2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

RPG: Robust Policy Gating for Smooth Multi-Skill Transitions in Humanoid Fighting

Yucheng Xin[1,2*], Jiacheng Bao[3,2*], Yubo Dong[4,2], Xueqian Wang[1], Bin Zhao[2,3]
Xuelong Li[3], Junbo Tan[1†], Dong Wang[2†]

***Abstract*— Humanoid robots have demonstrated impressive motor skills in a wide range of tasks, yet whole-body control for humanlike long-time, dynamic fighting remains particularly challenging due to the stringent requirements on agility and stability. While imitation learning enables robots to execute human-like fighting skills, existing approaches often rely on switching among multiple single-skill policies or employing a general policy to imitate input reference motions. These strategies suffer from instability when transitioning between skills, as the mismatch of initial and terminal states across skills or reference motions introduces out-of-domain disturbances, resulting in unsmooth or unstable behaviors. In this work, we propose RPG, a hybrid expert policy framework, for smooth and stable humanoid multi-skills transition. Our approach incorporates motion transition randomization and temporal randomization to train a unified policy that generates agile fighting actions with stability and smoothness during skill transitions. Furthermore, we design a control pipeline that integrates walking/running locomotion with fighting skills, allowing humanlike long-time combat of arbitrary duration that can be seamlessly interrupted or transit action policies at any time. Extensive experiments in simulation demonstrate the effectiveness of the proposed framework, and real-world deployment on the Unitree G1 humanoid robot further validates its robustness and applicability.**

I. INTRODUCTION

Humanoid robots have shown remarkable progress in recent years, demonstrating agile locomotion [1], [2], [3], [4], [5], dexterous manipulation [6], [7], [8], [9], [10], [11], [12] , and even athletic motions [13], [14], [15]. A complex, practical, and compelling testbed task is enabling humanoid robots to perform human-like fighting actions such as jumping [4], [14], [2], punching [16], sword swings, and kicking [15], [17]. Such task requires whole-body coordination, rapid skill switching, and stable contact control, which place much higher demands on robustness and agility compared to conventional locomotion or manipulation. An ideal humanoid whole-body control policy for fighting would resemble controlling a character in a role-playing game (RPG), where a user can seamlessly trigger and execute diverse combat skills in real time. Achieving this level of flexible and fluid multi-skill control in physical humanoid robots, however, remains a significant challenge.

Recent advances in imitation learning have enabled robots to learn complex motion skills from reference motion trajectories. While effective for single-skill learning [13], [15], extending imitation learning to multiple diverse skills is non-trivial. Naive solutions, such as switching between separate policies or using a general policy to mimic diverse reference motions [16], [18], [19], [17], [20], often lead to instability in policy deployment. This is primarily due to mismatched initial and terminal states between reference sequences, leading to out-of-domain disturbances during skill transitions. As a result, robots suffer from jerky or unsmooth motions, particularly when attempting to concatenate different fighting skills for practical human-like combat.

To overcome these challenges, we propose Robust Policy Gating (RPG), a hybrid expert policy framework designed for multi-skill imitation learning with smooth skill transitions. Our method first trains multiple expert policies on distinct categories of fighting skills. During training, we introduce policy-transition randomization and temporal randomization to explicitly simulate mid-sequence truncations and arbitrary skill switches, forcing each expert to learn robustness against discontinuous motions. Once the experts converge, we freeze them and train a lightweight gating network that outputs weighted combinations of their actions, regularized with torque and contact smoothness objectives. The resulting hybrid expert policy controller enables fluid switching across skills, even under abrupt transitions.

Extensive experiments in simulation demonstrate that RPG effectively generates agile and seamless fighting behaviors, even under abrupt action switches. Beyond isolated fighting skills, we design a control pipeline that integrates locomotion and combat behaviors. When no fighting action is triggered, the robot maintains a locomotion state; when a skill is commanded, the system seamlessly transitions to the specified action, supporting arbitrary interruptions and varying durations. This design provides a game-like interface for controlling humanoid robots, akin to RPG game combat mechanics. Furthermore, we validate the approach on the Unitree G1 humanoid robot, confirming that RPG transfers successfully to real hardware and enables robust execution of high-agility fighting motions.

In conclusion, our contributions are as follows:

1. We introduce Robust Policy Gating (RPG), the framework enabling smooth and robust multi-action transitions for humanoid fighting behaviors.

This work was supported by Shanghai AI Laboratory, the Natural Science Foundation of Shenzhen (No.JCYJ20230807111604008, No. JCYJ20240813112007010), the Natural Science Foundation of Guangdong Province (No.2024A1515010003) and Cross-disciplinary Fund for Research and Innovation (No. JC2024002) of Tsinghua SIGS.

[1]Center of Artificial Intelligence and Robotics, Shenzhen International Graduate School, Tsinghua University, China, {xin-yc23@mails., wang.xq@sz., tjblql@sz.}tsinghua.edu.cn,
[2]Shanghai AI Laboratory, [3]Northwestern Polytechnical University, [4]Shanghai Jiao Tong University
[†]Corresponding author, *indicates equal contribution

