Proc of the 2026 IEEE International
Conference on Real-time Computing and Robotics
August 15-21, 2026, Seychelles

PhysiFlow: Physics-Aware Humanoid Whole-Body VLA via Multi-Brain Latent Flow Matching and Robust Tracking

Weikai Qin*, Sichen Wu*, Ci Chen, Mengfan Liu, Linxi Feng, Xinru Cui, Haoqi Han, Hesheng Wang†

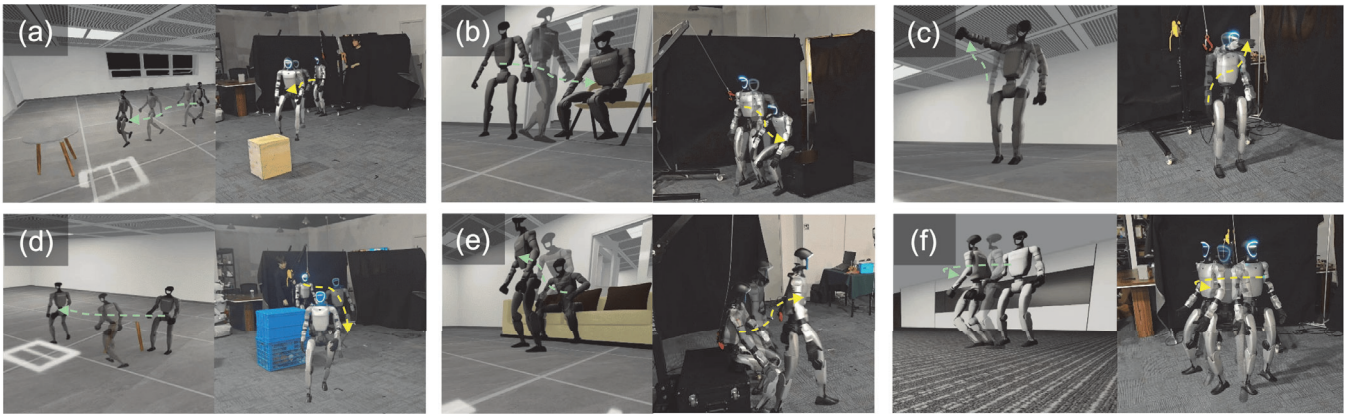

Fig. 1: Introducing **PhysiFlow**, a multi-brain VLA humanoid system that operates on Unitree G1 robots and performs end-to-end VLA humanoid whole body control in large spaces. The proposed system achieves consecutive tasks autonomously, including **(a-c)** walking to the designated item, sitting on the proposed item, and raising arm; **(d-f)** circling the designated item, standing up from the specific item and turning right.

***Abstract*— In the domain of humanoid robot control, the fusion of Vision-Language-Action (VLA) with whole-body control is essential for semantically guided execution of real-world tasks. However, existing methods encounter challenges in terms of low VLA inference efficiency or an absence of effective semantic guidance for whole-body control, resulting in instability in dynamic limb-coordinated tasks. To bridge this gap, we present a semantic-motion intent guided, physics-aware multi-brain VLA framework for humanoid whole-body control. A series of experiments was conducted to evaluate the performance of the proposed framework. The experimental results demonstrated that the framework enabled reliable vision-language-guided full-body coordination for humanoid robots.**

I. INTRODUCTION

Humanoid robots, which possess a similar physical structure to humans, have the potential to execute a variety of tasks and motions that are prevalent in human daily life. Nowadays, great efforts have been made to enhance their locomotion and manipulation capabilities: advances in learning-based [1], [2] and model-based [3] whole-body control methodologies have enabled humanoid robots to perform a diverse range of tasks with high flexibility and precision in execution [4], [5]. Meanwhile, VLA models have gained widespread application in robotic manipulation [6]. By unifying visual perception, language understanding, and action generation within an end-to-end framework, they eliminate the interface barriers of traditional modular systems [7]. Besides, they endow robots with intuitive human-robot interaction and autonomous semantic task understanding capabilities.

The fundamental challenge encountered in the field of humanoid VLA is the integration of semantic perception, high-frequency motion generation, and physics-aware control in existing frameworks. Traditional VLA-driven systems suffer from low inference efficiency and poor adaptability with edge deployment [8], [9], while learning-based whole-body control methods often lack effective guidance from vision-language semantics [10], [11]. Consequently, humanoid robots often exhibit instability and task failure in vision-and-language-guided dynamic whole-body control scenarios that demand simultaneous upper and lower limb coordination and balance maintenance [12]. To address this problem, we present a semantic-motion intent guided, physics-aware multi-brain VLA framework for humanoid whole-body control, which enables reliable full-body motion performance with real-time inference and physical stability via semantic-motion integration.

The key innovation of this study lies in the bio-inspired hierarchical multi-brain architecture, which is centred on the semantic latent vector. This decouples and coordinates three functional brains for the purpose of humanoid whole-body

Weikai Qin, Sichen Wu, Ci Chen, Mengfan Liu, Linxi Feng, Xinru Cui and Haoqi Han is with the department of Automation Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China (e-mail: LingMisaki@sjtu.edu.cn, 15182340716@163.com, 21813007@zju.edu.cn, mengfan_liu@sjtu.edu.cn, linxif2008@sjtu.edu.cn, cxr0726@sjtu.edu.cn and hhq123@sjtu.edu.cn)

*Equal contribution

†Hesheng Wang is with the Department of Automation, Key Laboratory of System Control and Information Processing of Ministry of Education, Key Laboratory of Marine Intelligent Equipment and System of Ministry of Education, Shanghai Engineering Research Center of Intelligent Control and Management, Shanghai Jiao Tong University, Shanghai, 200240, China (e-mail: wanghesheng@sjtu.edu.cn)

