patterns across the robot's joints (Fig. 1), making it difficult for a single policy to model them jointly. Second, segmented datasets typically lack transitions between

*Equal contribution.

H. Kee is with NVIDIA Corporation, Seoul 06164, Korea (e-mail: khogun@nvidia.com).

W. Oh, J. Kim, J. Jeong, H. Son, H. Lee, M. Kang and S. Oh are with the Department of Electrical and Computer Engineering and ASRI, Seoul National University, Seoul 08826, Korea (e-mail: {wooseok.oh, jooyoung.kim, jaeyeon.jeong, hyeondal.son, hosung.lee, minjae.kang}@rllab.snu.ac.kr, songhwai@snu.ac.kr).

(Corresponding Author : Songhwai Oh)

This work was supported in part by the Institute of Information communications Technology Planning Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00336738, Development of Complex Task Planning Technologies for Autonomous Agents, 50%) and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25547694, Embodiment-Agnostic Intelligence and Knowledge Transfer System for Autonomous Complex Task Execution in Heterogeneous Humanoids, 50%).

behavior segments, making it difficult for policies to learn when and how to switch between behaviors during long-horizon execution.

To address these challenges, we propose Responsibility-Induced Specialized Experts (RISE), a responsibility-driven multi-expert VLA framework designed to learn long-horizon humanoid behaviors from segmented datasets. By introducing multiple action experts in the VLA action head, RISE allows different experts to focus on locomotion- and manipulation-dominant action distributions while sharing the same perception-language backbone.

A key idea of RISE is to derive expert responsibilities—reflecting how well each expert reconstructs a given action—from the flow matching training objective. During training, experts that better reconstruct the target action receive higher responsibility and are updated more strongly. This responsibility-weighted learning refines expert specialization after an initial energy-guided warm-up, so that expert assignment is shaped by reconstruction performance rather than fixed task labels.

While the proposed framework is broadly applicable to segmented behavioral datasets, we study humanoid loco-manipulation tasks composed of locomotion and manipulation behaviors. As different experts specialize in different behavioral patterns, shifts in the gating weights provide a useful signal for detecting behavior transitions during inference, enabling the policy to sequence behaviors and execute long-horizon tasks despite being trained only on segmented data.

We evaluate our approach in both simulation and real-world humanoid experiments on loco-manipulation tasks. The results show that responsibility-driven expert policies can effectively leverage segmented datasets to learn humanoid loco-manipulation behaviors composed of multiple behavioral segments, improving performance on long-horizon tasks that would otherwise require expensive long-horizon demonstrations.

We summarize our contributions as follows:

- We propose Responsibility-Induced Specialized Experts (RISE), a multi-expert VLA framework for learning long-horizon humanoid behaviors from segmented datasets.
- RISE encourages each expert to specialize in a distinct behavior pattern by learning responsibilities derived from the flow-matching objective.
- We show that meta-controller gating dynamics can serve as a transition signal, enabling long-horizon humanoid loco-manipulation from segmented datasets.

II. Related Work

A. Humanoid Loco-Manipulation Frameworks

Humanoid loco-manipulation involves high-dimensional control with strong coupling between balance and manipulation. To improve training stability and data efficiency, many works adopt an upper–lower body decomposition during data collection and learning. In this paradigm, upper-body behaviors are obtained through motion capture or teleoperation, while lower-body locomotion is trained to remain robust under upper-body perturbations via reinforcement learning or trajectory tracking.

Representative works combine upper-body imitation or teleoperation with robust locomotion learning [7]–[11], allowing stable balance control to be learned independently from task-specific upper-body behaviors. Our work follows the same upper–lower decomposition in both data collection and system design.

B. Vision-Language-Action Models

Vision–Language–Action (VLA) models learn policies that map visual observations and language instructions directly to robot actions. Prior work includes RT-2 [1], π_0 [2], $\pi_{0.5}$ [3], OpenVLA [12], and UniVLA [13], which explore scalable multimodal policies for robotic control.

Recent work incorporates Mixture-of-Experts (MoE) architectures into VLA models to improve scalability and better capture heterogeneous robotic behaviors through multiple experts. ForceVLA [14] introduces force-aware experts for manipulation policies, while FedVLA [15] employs dual-gating MoE to support federated training across heterogeneous robot datasets. Our approach also adopts a multi-expert architecture, but differs in how routing supervision is obtained and how it is used. Rather than relying only on learned gating or task-specific expert design, RISE derives responsibilities from per-expert flow-matching errors, uses them to supervise the meta-controller, and exploits the resulting gating dynamics as a transition signal for sequencing segmented humanoid behaviors.

MoDE [16] introduces a mixture of expert denoisers within diffusion transformer policies for multitask learning. Despite structural similarity, our approach focuses on inducing behavioral specialization and leveraging learned gating dynamics to detect transitions between behaviors. In particular, we derive expert responsibilities from reconstruction performance and use them to supervise the meta-controller.

C. Humanoid Vision-Language-Action Systems

Recent work extends VLA paradigms to humanoid platforms. Demonstration-based approaches map vision and language inputs directly to actions for manipulation or loco-manipulation, including GR00T N1 [4] and WholeBodyVLA [17]. Other work focuses on language-conditioned whole-body motion generation without dexterous manipulation, such as LeVERB [18] and Humanoid-VLA [19].