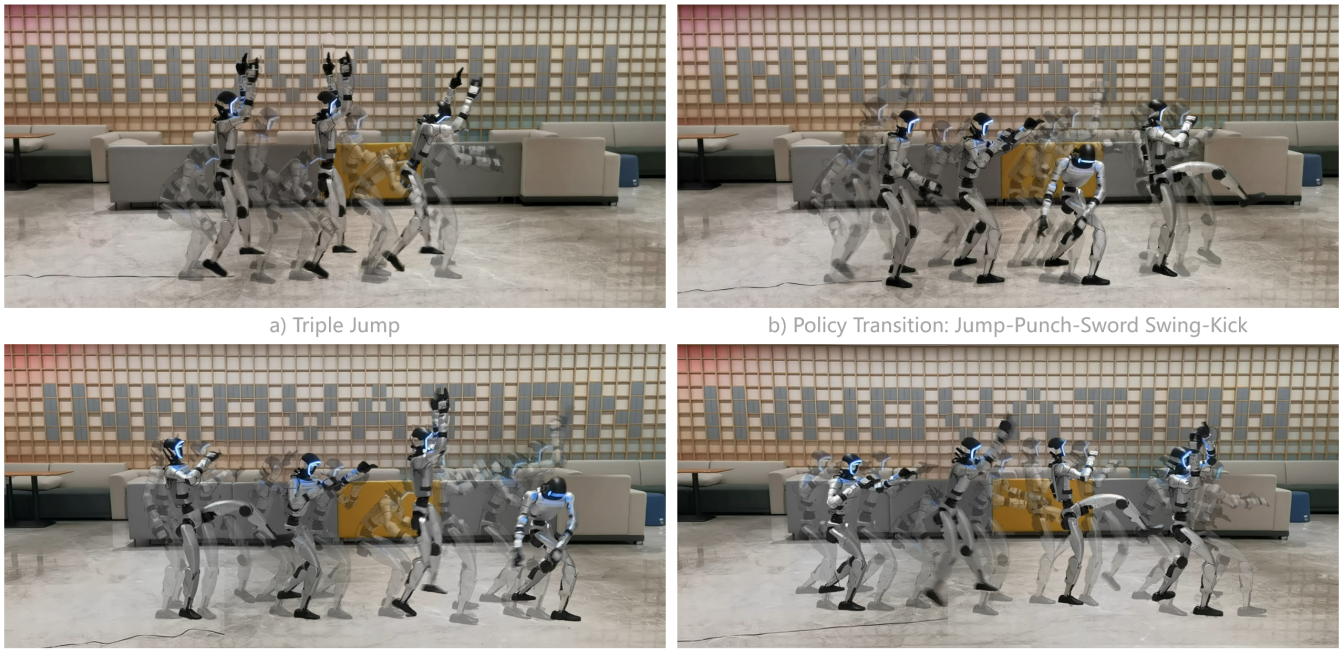

Fig. 1. **Policy Transition Demonstration.** We conducted policy transition tests for robotic combat motions using the proposed RPG. Punching and Sword Swing motions primarily involve the upper body, whereas Jumping and Kicking motions are mainly lower-body actions. We demonstrate 4 distinct policy transition combinations here, highlighting the motion capabilities during transition between upper and lower-body strategies. **a)** Jumping introduces significant instability for humanoid robot. However, by repeatedly executing the jumping policy in succession, we achieved a triple jump sequence. **b)-d)** Other policy transition combinations.

Fig. 2. **Recovery Tests.** Due to the short duration of the Jumping motion, it was excluded from consideration. For the other motions, the robot demonstrated the ability to avoid foot sliding and resume a stable, human-like standing posture from any interrupted state during execution.

2. We introduced a novel method incorporating both policy-transition and temporal randomization during expert policies training, and a new gating network for smoothness regularization, achieving robustness against disturbances during policy transitions and stable fusion of multiple expert policies.
3. We designed a pipeline that integrates locomotion and fighting skills, allowing users to control the robot for prolonged combat in a manner similar to playing an action RPG game.

II. RELATED WORKS

A. *Imitation Learning for Humanoid Whole-Body Control.*

Recent advances in imitation learning have empowered humanoid robots to acquire dynamic and naturalistic skills by tracking and synthesizing complex human motions. For agile behaviors, works such as [13], [21] focus on jumping, while PBHC [15] demonstrates multi-step motion reproduction for Kungfu and dancing. BeyondMimic [14] further introduces a guided diffusion framework that enables humanoids to execute highly challenging motions, including spins, sprinting, and cartwheels, while also supporting zero-shot task-specific control.

Expressive and resilient skills have also been explored: Exbody [22] and Exbody2 [23] focus on dance imitation, whereas Embrace Collisions [24] tackles recovery behaviors. Humanoid parkour has been showcased in [2], [25]. For contextual imitation, Videomimic [26] leverages everyday human videos with environment reconstruction to teach humanoids real-world tasks such as stair climbing, sitting, and standing.

On scalable and general motion tracking, GMT [19] employs adaptive sampling and a motion mixture-of-experts (MoE) to capture diverse skills with a single controller, while

BumbleBee [27] distills clustered experts into a generalist policy for whole-body humanoid control. HuB [17] further integrates motion refinement, balance-aware learning, and robustness training to master extreme balancing poses such as the Swallow Balance and Bruce Lee's Kick.

Beyond full-body imitation, recent efforts extend to manipulation and sports. OKAMI [28] and HumanPlus [29] emphasize upper-body skills for object interaction, and HITTER [20] combines model-based planning with reinforcement learning to achieve agile, stable table-tennis rallies.

B. Policy Switching and Motion Transition.

Enabling diverse robot skills often involves skill composition or switching mechanisms, such as R2S2's composable skill library [8] or automatic gait discovery [30]. ASE [31] learns a continuous latent skill space through adversarial imitation, enabling smooth skill transitions without predefined segmentation. Other methods use contextual expert switching: [32] uses uncertainty to switch between RL and BC. For robustness, [33] formalizes policy selection as a multi-armed bandit problem.

Generating smooth and stable transitions between motions is critical for robust robotic performance and has been addressed through various methods. These include dedicated transition policies like the Expert Composer [34] for skill sequencing, as well as techniques that engineer smoothness directly into controllers through regularization [35], command interpolation [10], and comfort-oriented rewards [36]. Our Robust Policy Gating (RPG) framework builds on this line of work by achieving seamless transitions between diverse skills through a specific randomization method and a learned gating network.