2026 IEEE International Conference on Real-time Computing and Robotics (RCAR) | 979-8-3195-2313-6/26/$31.00 ©2026 IEEE | DOI: 10.1109/RCAR70436.2026.11765856

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on November 20,2026 at 04:37:04 UTC from IEEE Xplore. Restrictions apply.

control. The Neocortical Brain fuses task semantics of what to do and motion intent of how to do via a curriculum-based conditional variational autoencoder (CVAE) based on SigLIP [13] with LoRA lightweight adaptation. The Basal Ganglionic Brain generates $50\,\mathrm{Hz}$ high-frequency motion sequences through latent vector-driven flow matching. The Cerebellar Brain is a motion tracker that enforces physics-aware constraints via joint fine-tuning and tracking error backpropagation. This design successfully overcomes the conventional conflict between inference efficiency, semantic generalization, and dynamic balance in VLA systems, effectively adapting to full-body tasks guided by vision and language that require upper and lower limb coordination.

The proposed framework is validated on the Unitree G1 humanoid robot through a combination of simulation and real-world experimentation. A series of tests are conducted in order to evaluate whole-body control scenarios. These include the ability to sit on or circle a specific item, which is shown in Fig. 1. In the context of simulation, the proposed method has been shown to outperform existing baseline approaches with regard to task success rate. The reliability of the system is further demonstrated by real-world experiments. The key contributions are summarized as follows:

1) A novel bio-inspired multi-brain VLA framework is proposed, decoupling high-level semantic-motion intent inference from low-level high-frequency motion generation and stable tracking control for humanoid whole-body control.
2) In Neocortical Brain, a two-phase CVAE curriculum based on SigLIP with LoRA lightweight adaptation is employed to generate the modality-invariant semantic latent vector with motion intent.
3) Driven by latent vector, we introduce a physics-aware flow-matching training paradigm by fusing motion tracking with joint fine-tuning for dynamic and consistent motion generation.

II. RELATED WORK

A. VLA for Robotics Learning

VLA models have become the mainstream of end-to-end robotic learning, which unify multi-modal perception and action generation to break the interface barriers of traditional modular systems [14]. For fixed-base dual-arm manipulation, representative works like π_0 [15] and CO-RFT [16] have achieved excellent generalization and trajectory quality, but they pose significant challenges when migrated to humanoid robots: insufficient inference speed for high-frequency full-body control [17], and excessive computing overhead that exceeds on-board edge device limits [18]. To solve these issues, recent humanoid VLA works widely adopt dual-brain decoupled architectures, such as Galaxea G0 [19] and LeVERB [20], which effectively improve inference efficiency [21]. However, existing evaluations are mostly limited to simple tabletop tasks, lacking physics-aware stable control for complex whole-body humanoid motion with coordinated limb movement [22]. In this work, we propose a multi-brain VLA framework to fill this gap, enabling robust whole-body control for humanoid robots in complex scenarios.

B. Humanoid Whole-Body Control

Whole-body control of humanoid robots remains a formidable challenge, due to their highly nonlinear dynamics and multi-joint coordination requirements [1], [2]. Although traditional model-based approaches [3], [4] are effective for basic motions in structured environments, it is often difficult to balance flexibility and stability in complex scenarios. Recent advances in learning-based control, such as RL and teacher-student frameworks, have demonstrated significant progress in flexible motions and multiple tasks [5], [10], [11]. TWIST leverages an RL and behaviour cloning (BC) teacher-student pipeline with real-time motion capture (MoCap) retargeting, enabling it to execute coordinated skills like whole-body manipulation and locomotion [23], while ASAP proposes a two-stage framework with a delta action model and asymmetric actor-critic training to compensate sim2real physics mismatch [24]. However, these methods fundamentally rely on motion tracking and teleoperation, lacking autonomous comprehension of vision input and language guidance [25]. The existence of these high-level perceptual decision-making abilities is of critical importance for humanoid robots to transition into domestic service scenarios, which represents the primary gap in current research.

III. METHOD

A. Multi-brain Architecture

We propose a bio-inspired hierarchical multi-brain VLA framework which is illustrated in Fig. 2 for humanoid whole-body control, focusing on semantic-motion intent alignment, efficient motion generation, and physics-aware dynamic action execution. The full framework is described as follows.

1) Neocortical Brain: Semantic-Motion Intent Alignment

The Neocortical Brain Module employs a two-stage curriculum-based CVAE as its VLA prior model. Inspired by LeVERB [20]'s residual CVAE architecture, we introduce the latent vector z_{vl} as the primary training objective. Its core function is to integrate the semantic instruction of what to do with the motion intent of how to do. Furthermore, in the inference stage, it can generate $10\,\mathrm{Hz}$ semantic-motion intent z_{vl} directly from current vision input and language guidance alone, without knowing any future motion sequences and the current state, making it ideal for real-time control scenarios. The detailed structural design and training specifics of z_{vl} are elaborated on in subsequent sections.

2) Basal Ganglionic Brain: Semantic-Motion Intent Driven Flow Matching

Inspired by the flow matching mechanism in $\pi_{0.5}$ [14], this module innovatively replaces traditional direct vision-language encoders with the $10\,\mathrm{Hz}$ latent vector z_{vl}, which is trained in Neocortical Brain. Along with the robot's current state, it forms the input to the flow-matching model, enabling

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on November 20,2026 at 04:37:04 UTC from IEEE Xplore. Restrictions apply.