Our work follows the demonstration-based humanoid VLA paradigm but focuses on learning from segmented behavioral datasets. We study how expert specialization enables robust integration of locomotion and manipulation behaviors when learning from segmented behavioral datasets.

III. Problem Formulation

We study humanoid loco-manipulation as learning a goal-conditioned policy

$$\pi_\theta : \mathcal{G} \times \mathcal{S} \to \mathcal{A}, \tag{1}$$

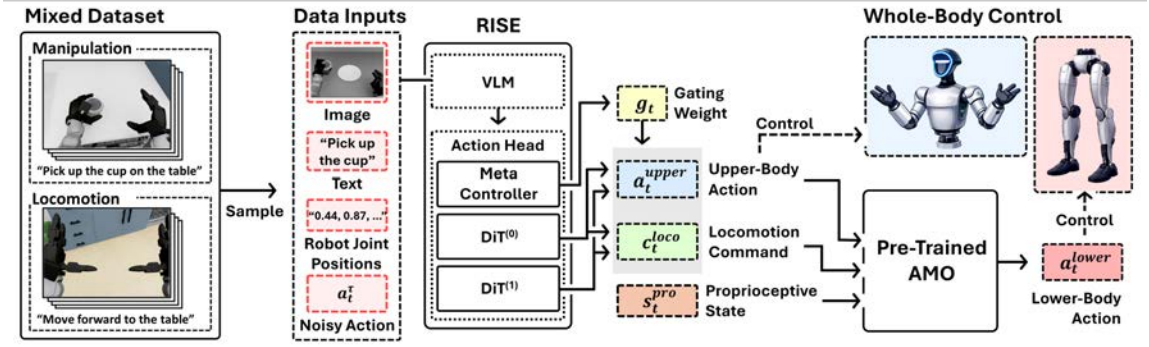

Fig. 2. **Overview of RISE.** RISE is trained on a mixed dataset containing independently collected locomotion- and manipulation-dominant demonstrations. Given an image, language instruction, robot joint positions, and a noisy action (red boxes denote quantities sampled from the dataset during training), RISE predicts expert gating weights together with an upper-body action and a locomotion command. The upper-body action and locomotion command are then provided to the AMO [7] locomotion policy to generate the corresponding lower-body action. This hierarchical interface produces whole-body joint targets and enables real-time humanoid control during task execution.

where the goal $g \in \mathcal{G}$ is provided as a natural-language instruction. At time step t, the agent receives a state $s_t \in \mathcal{S}$ defined as

$$s_t = \left(s_t^{\text{pro}},\, o_t\right), \tag{2}$$

where s_t^{pro} is the robot proprioceptive state and o_t is an exteroceptive observation of the environment. We use

$$s_t^{\text{pro}} = \left[\theta_t,\ \omega_t,\ q_t,\ \dot{q}_t,\ a_{t-1}\right], \tag{3}$$

which includes the base orientation θ_t, base angular velocity ω_t, joint positions/velocities $(q_t, \dot{q}_t)$, and the previous action a_{t-1}.

The action space $\mathcal{A}$ consists of whole-body joint position targets. We denote an action as

$$a_t = \left(a_t^{\text{upper}},\, a_t^{\text{lower}}\right) \in \mathcal{A}, \tag{4}$$

where a_t^{upper} and a_t^{lower} are the target joint positions for the upper-body and lower-body joints, respectively. We generate a_t via a hierarchical policy that separates high-level decision-making from robust lower-body control. Given a language instruction l, observation o_t, and joint positions q_t, the VLA module outputs a VLA action consisting of upper-body targets and a high-level locomotion command:

$$u_t = (a_t^{\text{upper}}, c_t^{\text{loco}}) \sim \pi_{\text{VLA}}(\cdot \mid l, o_t, q_t). \tag{5}$$

The locomotion command includes the linear velocities in the x and y directions, as well as the angular velocity. Conditioned on these VLA outputs and proprioception, a separate locomotion policy produces lower-body targets:

$$a_t^{\text{lower}} \sim \pi_{\text{loco}}(\cdot \mid a_t^{\text{upper}}, c_t^{\text{loco}}, s_t^{\text{pro}}). \tag{6}$$

We then form the whole-body target by concatenating the upper- and lower-body targets,

$$a_t = \left(a_t^{\text{upper}},\ a_t^{\text{lower}}\right). \tag{7}$$

IV. Methods

We propose Responsibility-Induced Specialized Experts (RISE), a VLA framework consisting of a multi-expert action head and a meta-controller (Fig. 2). RISE models heterogeneous behaviors using multiple experts and learns expert weighting through a meta-controller. During training, expert responsibilities are derived from each expert's flow-matching prediction error and used to supervise the meta-controller. This responsibility-driven supervision encourages experts to specialize in different behavioral patterns. The overall training and inference pipeline of RISE is illustrated in Fig. 3.

A. Flow Matching with Multi-Expert Action Head

RISE builds on the GR00T backbone. We reuse the text tokenizer, vision encoder, and VLM from GR00T-N1.5, and modify only the action head by replacing the single DiT architecture with a set of M DiT experts and an additional meta-controller.