III. METHODS

A. Framework.

Our proposed Robust Policy Gating (RPG) framework is designed to learn a unified control policy that seamlessly executes and transitions between multiple dynamic fighting skills. The methodology consists of four core components: (i) data acquisition and motion retargeting, (ii) robust expert policies training via policy-transition and temporal randomization, (iii) a gating network for smooth policy fusion, (iv) sim-to-sim and sim-to-real validation and integration with a locomotion policy for a complete control pipeline. The overall framework is depicted in Fig. 3.

B. Data Process.

We begin by collecting motion references from two sources: (i) existing public motion datasets [37], and (ii) video recordings obtained from demonstrations or online footage. To convert raw video into motion sequences, we employ the GVHMR [38] framework, which extracts 3D human motions from monocular video. The resulting motion clips are then processed using the PHC retargeting method [39], adapting the human motions to the humanoid robot's kinematic structure while preserving key dynamics. From this pipeline, we obtain a set of representative actions relevant to humanoid fighting: punching, jumping, sword swings, and kicking. These action segments form the basis for policy training, generating kinematically feasible reference motion sequences $\mathcal{M}_m^r, m \in M$, where $M = \{j, p, s, k\}$ represents motions of jumping, punching, sword swing and kicking.

C. Expert Policy Training with Randomization Strategies.

For each action category, we train a dedicated policy $\pi^m, m \in M$ using Proximal Policy Optimization (PPO). Each policy is optimized to imitate its corresponding reference motion $\mathcal{M}_m^r$. All policies share the same environment, but only one policy is updated per step. This means that the update of each policy is unaffected by the other policies, but only influenced indirectly through the changes in the robot's state caused by other policies in the simulation environment.

For imitation learning–based multi-policy-transfer smoothness and robustness, the primary challenge is that out-of-domain disturbances during policy switching arise from large discontinuities in the reference motions encoded by different policies. To address this issue, we take a twofold approach: (i) introducing policy-transition scenarios during training, and (ii) injecting sudden perturbations to the robot's body state through temporal discontinuities in the reference motions. In particular, to improve the smoothness and robustness against action transitions, we introduce two forms of randomization operation during training:

1) **Randomization for Policy Transition Robustness:** A key innovation in our training pipeline is the incorporation of a randomization method to explicitly train for transition scenarios, moving beyond naive single-skill imitation. During training, an episode does not necessarily consist of a single full motion playback. The length of a complete episode T_e is defined as the sum of the durations of the all categories of reference motions $T_m, m \in M$. At any timestep t, with fixed probability $p_{trans} \in [0, 1]$, a transition event is triggered stochastically as follows:

$$b_t = \mathbf{Bernoulli}(p_{trans})$$

If $b_t = 1$, the current policy π^m and reference motion M_m^r is abruptly truncated. Based on the current simulator and robot state, the update of the original policy network π^m is frozen, and the training randomly switches to another action-specific policy $\pi^{m'}$, where m' is sampled uniformly from the motion set M, with the reference motion $\mathcal{M}_t$ reset accordingly.

$$\pi_{t+1} = \begin{cases} \pi^{m'}, \text{if } \mathrm{b_t} = 1 \\ \pi^m, \text{otherwise} \end{cases}$$

$$\mathcal{M}_{t+1} = \begin{cases} \mathcal{M}^r_{m',t=0}, \text{ if } \mathrm{b_t} = 1 \\ \mathcal{M}^r_{m,t=t+1}, \text{ otherwise} \end{cases}$$

This process forces each expert π^m to learn to start and terminate from a broad distribution of states, effectively simulating the situation that frequent interruptions and transitions in combat scenarios, and improving the robustness when policy switching.

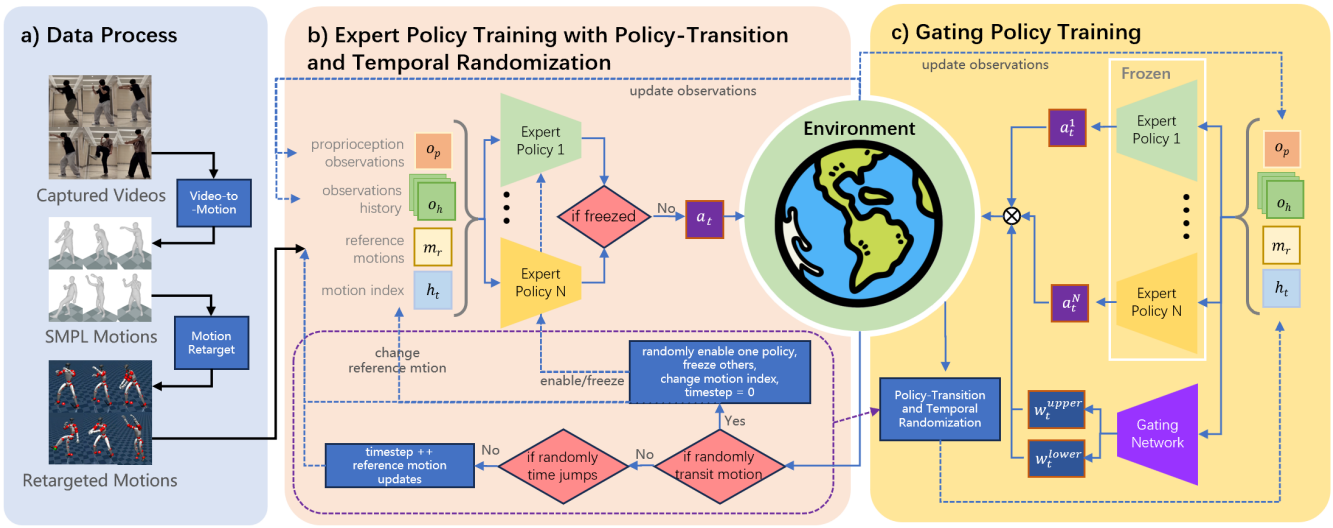

Fig. 3. **Framework. a)** Collection of human demonstration data for combat motions (jumping, punching, sword swing, kicking), with video-based motion conversion and retargeting to adapt to the Unitree G1 humanoid robot; **b)** Expert networks are trained for each motion type via imitation learning. All experts are loaded concurrently, though only one policy is updated per step. Policy-transition and temporal randomization are introduced: at any time step, the system may stochastically switch to a new motion, freezing the current policy, resetting the reference motion, and enabling a new policy to update. Additionally, random time-jumping in reference motion sequences is applied to simulate abrupt motion changes and enhance robustness. **c)** After the expert policies convergence, the expert policies are frozen while the randomization method remains active. A gating network is trained to blend outputs from the experts. The action output is decomposed into components for upper and lower bodies, and the gating network computes weighted combinations for each. The gating network is optimized with smoothness regularization to improve motion fluency during policy transition.