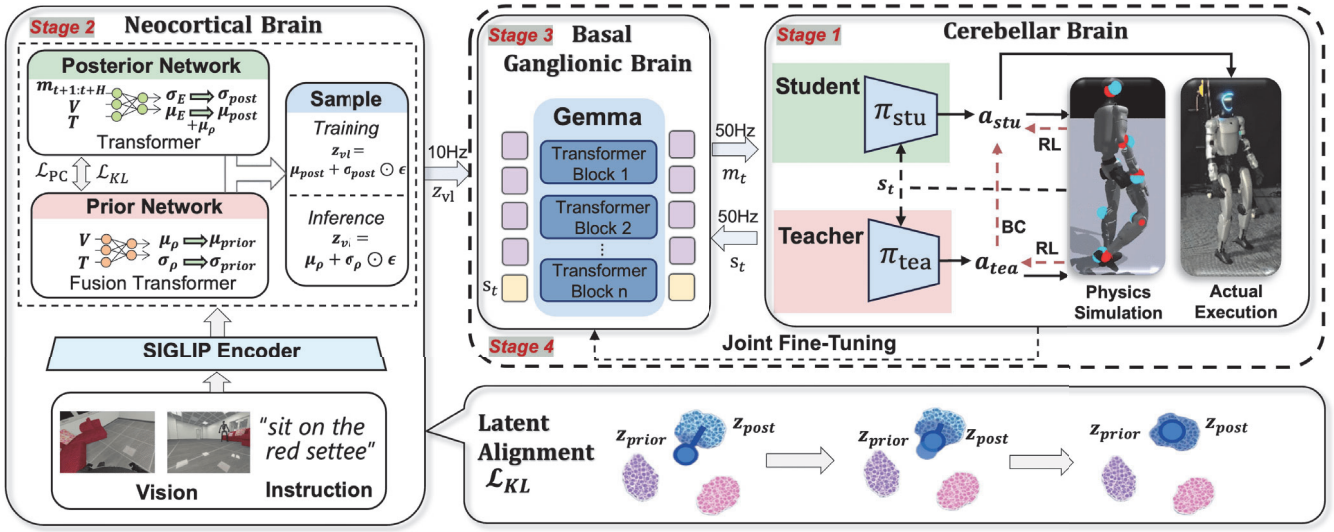

Fig. 2: **The overall pipeline of PhysiFlow.** This bio-inspired architecture decouples semantic reasoning from physics-aware execution. **(a) Neocortical Brain:** A curriculum-based CVAE processes vision and language to synthesize a 10 Hz latent vector z_{vl}, aligning task semantics with motion intent. **(b) Basal Ganglionic Brain:** Conditioned on z_{vl} and robot states, a flow-matching model generates 50 Hz motion sequence m_t for continuity. **(c) Cerebellar Brain:** A robust motion tracker enforces physical constraints, translating these chunks into stable motor commands for closed-loop whole-body control.

continuous motion sequence generation guided by both semantic and motion-trend intent information. This approach ensures generated motion sequences not only meet task semantic requirements, but also inherit logical consistency and continuity in motion naturally. The model uses the low-parameter action expert network Gemma [26], which leverages the capability of flow matching to generate high-frequency motion sequence chunks. This ensures 50 Hz real-time inference performance, which is ideal for the high-frequency feedback demands of full-body control in humanoid robots. During training, the model learns flow field mappings from noisy motion sequences to real motion sequences. It uses a beta distribution time-step sampling strategy ($\alpha = 1.5$, $\beta = 1$) to enhance training stability.

3) Cerebellar Brain: Motion Tracking with Joint Fine-Tuning

The motion tracker is used in this module to achieve stable, high-precision tracking of input motion sequences. The pre-trained motion tracking model is optimised using a combination of RL and teacher-student learning strategies, while domain randomisation enhances sim2real capabilities. During deployment, the motion tracking module receives 50 Hz motion sequence inputs. These are converted by an internally running 1000 Hz PD controller into executable motor commands for the Unitree G1 humanoid robot, achieving closed-loop control from semantic intent to physical action a_t.

During training, a sequential optimization pipeline is employed to ensure physically viable, high-frequency control. As illustrated in Fig. 2, the process is divided into four distinct stages. In stage 1, we first pre-train the Cerebellar Brain on a physics-informed motion dataset to build a robust motion tracker. Then, in stage 2, the Neocortical Brain CVAE is trained via two-phase curriculum learning with staged loss and LoRA adaptation. For the stage 3, the Basal Ganglionic flow-matching module is trained with a Gemma [26] decoder through sampling, path building, and the MSE loss minimization process. Finally in stage 4, we propose a joint fine-tuning strategy for the flow matching model and the frozen motion tracking module. The error between executed and reference actions from the two sequential modules is backpropagated to fine-tune the flow matching model, ensuring its generated motions are physics-aware and dynamically consistent with motion tracking constraints. This hierarchical paradigm dynamically bridges abstract semantic reasoning with stable physical execution.

B. The design of the Neocortical Brain

To achieve multi-modal semantic alignment across vision, language, and motion intent, we build our model upon a CVAE framework with dedicated feature encoding and fusion modules. The vision encoder employs a pre-trained SigLIP ViT-B/16 model, which processes images from the first-person view V_t^{ego} and the third-person view V_t^{exo} to comprehensively capture scene information. The text encoder also originates from SigLIP, leveraging its contrastive pre-training advantage to encode text instructions T into semantically consistent textual features. The two feature types are further integrated via a Fusion Transformer to generate a unified vision-language representation, which is projected to parameterize the prior and posterior distributions of the latent verb z_{vl}.