At time step t, the VLM backbone encodes the language instruction and observation into a latent feature:

$$z_t = f_{\text{VLM}}(l,\ o_t). \tag{8}$$

Following GR00T, we train RISE with flow matching over the VLA action $u_t = (a_t^{\text{upper}}, c_t^{\text{loco}})$. Given a ground-truth VLA action u_t and noise $\epsilon \sim \mathcal{N}(0, I)$, RISE is trained to predict the vector field $u_t - \epsilon$. For a flow-matching timestep $\tau \in [0, 1]$, we construct the interpolated VLA action as

$$u_t^\tau = \tau u_t + (1 - \tau)\epsilon.$$

Since the action head of RISE consists of multiple DiT experts, each expert predicts an expert-specific flow at flow-matching timestep τ:

$$\hat{v}_{t,\tau}^{(i)} = f_{\text{DiT}}^{(i)}(z_t, u_t^\tau, q_t, \tau), \qquad i \in \{0, \ldots, M-1\}. \tag{9}$$

Conditioned on z_t, the meta-controller outputs expert gating weights. In the general M-expert case, it produces a gating distribution $\mathbf{g_t} \in \Delta^{M-1}$:

$$\mathbf{g_t} = f_{\text{meta}}(z_t), \qquad g_t^{(i)} \geq 0,\ \sum_{i=0}^{M-1} g_t^{(i)} = 1. \tag{10}$$

RISE then forms the final flow prediction by mixing the experts' predicted vector fields using $\mathbf{g_t}$:

$$\hat{v}_{t,\tau}^{\text{mix}} = \sum_{i=0}^{M-1} g_t^{(i)}\, \hat{v}_{t,\tau}^{(i)}. \tag{11}$$

At inference time, we initialize $u_t^0 \sim \mathcal{N}(0, I)$ and generate the VLA action via forward Euler integration:

$$u_t^{\tau_{k+1}} = u_t^{\tau_k} + \frac{1}{K}\hat{v}_{t,\tau_k}^{\text{mix}}, \tag{12}$$

where $\tau_k = k/K$ for $k \in \{0, \ldots, K-1\}$. We use $K = 4$, following the GR00T implementation.

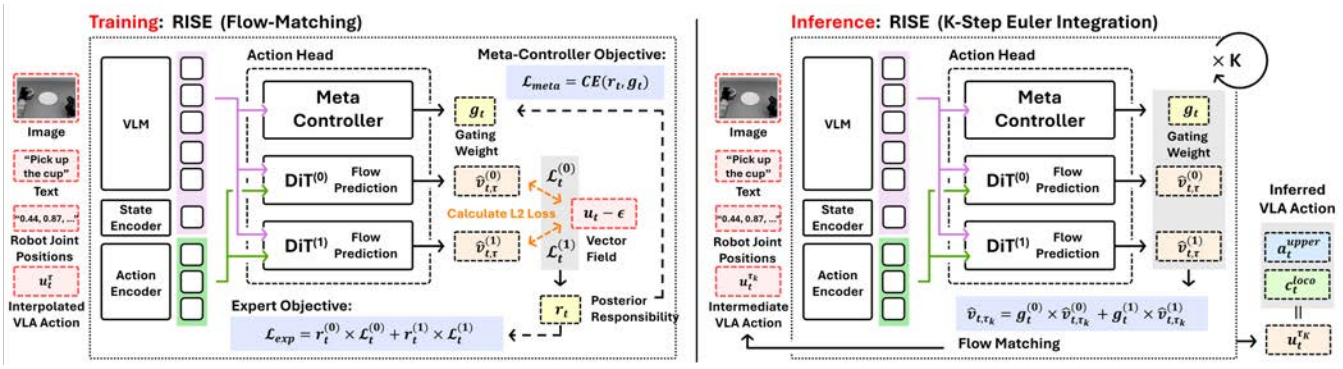

Fig. 3. **Training and inference of RISE.** RISE conditions on an image, text instruction, robot joint positions, and a noisy VLA action to learn a meta-controller and two DiT experts for flow matching. During training, the experts predict the vector fields and their prediction losses are weighted by the expert responsibilities computed from the per-expert flow-matching errors. At inference time, RISE runs a K-step Euler integration, where the VLA action is iteratively integrated by mixing the two experts' flow predictions using the gating weights output by the learned meta-controller. Quantities outlined in red indicate elements sampled from the dataset.

B. Training with Expert Responsibility

To train a flow-matching policy with multiple DiT experts, we encourage specialization across experts and supervise the meta-controller using expert-loss-derived responsibilities. The key principle is: *the expert that better predicts the target vector field should receive higher responsibility*.

Given a ground-truth VLA action $u_t = (a_t^{\text{upper}}, c_t^{\text{loco}})$, we sample noise $\epsilon \sim \mathcal{N}(0, I)$ and the target vector field becomes $u_t - \epsilon$. We obtain the flow-matching loss of each expert i as an expectation over the sampled noise ϵ and flow-matching timestep τ, where $\xi \sim \text{Beta}(1.5, 1.0)$, $\tau = s(1-\xi)$, and $s = 0.999$:

$$\mathcal{L}_t^{(i)} = \mathbb{E}_{\tau,\epsilon}\left[\left\|\hat{v}_{t,\tau}^{(i)} - (u_t - \epsilon)\right\|^2\right], \qquad i \in \{0, \ldots, M-1\}. \quad (13)$$

We obtain a responsibility from the expected per-expert flow-matching losses using sparsemax [20]:

$$\mathbf{r_t} = \text{sparsemax}\left(-\frac{1}{\beta}\left[\mathcal{L}_t^{(0)}, \ldots, \mathcal{L}_t^{(M-1)}\right]\right), \qquad \mathbf{r_t} \in \Delta^{M-1}, \quad (14)$$

where sparsemax maps scores to a probability vector on the simplex that can be sparse, and $\beta > 0$ controls the sparsity of the assignment. Using sparsemax encourages a sparse responsibility assignment across experts, which promotes expert specialization according to distinct behavior patterns.