2) **Randomization for Temporal Discontinuities:** When sampling a reference state from $\mathcal{M}^r_m$, we do not strictly follow consecutive frames. With a fixed probability p_{jump}, the reference motion at the next timestep $\mathcal{M}_{t+1}$ is randomly replaced by the reference motion from $k \sim \mathcal{U}\{1, K\}$ steps ahead.

$$\mathcal{M}_{t+1} = \begin{cases} \mathcal{M}^r_{m,t=t+k}, & \text{if } \textbf{Bernoulli}(\text{p}_{\text{jump}}) = 1 \\ \mathcal{M}^r_{m,t=t+1}, & \text{otherwise} \end{cases}$$

This compels the policy to handle non-smooth, discontinuous reference commands, significantly improving its robustness to timing misalignments during transitions.

The aforementioned randomization techniques are formalized in pseudocode as presented in Algorithm 1. This approach of imitation learning with randomized cross-policy and cross-temporal domain sampling effectively simulates two critical aspects of combat motions: the necessity to terminate ongoing actions and transition to new strategies at any moment, and the ability to handle abruptly changing motion references. Consequently, it significantly enhances the robustness of each individual expert policy.

For the above expert networks, we employ an Asymmetric PPO algorithm for imitation learning training. All expert networks have an identical architecture. Upon completion of the described training process, we obtain multiple expert networks that can operate independently to execute their respective reference motions. In the Asymmetric PPO framework, the observations for the actor and critic networks are detailed in Table I. All networks employ an MLP architecture. The output of all networks is a 23-dimensional joint actions $a \in \mathbb{R}^{23}$ for the robot.

Algorithm 1 Randomization Strategies

Load policies for all categories of motions
Choose one policy $\pi^m, m \in M$ to be activated, others are frozen
for each timestep t in an episode **do**
  Sample current reference frame $\mathcal{M}_t = \mathcal{M}^r_{m,t}$
  Execute action $a_t \sim \pi^m(\mathcal{S}_t)$
  Update policy π^m
  — Randomization for Policy Transition —
  if Bernoulli(p_{trans}) = 1 **then**
    Freeze update of π^m
    Sample New Reference Motion $m' \in M$
    Activate a new policy $\pi^m = \pi^{m'}$
    Reset reference motion $\mathcal{M}_{t+1} = \mathcal{M}^r_{m',t=0}$
  end if
  — Randomization for Temporal Discontinuities —
  if Bernoulli(p_{jump}) = 1 **then**
    Replace $\mathcal{M}^r_{m,t+1}$ with $\mathcal{M}^r_{m,t+k}$, $k \sim \mathcal{U}(1, K)$
  end if
  Update States $\mathcal{S}_t$
end for

We define many of the reward terms in the following format:

$$\mathcal{R} = \sum_i w^e_i R(r_i)$$

$$R(r_i) = \exp\left(-||r_i - r^{ref}_i||^2_2 / \sigma_{r_i}\right)$$

where the w^e is weight for expert policies training, mean-

TABLE I

OBSERVATION SPACE FOR THE EXPERT NETWORKS

Observation States	**Actor Dims**	**Critic Dims**
Joint Positions	23	23
Joint Velocities	23	23
Root Angular Velocity	3	3
Root Projected Gravity	3	3
Actions	23	23
Reference Motion Phase	1	1
History (above)	76×4	76×4
Reference Joint Positions	23	23
Reference Joint Velocities	23	23
Reference Body Positions	81	81
Future Reference Joint Positions	23×2	23×2
Future Reference Body Positions	81×2	81×2
Root Linear Velocity	-	15
Body Position Difference	-	81
Randomized Base CoM offset	-	3
Randomized Base Link Mass	-	22
Total Dimensions	**715**	**836**

while w^g firstly mentioned in Table. II is for the gating policy training. The update rule for the σ_{s_i} value follows the same definition as presented in the PBHC [15].

The reward function used during training is outlined in Table II. The reward is primarily divided into two components: one for motion tracking, and another for ensuring smooth task execution. For the latter component, we specifically introduce penalties on the absolute values of output torques and their temporal differences to ensure motion fluency. Furthermore, to prevent biologically implausible phenomena such as foot sliding during policy transitions, where feet might drag along the ground to reach target positions, we designed additional reward terms that penalize foot-ground sliding and encourage prolonged foot aerial phases. This design promotes lifting the feet during movement, better mimicking natural human motion characteristics.