Inspired by the residual CVAE design in LeVERB [20], we construct the prior and posterior networks as follows. The prior network $p_\rho(\mu_\rho, \sigma_\rho^2|V,T)$ of the CVAE outputs the mean μ_ρ and diagonal variance σ_ρ^2, modeling scene semantics and task objectives, where $\mu_{prior} = \mu_\rho$, $\sigma_{prior}^2 = \sigma_\rho^2$. The posterior network $q_E(\mu_E, \sigma_E^2|V,T,m_{t+1:t+H})$ additionally

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on November 20,2026 at 04:37:04 UTC from IEEE Xplore. Restrictions apply.

takes the privileged future motion sequences $m_{t+1:t+H}$ (only accessible during training) as input, predicting mean μ_E and diagonal variance σ_E^2 to capture motion intent details that cannot be inferred from vision-language inputs alone.

During training, we construct the posterior distribution via residual fusion to decouple semantic encoding and motion detail completion: the posterior variance $\sigma_{post}^2 = \sigma_E^2$ is directly inherited from the posterior network, and the posterior mean is defined as $\mu_{post} = \mu_\rho + \mu_E$. The standard reparameterization trick is applied for gradient backpropagation with stage-specific sampling rules: in the training stage, the 256-dimensional latent verb z_{vl} is sampled from the posterior distribution as $z_{vl} = \mu_{post} + \sigma_{post} \odot \epsilon$ to leverage privileged motion information, where $\epsilon \sim \mathcal{N}(0, I)$. At inference, with no access to future motion sequences, we sample z_{vl} directly from the prior network's output distribution as $z_{vl} = \mu_\rho + \sigma_\rho \odot \epsilon$.

The total loss function focuses on core CVAE optimization objectives while integrating our proposed semantic alignment and regularization terms, expressed as

$$\mathcal{L} = \mathcal{L}_{Recon} + \beta\mathcal{L}_{KL} + \mathcal{L}_{PC} + \mathcal{L}_{VL} + \mathcal{L}_{Aux} + \lambda_{Disc}\mathcal{L}_{Disc} \tag{1}$$

The specific definition of each loss term is as follows:

$$\mathcal{L}_{Recon} = \sum_{t=1}^{H} w_t \cdot \|\hat{m}_t - m_t\|_2^2 \tag{2}$$

$$\mathcal{L}_{KL} = D_{KL}\big(\mathcal{N}(\mu_{post}, \sigma_{post}^2) \parallel \mathcal{N}(\mu_\rho, \sigma_\rho^2)\big) \tag{3}$$

$$\mathcal{L}_{PC} = \sum_{t=1}^{H} \|\hat{m}_t^{\mu_{prior}} - \hat{m}_t^{\mu_{post}}\|_2^2 \tag{4}$$

$$\mathcal{L}_{VL} = -\log \frac{\exp(\cos(\mu_\rho, t^+)/\tau)}{\sum_j \exp(\cos(\mu_\rho, t_j)/\tau)} \tag{5}$$

where $\mathcal{L}_{Recon}$ is the time-step weighted reconstruction loss on the motion increment m_t, emphasizing short-term motion intent. $\mathcal{L}_{\text{KL}}$ aligns the posterior $\mathcal{N}(\mu_{\text{post}}, \sigma_{post}^2)$ with the prior $\mathcal{N}(\mu_\rho, \sigma_\rho^2)$ to regularize latent space, ensuring the prior network approximates the posterior's motion encoding capability for inference. $\mathcal{L}_{PC}$ distills posterior knowledge by aligning $\hat{m}_t^{\mu_{prior}}$ and $\hat{m}_t^{\mu_{post}}$, enabling the prior to mimic the posterior's motion intent at inference without access to future sequences, where $\hat{m}_t^{\mu_{prior}}$ is motion reconstructed from the prior mean μ_ρ under inference-like conditions, and $\hat{m}_t^{\mu_{post}}$ is motion reconstructed from the posterior mean μ_{post} with privileged future information. $\mathcal{L}_{VL}$ applies InfoNCE alignment between μ_ρ and text embeddings to bind the latent prior to language semantics and improve inference-time retrieval performance, while the t^+ is the matched text embedding for the current sample, t_j denotes the candidate text embeddings in the contrastive set. $\mathcal{L}_{Disc}$ is a modality-adversarial loss with a Gradient Reversal Layer (GRL) that enforces modality invariance of z_{vl} for enhanced cross-modal robustness, which is a mature paradigm in the field and thus no formulation is needed. $\mathcal{L}_{Aux}$ aggregates auxiliary objectives including instruction classification, z-use margins, σ_ρ regularization, which are all basic optimization terms and no separate formulation is needed.

We adopt a two-phase curriculum training strategy: In the first phase, we prioritize posterior convergence by optimizing only $\mathcal{L}_{Recon}$ and $\mathcal{L}_{KL}$, which ensures accurate motion intent capture and prevents posterior collapse caused by premature alignment to the prior. The second phase gradually introduces $\mathcal{L}_{PC}$, $\mathcal{L}_{VL}$ and $\mathcal{L}_{Aux}$ through weight warming to enable efficient knowledge distillation from posterior to prior. Furthermore, lightweight parameter adaptation with LoRA is implemented, which involves freezing the SigLIP encoder and training only low-rank increments. This design allows z_{vl} to effectively fuse cross-modal semantics with motion trend logic, thereby enhancing semantic-motion intent adaptability. This design lays a robust foundation for intent-driven flow matching in the subsequent Basal Ganglionic Brain and motion tracking in the Cerebellar Brain.