Then we train experts with a responsibility-weighted flow-matching objective:

$$\mathcal{L}_{\text{exp}} = \sum_{i=0}^{M-1} r_t^{(i)}\, \mathcal{L}_t^{(i)}. \quad (15)$$

This weighting allocates larger gradients to the expert that better explains the target vector field under the current context, thereby encouraging specialization across largely separated behavioral patterns.

The meta-controller predicts expert gating weights over the M experts, denoted by $\mathbf{g_t} \in \Delta^{M-1}$, from backbone features. We supervise it to match the responsibility:

$$\mathcal{L}_{\text{meta}} = \text{CE}(\mathbf{r_t},\ \mathbf{g_t}) = -\sum_{i=0}^{M-1} r_t^{(i)} \log g_t^{(i)}. \quad (16)$$

We stop gradients through $\mathbf{r_t}$ when computing $\mathcal{L}_{\text{exp}}$ and $\mathcal{L}_{\text{meta}}$ to avoid degenerate gating updates that minimize the loss without promoting expert specialization.

While the above formulation supports M experts in general, we instantiate $M = 2$ to focus on humanoid loco-manipulation tasks and encourage each expert to specialize in locomotion and manipulation-dominant behavior. Let experts 0 and 1 denote the locomotion and manipulation-specialized experts, respectively.

Early in training, responsibilities derived from flow-matching objectives can be unreliable because both experts are inaccurate. To break symmetry, we warm-start the responsibility using an energy-guided heuristic computed from ground-truth actions.

Let the whole-body action from dataset be $a_t = (a_t^{\text{upper}}, a_t^{\text{lower}})$ and dataset-level means be μ^{upper} and μ^{lower}. We define a pseudo responsibility based on an energy score:

$$\tilde{\mathbf{r}}_\mathbf{t} = \text{softmax}\left(\alpha \frac{\left\|a_t^{\text{lower}} - \mu^{\text{lower}}\right\|_2}{\sigma^{\text{lower}}}, \alpha \frac{\left\|a_t^{\text{upper}} - \mu^{\text{upper}}\right\|_2}{\sigma^{\text{upper}}}\right), \quad (17)$$

where $\alpha > 0$ controls the sharpness, and σ^{upper} and σ^{lower} are dataset-level normalization scales for the upper and lower-body actions, respectively. This pseudo responsibility increases $\tilde{r}_t^{(0)}$ when lower-body activity dominates (e.g., walking), and increases $\tilde{r}_t^{(1)}$ when upper-body activity dominates (e.g., grasping).

During warm-up, we use $\tilde{\mathbf{r}}_\mathbf{t}$ to guide both expert learning and meta-controller supervision:

$$\mathcal{L}_{\text{exp}}^{\text{warm}} = \tilde{r}_t^{(0)}\, \mathcal{L}_t^{(0)} + \tilde{r}_t^{(1)}\, \mathcal{L}_t^{(1)}, \quad (18)$$

$$\mathcal{L}_{\text{meta}}^{\text{warm}} = \text{CE}(\tilde{\mathbf{r}}_\mathbf{t},\ \mathbf{g_t}). \quad (19)$$

We further align the flow-matching responsibility $\mathbf{r_t}$ with the energy-guided pseudo responsibility $\tilde{\mathbf{r}}_\mathbf{t}$ by minimizing a KL divergence between the corresponding two-class distributions:

$$\mathcal{L}_{\text{align}} = D_{\text{KL}}(\tilde{\mathbf{r}}_\mathbf{t} \parallel \mathbf{r_t}). \quad (20)$$

This alignment stabilizes the hand-off from energy-guided gating to error-based gating.

We apply the energy-guided terms only during the warm-up stage. Concretely, the warm-up objective is

$$\mathcal{L}^{\text{warm}} = \lambda_{\text{exp}}\, \mathcal{L}_{\text{exp}}^{\text{warm}} + \lambda_{\text{meta}}\, \mathcal{L}_{\text{meta}}^{\text{warm}} + \lambda_{\text{align}}\, \mathcal{L}_{\text{align}}. \quad (21)$$

After warm-up, we replace the warm-up losses with the main objective:

$$\mathcal{L}^{\text{main}} = \lambda_{\text{exp}}\, \mathcal{L}_{\text{exp}} + \lambda_{\text{meta}}\, \mathcal{L}_{\text{meta}}. \quad (22)$$

This objective jointly trains the shared VLM backbone, the M DiT experts, and the meta-controller, yielding a behavior-specialized whole-body control policy for long-horizon humanoid loco-manipulation.

C. Task Transition via Expert Gating

The meta-controller outputs an expert-gating distribution $\mathbf{g_t} \in \Delta^1$ (Eq. 10). In our two-expert instantiation, Expert 0 and Expert 1 specialize in locomotion- and manipulation-dominant behaviors, respectively. Since $g_t^{(0)} + g_t^{(1)} = 1$, we apply a short moving average to suppress temporary fluctuations and use the smoothed manipulation-expert gating weight, denoted by $\bar{g}_t^{(1)}$, as the task-transition signal.