TABLE II

REWARDS OF EXPERT POLICIES

Rewards	**Expression**	w^e	w^g
Joint Positions	$R(q_t)$	1.0	0.3
Joint Velocities	$R(\dot{q}_t)$	1.0	0.3
Body Positions	$R(p_t)$	2.0	0.6
Body Rotations	$R(\theta_t)$	0.5	0.2
Body Velocities	$R(\dot{p}_t)$	0.5	0.2
Body Ang-Velocities	$R(\dot{w}_t)$	0.5	0.1
Upper Body Positions	$R(p_t^{up})$	4.0	0.2
Feet Positions	$R(p_t^{feet})$	1.0	0.6
Max Joint Positions	$\exp\left(-\|\|q_t - q_t^r\|\|_\infty/\sigma_{mj}\right)$	1.0	1.0
Joint Position Limits	$\mathbb{I}(q_t \notin [q_{min}, q_{max}])$	-10.0	-5.0
Joint Velocity Limits	$\mathbb{I}(\dot{q}_t \notin [\dot{q}_{min}, \dot{q}_{max}])$	-5.0	-3.0
Joint Torque Limits	$\mathbb{I}(\tau_t \notin [\tau_{min}, \tau_{max}])$	-5.0	-4.0
Feet Contact Forces	$\min(\|\|F^{feet} - 400\|\|_2^2, 0)$	-1e-2	-1e-2
Feet Air Time	$\mathbb{I}[T_{air} > 0.3]$	-1.0	-1.0
Feet Slipping Penalty	$\|\|v^{feet}\|\|_2^2 \cdot \mathbb{I}[\|\|F^{feet}\|\|_2 > 1]$	-3.0	-2.0
Torque	$\|\|\tau\|\|_2^2$	-1e-6	-1e-6
Collision Penalty	$\mathbb{I}_{collision}$	-30.0	-20.0
Termination Penalty	$\mathbb{I}_{termination}$	-300.0	-300.0
Alive	1	0.8	0.8
Torque Difference	$\|\|\tau_t - \tau_{t-1}\|\|_2^2$	-	-0.5
Task ID Check	$\exp(-\|\|c_t - c_t^r\|\|_2^2)$	-	2.0

During the training process, we also incorporated elements of curriculum learning by progressively tightening the precision requirements for motion tracking errors, thereby enhancing the convergence effectiveness of the policy.

D. Gating Policy Training.

Once the individual expert policies are obtained, and to involve the smoothness constraints on controlling, we freeze their parameters and design a gating network to produce smooth transitions while leveraging the motion capabilities of existing expert policies. Since the upper and lower limbs play different roles in the tracking task during motion execution, we define the dimension of the gating network output $\hat{w}$ as twice the number of expert policies $\hat{w} \in \mathbb{R}^{2N}$. Specifically, separate weight coefficients are assigned to the outputs of each expert policy for the upper and lower limbs, which are then applied independently to generate the final actions. The gating network receives the current robot states $\mathcal{S}_t$, which are the same as the input states of expert policies, and an optional task embedding c_t that encodes the high-level combat instruction. It outputs a weight vector $\hat{w}_t = [\hat{w}_t^{u,1}, \ldots, \hat{w}_t^{u,N}, \hat{w}_t^{l,1}, \ldots, \hat{w}_t^{l,N}]$ through a softmax layer, where N denotes the number of expert policies, $\{u, l\}$ represents the weights for upper and lower body actions. The final control action a^g is then computed as a weighted mixture of the experts:

$$a_t^g = \sum_p \sum_m \hat{w}_t^{p,m} \pi_p^m(\mathcal{S}_t, c_t), p \in \{u, l\}, m \in M$$

specifically, c_t is an N-dimensional one-hot vector, where the i-th entry is set to 1 for specific motion $\mathrm{index}(m)$, and all other entries are 0.

$$c_{t,i} = \begin{cases} 1, & \text{if i} = \text{index(m)} \\ 0, & \text{otherwise} \end{cases}$$

The training procedure of the gating network follows the same scheme as that of the expert networks, adopting an asymmetric PPO structure. The actor policy is employed as a MLP network, followed by a softmax output. This lightweight architecture ensures real-time inference on physical hardware. In addition, we regularize the gating output with a temporal smoothness constraint to discourage rapid fluctuations between experts, which can otherwise result in discontinuous or unstable motions.

Its observation state is identical to that listed in Table I, but further augmented with a task vector c_t in actor network. During algorithm updates, the gating network places greater emphasis on task scheduling, as well as the smoothness and stability of output torques. Therefore, its reward design differs from that of the expert networks. As shown in Table II, the single-step reward is defined as:

$$\mathcal{R} = \sum_i w_i^g R(r_i)$$

E. Validation and Deployment.

After completing the training of all experts and gating policies, we validated our framework in the MuJoCo simulation

environment and subsequently deployed it onto the Unitree G1 humanoid robot.

Additionally, we specifically designed a control pipeline for robotic combat tasks. In the real-world control pipeline, action commands are issued via a game joystick similar to playing an RPG action game: holding down a specific button triggers the execution of the corresponding action policy, while the robot defaults to locomotion mode when no command is given. For locomotion, we integrate a policy trained with the `RoboMimic` framework, ensuring stable operation of the robot outside combat tasks. We implemented a complete control pipeline that enables RPG-style interaction: the robot can execute any fighting skill on demand, smoothly stop and switch between different actions at arbitrary timesteps, and seamlessly return to a locomotion state when no fighting commands are issued.

IV. Experiments

A. Experimental Setup.

We ultimately recorded monocular video data and converted them into SMPL-format motions, which were then retargeted to the Unitree G1 humanoid robot. Four representative fighting skills are considered: *punching*, *jumping*, *sword swing*, and *kicking*. Both the *punching* and *sword swing* motions are composed of multiple consecutive striking actions combined to form a complete movement. Based on the proposed RPG framework, we obtained multiple imitation-learning expert policies capable of executing individual actions, as well as a multi-policy control pipeline that enables stable and seamless policy transitions. The experiments were trained on the IsaacGym simulation platform using an RTX 4090 GPU, with sim-to-sim policy validation performed in MuJoCo prior to deployment. The control frequency on the real robot is 50 Hz. Finally, the proposed framework was successfully deployed on the real Unitree G1 robot. As illustrated in Fig. 1 and Fig. 2, the final demonstration shows that the robot not only executes individual expert actions effectively, but also achieves smooth and stable transitions across multiple action policies.

The proposed RPG framework makes two primary contributions: (i) a training methodology incorporating policy-transition and temporal randomization mechanisms, and (ii) a gating network for policies merging. To validate the effectiveness of the proposed RPG framework, we conduct ablation studies, which will be discussed in the following sections.