C. The design of the Basal Ganglionic Brain

The Basal Ganglionic Brain is designed to bridge low-frequency semantic-motion intent from the Neocortical Brain and high-frequency humanoid stable control from Cerebellar Brain. To solve low inference frequency and cumulative error of autoregressive VLA models, we adopt conditional flow matching for real-time continuous motion generation. It takes dual conditional inputs including $10\,\text{Hz}$ semantic latent vector $z_{vl} \in \mathbb{R}^{256}$ from the Neocortical Brain and 38-dimensional Unitree G1 robot state $s_t \in \mathbb{R}^{38}$ including 3D root position, 6D orientation and 29 joint angles, outputting $50\,\text{Hz}$ motion sequence chunks $M_t = [m_{t,1}, \ldots, m_{t,l}]$ with $l = 10$ and stride=2, which ensures temporal continuity and smoothness of motion.

A lightweight single-stream Gemma [26] decoder is adopted for flow matching. This unifies the modelling of the token sequence $[z_{\text{vl}}, s_t, m_t]$, while motion tokens are generated by projecting the noise-data inference x_t with time embeddings and the decoder exclusively outputs the vector field v_θ for motion sequence tokens. During training, time steps $t \sim \text{Beta}(\alpha = 1.5, \beta = 1)$ and Gaussian noise $\varepsilon \sim \mathcal{N}(0, I)$ are sampled to construct the linear path $x_t = t \cdot \varepsilon + (1 - t) \cdot m_t$, with the objective to minimize the masked mean squared error $\|v_\theta(x_t, t, z_{\text{vl}}, s_t) - (\varepsilon - m_t)\|^2$. For inference, the model operates at 10 Hz, generating action chunks of length 10 via reverse ODE integration (Euler/Heun solver) from $t = 1$ to $t = 0$. To reduce latency and improve smoothness, only the first five frames of each generated chunk are fed into the downstream Cerebellar Brain. By executing these 5 frames per 10 Hz inference cycle, the system achieves an effective high-frequency control rate of 50 Hz. Optional weight-based fusion can also be applied across these overlapping frames to further improve temporal consistency.

D. The design of the Cerebellar Brain

The Cerebellar Brain acts as a robust motion tracking module, translating high-frequency motion sequence chunks into stable, physics-aware whole-body control. Inspired by the

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on November 20,2026 at 04:37:04 UTC from IEEE Xplore. Restrictions apply.

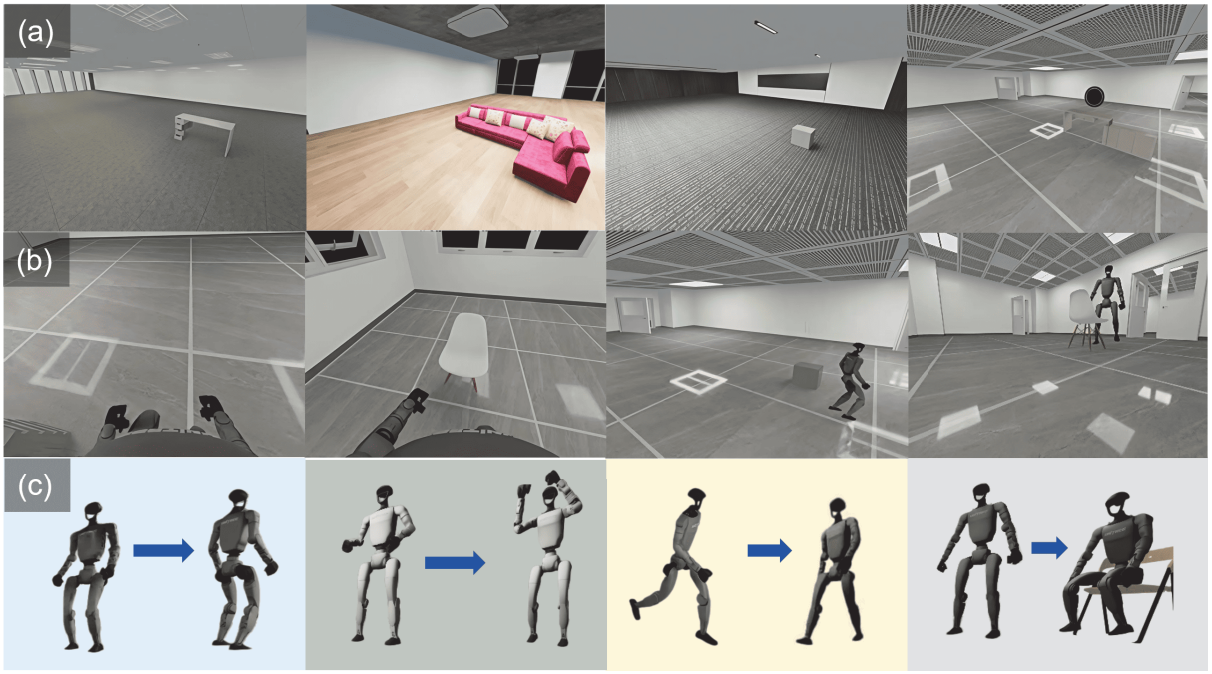

Fig. 3: **Visualization of the VLA dataset. (a)** Diverse visuals with various Scenes and Items. **(b)** Diverse camera angles with ego and exo views. **(c)** Diverse task from turning around to standing up

TWIST [23] architecture, we use a two-stage teacher-student RL framework to address the instability and conservative behavior common in single-stage policies. First, a privileged teacher policy learns smooth, coordinated whole-body movements via RL by utilizing future reference motions. Next, a deployable student policy learns from the teacher through a combination of RL and behavior cloning (BC), relying solely on real-time proprioceptive feedback and current reference frames.

After the initial pre-training on a physics-informed motion dataset, we fine-tune the student policy on our own task-specific dataset. This step improves dynamic consistency and physical realism, helping to bridge the sim-to-real gap and ensuring the robot can reliably perform complex tasks.