Using a confidence threshold $\delta \in (0.5, 1)$, we divide the gating weight into a locomotion-confident region, an intermediate ambiguous region, and a manipulation-confident region. We define the gating-region label $\hat{y}_t \in \{0, 1, 2\}$ as

$$\hat{y}_t = \begin{cases} 0, & \bar{g}_t^{(1)} < 1 - \delta, \\ 1, & 1 - \delta \le \bar{g}_t^{(1)} \le \delta, \\ 2, & \bar{g}_t^{(1)} > \delta, \end{cases} \tag{23}$$

where labels 0, 1, and 2 denote the locomotion-confident, ambiguous, and manipulation-confident regions, respectively.

Near the completion of a subtask, the robot generally becomes relatively stationary, and the gating weight empirically moves from a confident region toward the intermediate region. We therefore define a task-transition event only when the gating weight enters the ambiguous region from either confident region:

$$\mathbf{1}_{\text{trans}}(t) = \mathbf{1}\left[\hat{y}_{t-1} \in \{0, 2\} \wedge \hat{y}_t = 1\right]. \tag{24}$$

When a transition is detected, the system advances to the next predefined subtask instruction. A transition from the ambiguous region back into either confident region is ignored. Consequently, the same rule handles transitions between different dominant behaviors as well as consecutive subtasks with the same dominant behavior.

V. Experiments

We evaluate RISE on long-horizon humanoid tasks composed of sequential subtasks in both simulation and the real world. Our experiments assess whether RISE can (i) learn distinct behavior patterns via a multi-expert action head and (ii) autonomously coordinate transitions between locomotion-dominant and manipulation-dominant stages using the meta-controller in compositional task sequences.

A. Robot Platform and Experimental Setup

Robot platform. We conduct all experiments with a Unitree G1 humanoid. The base platform has 29 DoF for whole-body articulation. To enable dexterous manipulation, we attach two Unitree Dex3-1 hands, each a 3-finger hand with 7 DoF, resulting in 43 controllable DoF in total. The robot is controlled at 50 Hz using joint position targets tracked by low-level PD control. We use the same sensing and control interface in both simulation and real-world experiments.

Simulation environment. We build a kitchen environment in Isaac Sim by porting RoboCasa [21] scenes, including a table, sink, and trash bin with household objects on the table.

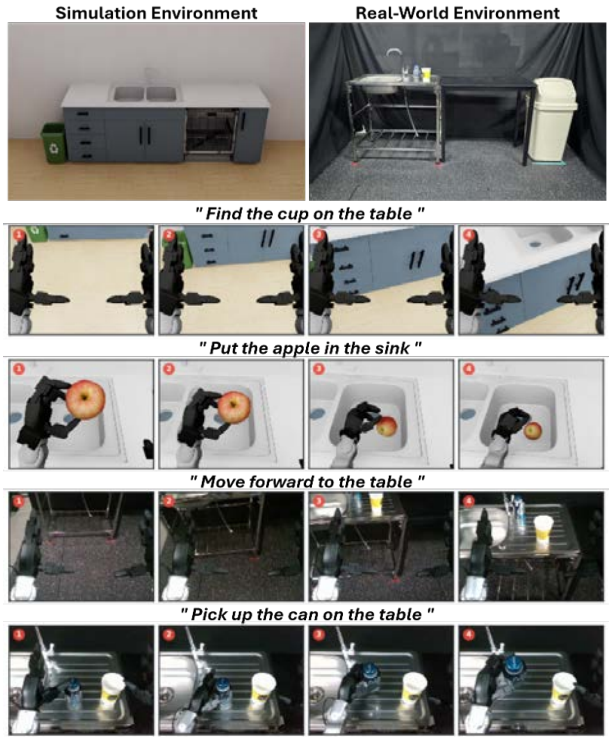

Fig. 4. **Simulation and real-world environments and trajectories.** Top left: simulated kitchen environment; top right: real-world environment. Rows 2–3: simulation locomotion/manipulation trajectory sequences from the dataset. Rows 4–5: real-world locomotion/manipulation trajectories.

Real-world environment. We construct a physical setup that mirrors the simulated layout. Fig. 4 illustrates the simulation and real-world environment layouts used for data collection and evaluation.

B. Loco-Manipulation Tasks and Evaluation Protocol

We evaluate RISE on long-horizon humanoid loco-manipulation tasks composed of an ordered sequence of subtasks specified by natural-language instructions. For each scenario, the subtask order is predetermined, and an episode is counted as successful only if the policy completes all subtasks in the prescribed order.

Simulation tasks. In simulation, we vary the subtask horizon and interaction target using three templates:

- **2-subtask (Move–Pick):** Move to the table and pick up the target object.
- **3-subtask (Pick–Move–Place):** Pick the object, move to the goal region, and place it.
- **4-subtask (Move–Pick–Move–Place):** Move to the table, pick the object, move to the goal region, and place it.

We evaluate three object categories (apple, bottle, cup). The placement goal is either the sink or the trash bin, as specified by the instruction.

Real-world tasks. For real-world experiments, we use task sequences that reflect the constraints and variability of physical execution. We evaluate:

- **2-subtask (Move–Pick):** Move to the table and pick up the target object.

- **4-subtask (Move–Pick–Move–Place):** Move to the table, pick the object, move to the goal region, and place it.
- **5-subtask (Pick–Place–Pick–Move–Place):** Pick and place an object at an intermediate location, then pick a second object, move to the goal, and place it.