B. Baseline and Metrics.

We employ the ASAP framework as the baseline for experimental comparisons, and all subsequent comparative experiments are conducted based on variants of the ASAP framework.

To evaluate the effectiveness of imitation learning for motion tracking, we introduce the following metrics to evaluate joint tracking performance: E_{mpjpe}(mean per-joint position error), E_{mpjae}(mean per-joint angle error), E_{mpjve}(mean per-joint velocity error), and metrics to assess root body tracking: E_{rootpe}(root position error), E_{rootre}(root rotation error), E_{rootve}(root velocity error).

To evaluate the robustness and control stability of the algorithm during cross-policy transitions, we propose a success rate metric denoted as $E^{a\to b}_{succ,n}$, which represents the success rate of the robot completing the transition from motion a to motion b across n experimental trials. The motion indexes can be described as j for jumping, p for punching, s for sword swing, and k for kicking.

Additionally, we introduce E_{maccj} (mean acceleration of joints) to evaluate the smoothness of robot control.

C. Main Results.

For the overall framework, our investigation focuses on the following key questions:

Q1: The impact of policy-transition and temporal randomization on individual expert networks.

To explore the impact of the RPG method on individual expert policies, we design a set of controlled experiments. The baseline approach involves using the ASAP framework to train imitation learning policies for each action separately. In contrast, the experimental group employs the RPG framework, where multiple expert networks are trained simultaneously using policy-transition and temporal randomization. After the training convergence, the corresponding individually trained expert networks from RPG are compared with those from the baseline.

TABLE III
Expert Policies Comparison of Tracking Effectiveness

Motions	Experiments	Metrics					
		E_{mpjpe}	E_{mpjae}	E_{mpjve}	E_{rootpe}	E_{rootre}	E_{rootve}
Jumping	baseline	**0.1492**	**0.0346**	**1.6895**	0.1521	1.5218	**0.2915**
	RPG (ours)	0.1594	0.0382	1.8231	**0.1387**	**1.3874**	0.3342
Punching	baseline	0.1428	0.0315	**1.5218**	0.1289	**1.2186**	**0.2489**
	RPG (ours)	**0.1321**	**0.0283**	1.6254	**0.1164**	1.3247	0.2753
Sword Swing	baseline	0.1689	**0.0321**	1.8927	**0.1298**	1.4672	0.3028
	RPG (ours)	**0.1524**	0.0368	**1.7346**	0.1423	**1.3429**	**0.2784**
Kicking	baseline	**0.1412**	**0.0302**	1.7825	0.1357	1.3984	**0.2631**
	RPG (ours)	0.1537	0.0339	**1.6428**	**0.1225**	**1.2873**	0.2896

We evaluated the policies trained for jumping, punching, sword swing, and kicking using the aforementioned motion tracking metrics. For lower-body dominant motions such as jumping and kicking, the tracking performance of RPG is generally weaker. This may be attributed to the greater disturbance impact on the lower limbs when handling abrupt motion transitions. Meanwhile, the baseline method shows relatively stronger performance in velocity tracking, likely because the robot's control is subject to safety constraints when sudden changes in reference motion cause sharp velocity variations.

However as the results, as shown in the table, indicate that there is no significant difference between the two methods in terms of motion tracking performance.

Q2: The effect of the proposed framework on robustness in multi-policy transitions.

To evaluate the robustness of the control methods, we aim to determine which approach enables the robot to maintain

better motion completion when facing out-of-domain states and disturbances during policy transitions. For this purpose, we employ the $E_{succ,n}$ metric to assess whether the robot can successfully execute motion switches without falling. In this set of baseline experiments, expert networks trained using the ASAP framework are directly integrated into the pipeline to perform policy switching. For the experimental group, the pipeline incorporates the framework trained with the RPG method. Each motion transition is repeated 20 times, and the success rate is calculated.

TABLE IV

SUCCESS RATE OF TRANSITION

Start (a)	$E^{a\to b}_{succ,20}$	**Target Motion (b)**			
		Jumping	Punching	Sword Swing	Kicking
Jumping	baseline	0.25	0.15	0.15	0.05
	RPG (ours)	**0.70**	**0.85**	**0.80**	**0.75**
Punching	baseline	0.35	0.60	0.50	0.45
	RPG (ours)	**0.80**	**0.95**	**0.75**	**0.75**
Sword Swing	baseline	0.30	0.50	0.45	0.65
	RPG (ours)	**0.70**	**0.95**	**0.90**	**0.90**
Kicking	baseline	0.30	0.65	0.40	0.50
	RPG (ours)	**0.70**	**0.90**	**0.75**	**0.85**

As shown in the results Table. IV, the success rate of motion transitions is closely related to the primary body parts involved in the movements. For example, transitions between lower-body motions—such as from jumping to kicking or from jumping to jumping—exhibit the highest level of difficulty. It also shows that RPG framework significantly improves the success rate of motion transitions. This demonstrates that the RPG framework exhibits stronger robustness in scenarios involving multi-policy switching.

Q3: The effect of the proposed framework on control smoothness in multi-policy transitions.

Ensuring smooth motion execution is critically important when handling abrupt action changes or cross-policy scenarios. Our experimental design follows the same setup as described for Q2, and we select E_{maccj} as the metric for evaluation. To assess control smoothness during policy transitions, we analyze a time window spanning from 50 timesteps before to 50 timesteps after each motion transition.