E. VLA Data Generation

In order to generate data for VLA humanoid whole-body control effectively, a series of practical methods were conducted in Isaac Lab [27]. Initially, we deployed Pico devices to remotely capture multiple motion sequences, including specific tasks such as sitting down on specified chairs, moving to a target location. Subsequently, the trained Cerebellar Brain's motion tracker is used to output physics-aware data that is dynamically consistent during actual execution. Furthermore, several scenarios were constructed in Isaac Lab, including laboratories and living rooms equipped with chairs, tables, and other relevant objects.

Within the Isaac Lab simulation environment, motion sequences captured from the experiment were replayed using Cerebellar Brain's motion tracker, while real-time images were captured from both first-person and third-person perspectives. In order to maximise data diversity, the objects and backgrounds in the simulation were automatically replaced during each replay of the motion sequences. The diversity of the VLA datasets can be seen in the Fig. 3.

IV. EXPERIMENT

A. The ablation of the Neocortical Brain

A comprehensive ablation study was conducted to validate the contribution of each core component, with one component ablated at a time. All experimental settings including the training schedule, hyperparameters, data preprocessing and dataset splits were fixed to ensure a fair comparison. All models were evaluated with a batch size of 48 and 16 evaluation batches. Ablated models are evaluated across three dimensions:

- **Reconstruction fidelity**: Measured by Recon. Post and Recon. Prior, quantifying trajectory prediction accuracy under posterior- and prior-conditioned inference.
- **Latent regularization & information utilization**: Evaluated via z-Use Gap, Text-Use Gap (Zero), Future-Shuffle Gap, and KL-CVAE, measuring latent usage, conditioning effectiveness, posterior dependence on future information, and latent distribution alignment.
- **Semantic alignment**: Assessed by Retrieval Top-1, Retrieval (Cross Ep.) and Retrieval (Same Task), quantifying latent-language matching in standard, cross-episode, and task-specific scenarios.

The specific ablation operations for each component are listed as follows:

- **VL alignment (VL off):** Ablated by setting all VL and VL-cond weights to zero.
- **PC distillation (PC Off):** Ablated by clearing PC-related weights.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on November 20,2026 at 04:37:04 UTC from IEEE Xplore. Restrictions apply.

TABLE I: Comparison of core metrics across different ablation settings.Best results are viewed in **bold black**.

Metric Name	Full	PC Off	Disc Off	VL Off	Lora Off	Textcond0	Curr. Off
Recon. Post	**0.017**	0.020	0.019	0.020	0.020	0.041	0.081
Recon. Prior	**0.023**	0.027	0.026	0.026	0.029	0.045	0.081
Future Shuffle Gap	**1.134**	0.679	0.852	0.940	0.958	0.516	0.001
KL CVAE	**36.095**	36.369	29.632	26.835	29.016	38.783	9.100
Retrieval (Cross Ep.)	**0.859**	0.879	0.872	0.037	0.869	0.846	0.846
Retrieval (Same Task)	**0.357**	0.364	0.365	0.086	0.362	0.349	0.352
Retrieval Top-1	**0.357**	0.363	0.361	0.016	0.361	0.349	0.350
Text-Use Gap (Zero)	**1.921**	1.119	1.546	3.706	4.642	0.454	1.319
z-Use Gap	**0.921**	1.030	0.885	0.723	0.763	0.784	3.397

- **Disc regularization (Disc Off):** Ablated by turning off the discriminator.
- **LoRA adaptation (Lora Off):** Ablated by disabling LoRA for vision and text encoders.
- **Text conditioning (Textcond0):** Ablated by setting its scale to zero.
- **Two-phase curriculum strategy (Curr. Off):** Ablated by adopting a uniform loss weighting scheme during training.

As demonstrated in Table I, the ablation results reveal the contribution of each component. For VL Off, its ablation induces substantial retrieval collapse, evidenced by Retrieval Top-1 declining from 0.357 to 0.016 and Retrieval (Cross Ep.) from 0.859 to 0.037, which proves its essential role in semantic grounding. For Textcond0, ablating it significantly reduces the Text Use Gap (Zero) from 1.921 to 0.454 and degrades retrieval performance, verifying language's function in shaping intent. PC Off causes minor retrieval changes but elevates reconstruction errors, drops Future Shuffle Gap and raises KL CVAE and z-Use Gap, degrading prior-posterior alignment. Disc Off barely affects metrics but reduces cross-modal performance by aligning vision-language and language-only latent spaces for scenarios without visual input. Lora Off leads to a moderate retrieval decline with consistent non-dominant gains in most metrics. Curr. Off drastically reduces Future Shuffle Gap to 0.001 and worsens reconstruction metrics, validating staged training. Overall, our full model balances semantic alignment, prior usability and latent stability, outperforming all ablated variants.

B. The ablation of the Basal Ganglionic Brain

The inference efficiency and motion smoothness of the Basal Ganglionic Brain are evaluated against autoregressive (AR) [14] and Denoising Diffusion Probabilistic Model (DDPM) [28] baselines as follows:

- **AR:** Generates motion sequences in a step-by-step autoregressive manner, with each frame dependent on the generation results of previous frames.
- **DDPM:** Generates valid motion sequences by gradually denoising random noise signals through a multi-step reverse diffusion process.