We consider two object categories (can, cup), with goal regions fixed by object type according to the collected dataset: cans are placed either on a table or into a trash can, while cups are placed either on a table or into a sink. Consequently, each task sequence is instantiated with two predefined object–goal configurations.

For each episode, we provide the language instruction and execute the policy until termination. We report success as the fraction of episodes in which all subtasks are completed in order.

C. Data Collection and Dataset Design

We collect demonstration data using a hybrid control setup aligned with our hierarchical framework. The lower body is controlled by the AMO [7] locomotion policy, while the upper body is operated via teleoperation for manipulation data collection. Specifically, we use XR-Teleoperate [22] with Meta Quest 3 and Apple Vision Pro to control upper-body motions during manipulation.

Each subtask is categorized as either locomotion or manipulation. For locomotion subtasks, data is collected by sending high-level locomotion commands via a remote controller to reach designated target regions. For manipulation subtasks, the robot is first positioned at an appropriate interaction location, after which teleoperation is used to execute the task.

Subtasks are recorded independently rather than as continuous long-horizon sequences. This allows us to explicitly separate locomotion and manipulation-dominant behaviors during data collection.

In simulation, the dataset comprises seven subtasks spanning locomotion behaviors (e.g., moving forward to search, moving left toward the sink or the trash bin) and manipulation behaviors (e.g., picking, placing on the table, putting into the sink, and throwing into the trash bin). For manipulation subtasks, we collect data separately for each object category (apple, bottle, cup). Specifically, we record picking, placing on the table, and putting into the sink for all three objects, while throwing into the trash bin is defined for apples only. We collect 200 trajectories per subtask per object, resulting in 2,600 simulated demonstrations in total.

In the real world, the dataset comprises nine subtasks spanning locomotion tasks (e.g., moving to the sink or trash bin) and object manipulation tasks (e.g., picking and placing a cup or can, throwing a can into the trash bin). Each task includes 50–100 trajectories, resulting in a total of 600 real-world demonstrations.

D. Baseline Methods and Ablation Study

We compare RISE against baseline and ablation variants to isolate the contributions of (i) the RISE architecture with multi-expert action heads and a meta-controller, (ii) responsibility-based expert learning from flow-matching errors, and (iii) the energy-guided warm-up used to stabilize early specialization. All methods share the same observation/action interfaces and are executed under the same hierarchical pipeline with a lower-body locomotion policy.

GR00T. A single-expert VLA baseline without expert decomposition or a meta-controller. Subtask progression is manually triggered by a human operator during execution.

RISE w/o warm-up. RISE trained from scratch using error-derived responsibilities, without any energy-guided warm-up.

RISE (energy-guided only). RISE trained using the warm-up objective $\mathcal{L}_{\text{exp}}^{\text{warm}}$ for the entire schedule. Expert updates are weighted only by the pseudo responsibility (i.e., no error-derived responsibility-based weighting via $\mathcal{L}_{\text{exp}}$).

RISE (ours). Full RISE with energy-guided warm-up followed by responsibility-weighted expert learning from flow-matching errors.

All methods are initialized from the same pre-trained GR00T-N1.5 model and fine-tuned on our collected dataset.

TABLE I

SIMULATION SUCCESS RATES (%) BY SUBTASK HORIZON.

Task	GR00T	RISE w/o warm-up	RISE
2-subtask (apple)	90	**100**	**100**
2-subtask (bottle)	40	**80**	**80**
2-subtask (cup)	**100**	90	**100**
3-subtask (apple / sink)	80	**100**	**100**
3-subtask (bottle / sink)	60	70	**80**
3-subtask (cup / sink)	80	90	**100**
3-subtask (apple / trash bin)	80	70	**90**
4-subtask (apple / sink)	90	**100**	80
4-subtask (bottle / sink)	30	**60**	**60**
4-subtask (cup / sink)	**90**	80	80
4-subtask (apple / trash bin)	70	**90**	**90**
Average	73.6	84.5	**87.3**

Fig. 5. Success rate by training steps in simulation experiments.

E. Simulation Results

We evaluate RISE on compositional humanoid loco-manipulation tasks in simulation and compare it against a single-expert GR00T baseline and an ablation trained without energy-guided warm-up. Table I reports success rates by subtask horizon (2/3/4 subtasks) using the best checkpoint selected from the training sweep shown in Fig. 5. Fig. 5 summarizes learning dynamics over fine-tuning steps.

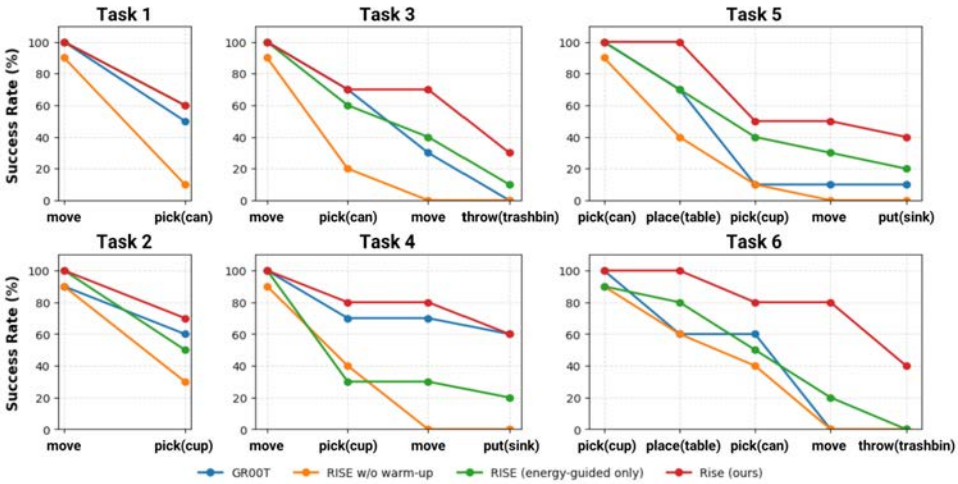

Fig. 6. Subtask-level progression across humanoid loco-manipulation tasks in the real world. Curves indicate the fraction of trials that successfully complete up to each subtask stage.