TABLE V

SMOOTHNESS COMPARISON WHILE TRANSITION

Start (a)	E_{maccj}	**Target Motion (b)**			
		Jumping	Punching	Sword Swing	Kicking
Jumping	baseline	1.8520	1.9533	1.8925	1.7633
	RPG w/o Gating	1.9197	1.8649	1.9024	1.5323
	RPG (ours)	**1.5235**	**1.6421**	**1.5875**	**1.4524**
Punching	baseline	1.9233	**1.2383**	1.8738	1.9642
	RPG w/o Gating	1.8125	1.3096	1.9482	1.9773
	RPG (ours)	**1.6237**	1.3529	**1.7821**	**1.6535**
Sword Swing	baseline	1.8735	**1.8127**	1.3525	1.8239
	RPG w/o Gating	1.7769	1.8322	1.2617	1.8254
	RPG (ours)	**1.5841**	1.9031	**1.1424**	**1.5432**
Kicking	baseline	1.9637	1.9826	1.8936	1.7235
	RPG w/o Gating	1.7689	1.8236	1.9341	1.8099
	RPG (ours)	**1.6539**	**1.6724**	**1.5937**	**1.4528**

As shown in the Table. V, the smoothness of motion is highly correlated with the range of motion. For instance, transitions involving punching can achieve relatively smooth results even with the baseline method. For motions with large amplitude action such as jumping and sword-swinging, the RPG generates significantly smoother motion executions. Even in cases where the smoothness of motions generated by RPG is inferior to that of the baseline, the difference between the two is not significant. The results indicate that the RPG method yields significantly smoother control outputs across the majority of motion transitions.

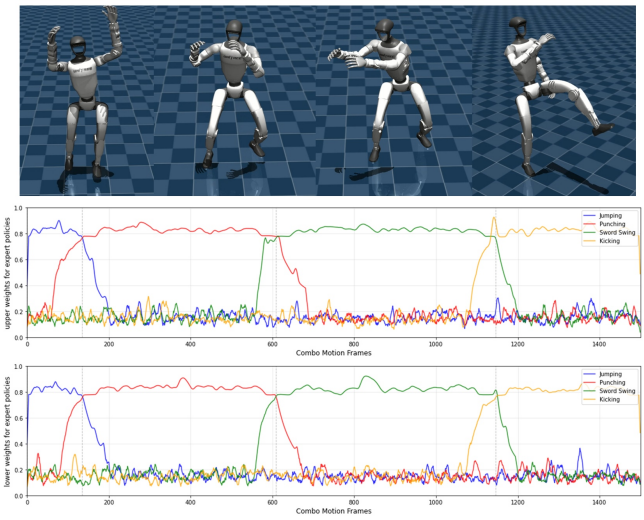

Fig. 4. Policy Transition Example: Jumping-Punching-Sword Swing-Kicking. The upper curve graph represents the variation of $\hat{w}_t^{u,m}$ along the time. The lower curve graph represents the $\hat{w}_t^{l,m}$ variation.

Q4: How does the gating policy perform in merging the expert policies?

We selected a motion transition combo as an example and recorded the gating policy output. The results are shown in Fig. 4. Since the weight values output by the gating policy account for the coordination between the upper and lower body of the robot, we recorded them separately. It can be observed that during action transitions, the weight values corresponding to the executed actions output by the gating network are more pronounced.

We also designed an ablation experiment comparing the performance with/without the gating mechanism. As shown in the Table. V, it tells that the gating policy significantly improves the control smoothness during policy transitions.

D. Discussion.

Based on the aforementioned experiments, it can be observed that the use of the RPG significantly enhances the stability and success rate of action policy transitions while improving the smoothness of robot control—all without substantially compromising the motion tracking performance of individual expert policies. This improvement can be attributed, in part, to the policy-transition and temporal randomization mechanisms, which expand the robot's state domain, and through RL reward constraints, enhance both

control smoothness and disturbance resistance in the face of abrupt motion changes.

V. CONCLUSION

This paper proposed RPG, a robust policy gating framework for multi-policies transition in humanoid fighting. RPG integrates policy-transition and temporal randomization mechanism for training with a smoothly regularized gating network, enabling stable and interruptible motion execution. Simulations and real-world tests on a Unitree G1 robot demonstrate that RPG achieves high transition success rates and improved control smoothness and robustness without compromising tracking accuracy. Meanwhile we developed a multi-policy control pipeline that enables robot combat tasks to be executed in a manner reminiscent of controlling a character in an action RPG game. Our approach provides a practical solution for dynamic multi-skill imitation in real-world humanoid applications.

In future work, we plan to integrate perception and recognition capabilities to enable the robot to autonomously track and engage targets in combat tasks.

REFERENCES

[1] I. Radosavovic, S. Kamat, T. Darrell, and J. Malik, "Learning Humanoid Locomotion over Challenging Terrain," Oct. 2024.

[2] Z. Zhuang, S. Yao, and H. Zhao, "Humanoid parkour learning," *arXiv preprint arXiv:2406.10759*, 2024.

[3] Z. Wang, J. Zhou, and Q. Wu, "Dribble Master: Learning Agile Humanoid Dribbling Through Legged Locomotion," May 2025.

[4] Y. Xue, W. Dong, M. Liu, W. Zhang, and J. Pang, "A Unified and General Humanoid Whole-Body Controller for Fine-Grained Locomotion," Feb. 2025.

[5] Y. Zhang, Z. Cao, B. Nie, H. Li, and Y. Gao, "Learning Robust Motion Skills via Critical Adversarial Attacks for Humanoid Robots," Jul. 2025.

[6] Q. Ben, F. Jia, J. Zeng, J. Dong, D. Lin, and J. Pang, "Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit," *arXiv preprint arXiv:2502.13013*, 2025.

[7] Z. Gu, J. Li, W. Shen, W. Yu, Z. Xie, S. McCrory, X. Cheng, A. Shamsah, R. Griffin, C. K. Liu, A. Kheddar, X. B. Peng, Y. Zhu, G. Shi, Q. Nguyen, G. Cheng, H. Gao, and Y. Zhao, "Humanoid Locomotion and Manipulation: Current Progress and Challenges in Control, Planning, and Learning," Jan. 2025.

[8] Y. Liu, Z. Zhang, H. Wang, and L. Yi, "Unleashing Humanoid Reaching Potential via Real-world-Ready Skill Space," May 2025.