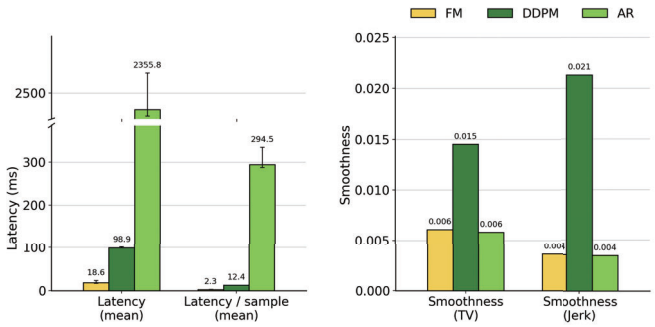

Fig. 4: **Performance benchmarking of the Basal Ganglionic Brain.** The proposed flow-matching (FM) paradigm is evaluated against autoregressive (AR) and Denoising Diffusion Probabilistic Model (DDPM) baselines.

As shown in Fig. 4, a mean latency of 18.65 ms and per-sample latency of 2.33 ms are achieved by our flow-matching (FM) approach, which is 5.3 times faster than DDPM and 126 times faster than AR. In terms of motion smoothness, a total variation of 0.0061 and jerk of 0.0036 are yielded by FM, matching the high smoothness of AR while outperforming DDPM by a significant margin. These results demonstrate that the FM paradigm effectively resolves the trade-off between inference speed and motion quality, enabling real-time $50\,\mathrm{Hz}$ high-frequency motion generation for humanoid whole-body control.

C. Simulation Experiments

To evaluate the proposed PhysiFlow framework, we conducted comprehensive simulation experiments using the Unitree G1 robot in Isaac Lab, with all inference executed on a single 48GB RTX 4090 GPU. We designed nine representative whole-body tasks, categorized into complex coordinated motions and regular basic tasks. LeVERB was selected as the primary baseline and locally reproduced within our environment to ensure a fair comparison. Task success rate serves as the core quantitative metric.

Table II presents the experimental results. While the reproduced baseline performs well on basic motions and short-distance navigation, PhysiFlow outperforms it in most categories, especially in tasks requiring continuous coordi-

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on November 20,2026 at 04:37:04 UTC from IEEE Xplore. Restrictions apply.

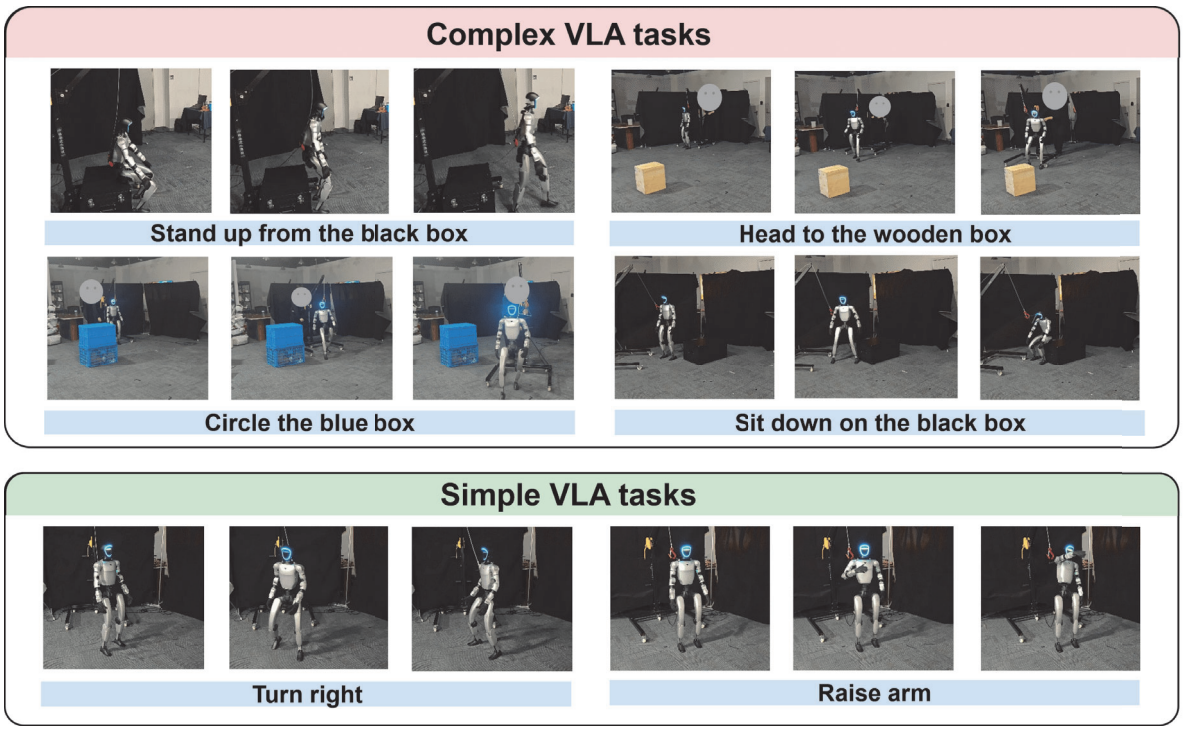

Fig. 5: **Real-world execution of semantically guided whole-body tasks by the Unitree G1 humanoid robot. Top:** Complex VLA maneuvers requiring continuous spatial navigation and dynamic multi-limb coordination. **Bottom:** Basic VLA tasks demonstrating responsive semantic execution and robust postural stability. These results validate the system's capacity to maintain physical compliance and dynamic consistency during unconstrained deployment.

TABLE II: Comparison of Success Rates (%) between LeVERB and PhysiFlow. Best results are viewed in **bold black**.