Across the simulation suite, RISE achieves the highest average success rate, improving performance from 73.6% with GR00T to 87.3%. The energy-guided warm-up further improves the average performance over RISE without warm-up, suggesting that early guidance helps stabilize expert specialization before the error-derived responsibilities become reliable. Remaining failures are primarily due to accumulated execution errors, such as imprecise approaches leading to failed grasps or placements, or mistimed transitions.

Fig. 5 also indicates that RISE learns more efficiently than GR00T, achieving higher success earlier in fine-tuning and maintaining an advantage throughout training. Overall, these results demonstrate that the multi-expert action head improves learnability for heterogeneous humanoid behaviors, and that energy-guided warm-up stabilizes training and improves average task performance.

F. Real-world Results

We evaluate RISE on compositional humanoid loco-manipulation tasks in the real world and compare it against a single-expert GR00T baseline, RISE without warm-up, and RISE with energy-guided training only. Tasks are grouped by subtask horizon (2, 4, and 5 subtasks). Success rates are summarized in Table II, and detailed task progression statistics are shown in Fig. 6.

Task transitions differ across methods. GR00T relies on manually triggered transitions from a human operator, while RISE and the energy-guided-only variant use a learned meta-controller. For RISE, the expert gating weights produced by the meta-controller during execution, along with the resulting task transitions, are shown in Fig. 7. Without the warm-up stage, however, the experts do not specialize well and the meta-controller fails to produce reliable transitions. Therefore, for RISE w/o warm-up we manually trigger task transitions during evaluation, following the same protocol as GR00T.

Across all methods, RISE achieves the highest overall performance with an average success rate of 50%, outperforming GR00T by 20 percentage points. While RISE and GR00T perform similarly on short-horizon tasks (2 subtasks), the gap widens as the horizon increases. On 4- and 5-subtask tasks, RISE is more consistent than the other methods, whereas GR00T fails more frequently as sequential execution becomes longer.

TABLE II

REAL-WORLD SUCCESS RATES (%) BY SUBTASK HORIZON.

Task	GR00T	RISE w/o warm-up	RISE (energy-guided only)	RISE
2-subtask (can)	50	10	**60**	**60**
2-subtask (cup)	60	30	50	**70**
4-subtask (can / trash bin)	0	0	10	**30**
4-subtask (cup / sink)	**60**	0	20	**60**
5-subtask (can / cup / sink)	10	0	20	**40**
5-subtask (cup / can / trash bin)	0	0	0	**40**
Average	30.0	6.7	26.7	**50.0**

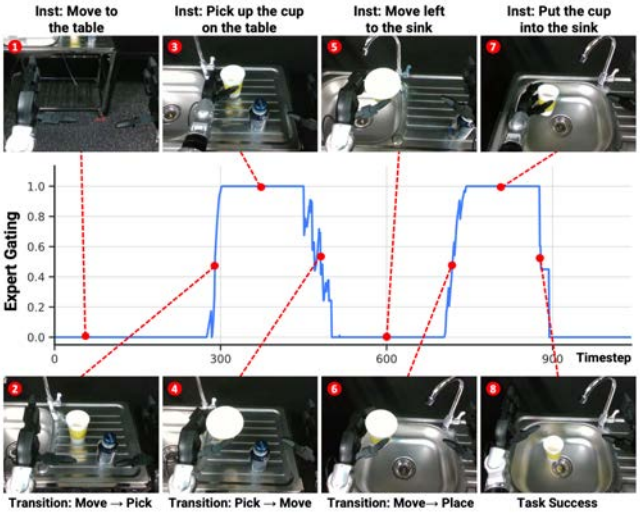

Fig. 7. **Task transitions via expert gating.** We visualize how the expert gating signal evolves over time during sequential task execution. At selected timesteps, we show the corresponding subtask label along with the aligned first-person observation.

A major bottleneck in real-world compositional loco-manipulation occurs at the locomotion-to-manipulation boundary. The success of the first manipulation subtask varies depending on whether locomotion is required beforehand. When manipulation is the first subtask (e.g., pick-first), the manipulation itself is relatively reliable. In contrast, when manipulation follows locomotion, failures occur more frequently because successful manipulation requires

reaching an interaction-ready pose (distance, yaw, and base alignment). Due to the limited and narrowly distributed real-world demonstrations, small deviations in the approach pose often lead to downstream grasp or placement failures.

Ablations further highlight the importance of responsibility-driven learning and warm-up in the real world. The energy-guided-only variant underperforms GR00T on average, indicating that heuristic gating alone is insufficient. Removing warm-up causes an even larger performance drop than in simulation, suggesting that early-stage stabilization is particularly important under noisier real-world dynamics and limited data.