[9] A. Schakkal, B. Zandonati, Z. Yang, and N. Azizan, "Hierarchical Vision-Language Planning for Multi-Step Humanoid Manipulation," Jun. 2025.

[10] W. Sun, L. Feng, B. Cao, Y. Liu, Y. Jin, and Z. Xie, "ULC: A Unified and Fine-Grained Controller for Humanoid Loco-Manipulation," Jul. 2025.

[11] C. Tessler, Y. Jiang, E. Coumans, Z. Luo, G. Chechik, and X. B. Peng, "MaskedManipulator: Versatile Whole-Body Control for Loco-Manipulation," May 2025.

[12] D. J. Agravante, A. Cherubini, A. Sherikov, P.-B. Wieber, and A. Kheddar, "Human-humanoid collaborative carrying," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 833–846, 2019.

[13] T. He, J. Gao, W. Xiao, Y. Zhang, Z. Wang, J. Wang, Z. Luo, G. He, N. Sobanbab, C. Pan *et al.*, "Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills," *arXiv preprint arXiv:2502.01143*, 2025.

[14] T. E. Truong, Q. Liao, X. Huang, G. Tevet, C. K. Liu, and K. Sreenath, "BeyondMimic: From Motion Tracking to Versatile Humanoid Control via Guided Diffusion," Aug. 2025.

[15] 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," Jun. 2025.

[16] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. Kitani, C. Liu, and G. Shi, "Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," *arXiv preprint arXiv:2406.08858*, 2024.

[17] T. Zhang, B. Zheng, R. Nai, Y. Hu, Y.-J. Wang, G. Chen, F. Lin, J. Li, C. Hong, K. Sreenath, and Y. Gao, "HuB: Learning Extreme Humanoid Balance," May 2025.

[18] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi, "Learning human-to-humanoid real-time whole-body teleoperation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 8944–8951.

[19] Z. Chen, M. Ji, X. Cheng, X. Peng, X. B. Peng, and X. Wang, "GMT: General Motion Tracking for Humanoid Whole-Body Control," Jun. 2025.

[20] Z. Su, B. Zhang, N. Rahmanian, Y. Gao, Q. Liao, C. Regan, K. Sreenath, and S. S. Sastry, "HITTER: A HumanoId Table TEnnis Robot via Hierarchical Planning and Learning," Aug. 2025.

[21] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Robust and versatile bipedal jumping control through reinforcement learning," *arXiv preprint arXiv:2302.09450*, 2023.

[22] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive whole-body control for humanoid robots," *arXiv preprint arXiv:2402.16796*, 2024.

[23] M. Ji, X. Peng, F. Liu, J. Li, G. Yang, X. Cheng, and X. Wang, "Exbody2: Advanced expressive humanoid whole-body control," *arXiv preprint arXiv:2412.13196*, 2024.

[24] Z. Zhuang and H. Zhao, "Embrace collisions: Humanoid shadowing for deployable contact-agnostics motions," *arXiv preprint arXiv:2502.01465*, 2025.

[25] Z. Zhuang, Z. Fu, J. Wang, C. Atkeson, S. Schwertfeger, C. Finn, and H. Zhao, "Robot parkour learning," *arXiv preprint arXiv:2309.05665*, 2023.

[26] A. Allshire, H. Choi, J. Zhang, D. McAllister, A. Zhang, C. M. Kim, T. Darrell, P. Abbeel, J. Malik, and A. Kanazawa, "Visual Imitation Enables Contextual Humanoid Control," May 2025.

[27] Y. Wang, M. Yang, W. Zeng, Y. Zhang, X. Xu, H. Jiang, Z. Ding, and Z. Lu, "From Experts to a Generalist: Toward General Whole-Body Control for Humanoid Robots," Jun. 2025.

[28] 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 *8th Annual Conference on Robot Learning*, 2024.

[29] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," *arXiv preprint arXiv:2406.10454*, 2024.

[30] W. Yu, F. Acero, V. Atanassov, C. Yang, I. Havoutis, D. Kanoulas, and Z. Li, "Discovery of skill switching criteria for learning agile quadruped locomotion," Feb. 2025.

[31] X. B. Peng, Y. Guo, L. Halper, S. Levine, and S. Fidler, "Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters," *ACM Transactions On Graphics (TOG)*, vol. 41, no. 4, pp. 1–17, 2022.

[32] N. S. Neggatu, J. Houssineau, and G. Montana, "Evaluation-Time Policy Switching for Offline Reinforcement Learning," Mar. 2025.

[33] D. K. Panda and W. Guo, "Robust Policy Switching for Antifragile Reinforcement Learning for UAV Deconfliction in Adversarial Environments," Jun. 2025.

[34] G. Christmann, Y.-S. Luo, and W.-C. Chen, "Expert Composer Policy: Scalable Skill Repertoire for Quadruped Robots," Mar. 2024.

[35] T. Huang, J. Ren, H. Wang, Z. Wang, Q. Ben, M. Wen, X. Chen, J. Li, and J. Pang, "Learning Humanoid Standing-up Control across Diverse Postures," Feb. 2025.

[36] Z. Wang, X. Yang, J. Zhao, J. Zhou, T. Ma, Z. Gao, A. Ajoudani, and J. Liang, "End-to-End Humanoid Robot Safe and Comfortable Locomotion Policy," Aug. 2025.

[37] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5442–5451.

[38] Z. Shen, H. Pi, Y. Xia, Z. Cen, S. Peng, Z. Hu, H. Bao, R. Hu, and X. Zhou, "World-Grounded Human Motion Recovery via Gravity-View Coordinates," in *SIGGRAPH Asia 2024 Conference Papers*, Dec. 2024, pp. 1–11.

[39] Z. Luo, J. Cao, A. Winkler, K. Kitani, and W. Xu, "Perpetual Humanoid Control for Real-time Simulated Avatars," Sep. 2023.