VI. CONCLUSION

We presented RISE, a responsibility-driven multi-expert Vision–Language–Action framework for learning long-horizon humanoid loco-manipulation from segmented demonstrations. RISE models heterogeneous behavioral patterns using multiple action experts and learns expert weighting through responsibilities derived from flow-matching prediction errors. This mechanism encourages experts to specialize in different behavioral regimes and enables the policy to coordinate locomotion and manipulation without requiring explicit transition annotations. Experiments in both simulation and on a real humanoid robot show improved long-horizon task success compared to a single-expert GR00T baseline and ablation variants trained on the same segmented data.

Despite these results, our current formulation focuses on sequential loco-manipulation, where locomotion and manipulation dominate at different stages of execution. Extending this approach to fully simultaneous whole-body behaviors—such as manipulation while actively walking or dynamically interacting with the environment—remains an open challenge. In such settings, additional mechanisms may be required to model more tightly coupled interactions and overlapping behavior distributions.

Looking forward, the responsibility-based framework provides a flexible foundation for scaling to richer behavior distributions. While our current instantiation focuses on expert specialization across locomotion and manipulation-dominant behaviors, the framework could also support other forms of expert decomposition, such as hand-specific or task-dependent specializations. With appropriate guidance signals, this responsibility-driven coordination mechanism may enable broader forms of behavior decomposition and integration for scalable humanoid whole-body learning.

REFERENCES

[1] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning (CoRL)*, 2023.

[2] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, "π_0: A Vision-Language-Action Flow Model for General Robot Control," *arXiv preprint*, vol. arXiv:2410.24164, 2024.

[3] 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," *arXiv preprint*, vol. arXiv:2504.16054, 2025.

[4] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. J. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, J. Jang, Z. Jiang, J. Kautz, K. Kundalia, L. Lao, Z. Li, Z. Lin, K. Lin, G. Liu, E. Llontop, L. Magne, A. Mandlekar, A. Narayan, S. Nasiriany, S. Reed, Y. L. Tan, G. Wang, Z. Wang, J. Wang, Q. Wang, J. Xiang, Y. Xie, Y. Xu, Z. Xu, S. Ye, Z. Yu, A. Zhang, H. Zhang, Y. Zhao, R. Zheng, and Y. Zhu, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[5] M. Seo, S. Han, K. Sim, S. H. Bang, C. Gonzalez, L. Sentis, and Y. Zhu, "Deep imitation learning for humanoid loco-manipulation through human teleoperation," *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, pp. 1–8, 2023.

[6] X. Huang, D. Batra, A. Rai, and A. Szot, "Skill transformer: A monolithic policy for mobile manipulation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.

[7] J. Li, X. Cheng, T. Huang, S. Yang, R. Qiu, and X. Wang, "Amo: Adaptive motion optimization for hyper-dexterous humanoid whole-body control," in *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[8] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive whole-body control for humanoid robots," in *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[9] M. Ji, X. Peng, F. Liu, J. Li, G. Yang, X. Cheng, and X. Wang, "Exbody2: Advanced expressive humanoid whole-body control," in *RSS 2025 Workshop on Whole-body Control and Bimanual Manipulation: Applications in Humanoids and Beyond*, 2025.

[10] J. Shi, X. Liu, D. Wang, O. Lu, S. Schwertfeger, C. Zhang, F. Sun, C. Bai, and X. Li, "Adversarial locomotion and motion imitation for humanoid policy learning," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[11] 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)*, 2025.

[12] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," in *Conference on Robot Learning (CoRL)*, 2024.

[13] Q. Bu, Y. Yang, J. Cai, S. Gao, G. Ren, M. Yao, P. Luo, and H. Li, "Univla: Learning to act anywhere with task-centric latent actions," in *Proceedings of Robotics: Science and Systems (RSS)*, 2025.

[14] J. Yu, H. Liu, Q. Yu, J. Ren, C. Hao, H. Ding, G. Huang, G. Huang, Y. Song, P. Cai, C. Lu, and W. Zhang, "Forcevla: Enhancing vla models with a force-aware moe for contact-rich manipulation," 2025.

[15] C. Miao, T. Chang, M. Wu, H. Xu, C. Li, M. Li, and X. Wang, "Fedvla: Federated vision-language-action learning with dual gating mixture-of-experts for robotic manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 6904–6913.

[16] M. Reuss, J. Pari, P. Agrawal, and R. Lioutikov, "Efficient diffusion transformer policies with mixture of expert denoisers for multitask learning," in *International Conference on Learning Representations, ICLR*, 2025.

[17] 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," *arXiv preprint arXiv:2512.11047*, 2025.

[18] 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*, vol. arXiv:2506.13751, 2025.

[19] 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*, vol. arXiv:2502.14795, 2025.

[20] A. F. T. Martins and R. F. Astudillo, "From softmax to sparsemax: A sparse model of attention and multi-label classification," in *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016, pp. 1614–1623.

[21] S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, A. Mandlekar, and Y. Zhu, "Robocasa: Large-scale simulation of everyday tasks for generalist robots," in *Robotics: Science and Systems (RSS)*, 2024.

[22] Unitree Robotics, "XR-Teleoperate: An open-source teleoperation framework and data collection toolkit for embodied intelligence," https://github.com/unitreerobotics/xr_teleoperate, 2024, accessed: 2026-02.