Task Type	Task Name	LeVERB [20]	PhysiFlow
Complex	Nav. (Long)	31.2	**63.6**
	Nav. (Short)	**78.5**	71.7
	Nav. & Sit	5.8	**18.1**
	Nav. & Circle	54.5	**69.2**
Regular	Stand up	88.6	**90.9**
	Locomotion	97.2	**100.0**
	turn around	68.7	**78.6**
	turn left	81.3	**81.8**
	raise arm	79.1	**100.0**
Average	Overall	65.0	**74.9**

nation. Overall, PhysiFlow achieves an average success rate of 74.9%, exceeding LeVERB's 65.0% by 9.9 percentage points. The proposed framework delivers major improvements in complex tasks. For example, it increases the success rate of the "Navigation & Circle" task from 54.5% to 69.2%, and more than doubles it for "Nav. (Long)" from 31.2% to 63.6%. These results clearly demonstrate that our multi-brain VLA architecture successfully combines semantic guidance with fast inference and stable motion.

D. Real-World Experiments

We deploy PhysiFlow on the physical Unitree G1 humanoid robot to validate real-world feasibility, motion stability, and semantic consistency. In this paper, we evaluate the hardware performance via open-loop execution of generated motion sequences on the robot, which provides a controlled and reproducible setting for assessing sim-to-real transfer of whole-body behaviors.

As shown in Fig. 5, the Unitree G1 robot completed various semantically guided tasks via the PhysiFlow architecture, proving its reliable sim-to-real transfer ability. The system achieved smooth movements, strong limb coordination in complex VLA tasks and excellent dynamic stability for basic commands, which confirms our flow-matching paradigm generates physically consistent motion sequence chunks. More importantly, PhysiFlow effectively bridges the divide between high-level multimodal semantic reasoning and low-level physical execution.

V. CONCLUSIONS

The paper introduces PhysiFlow, a physics-aware policy for humanoid whole-body VLA control, with a bio-inspired multi-brain architecture. This architecture comprises the Neocortical Brain, which fuses semantic-motion intent via curriculum-based CVAE, the Basal Ganglionic Brain, which generates high-frequency motion sequences through latent vector-driven flow matching, and the Cerebellar Brain, which enforces physics-aware constraints via joint fine-tuning. By

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on November 20,2026 at 04:37:04 UTC from IEEE Xplore. Restrictions apply.

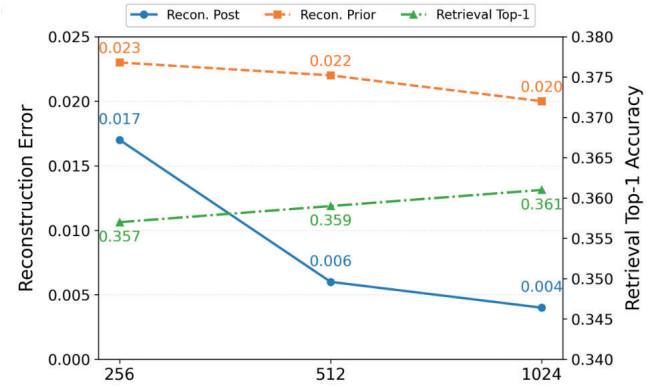

Fig. 6: **Performance metrics across different latent dimensions including 256, 512 and 1024.** The left axis tracks the posterior and prior reconstruction errors, while the right axis displays the Top-1 retrieval accuracy.

decoupling high-level semantic inference from low-level motion generation and stable tracking, the proposed framework resolves the trade-off between inference efficiency, semantic generalization, and dynamic balance in humanoid VLA systems. Experiments on Unitree G1 robots demonstrate significant improvements in task success rates and motion stability, and ablation studies validate the contribution of each core component. Future work may involve integrating world models and VLA into full-body control for humanoid robots, thereby addressing current challenges such as limited datasets, jitter and low success rates.

APPENDIX

A. Sensitivity Analysis of Latent Dimension

The dimensionality of the semantic latent vector z_{vl} directly affects the representation capability and inference efficiency of our multi-brain framework. To determine the optimal configuration, we conducted a sensitivity analysis comparing three candidate dimensions.

As shown in Fig. 6, increasing the latent dimension yields no significant improvement in core metrics including reconstruction error and retrieval accuracy. Meanwhile, higher dimensions introduce extra parameter overhead, increase memory consumption, and reduce inference speed. Importantly, restricting the dimension is not a compromise for low-cost computational solutions, but a strict necessity to meet the high-frequency inference requirements of humanoid robots. We thus select the 256-dim configuration for the final design, achieving the optimal balance between performance and real-time execution.

REFERENCES

[1] X. Cheng, Y. Ji, Chen, *et al.*, "Expressive whole-body control for humanoid robots," *arXiv preprint arXiv:2402.16796*, 2024.

[2] Z. Fu, Q. Zhao, Wu, *et al.*, "Humanplus: Humanoid shadowing and imitation from humans," *arXiv preprint arXiv:2406.10454*, 2024.

[3] J. Ramos and S. Kim, "Dynamic locomotion synchronization of bipedal robot and human operator via bilateral feedback teleoperation," *Science Robotics*, vol. 4, no. 35, p. eaav4282, 2019.

[4] Z. Gu, J. Li, Shen, *et al.*, "Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning," *arXiv preprint arXiv:2501.02116*, 2025.

[5] T. He, W. Xiao, T. Lin, *et al.*, "Hover: Versatile neural whole-body controller for humanoid robots," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9989–9996, IEEE, 2025.

[6] M. J. Kim, K. Pertsch, Karamcheti, *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[7] D. Qu, H. Song, Chen, *et al.*, "Spatialvla: Exploring spatial representations for visual-language-action model," *arXiv preprint arXiv:2501.15830*, 2025.

[8] J. Bjorck, F. Castañeda, Cherniadev, *et al.*, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[9] R. Zhang, M. Dong, Zhang, *et al.*, "Mole-vla: Dynamic layer-skipping vision language action model via mixture-of-layers for efficient robot manipulation," *arXiv preprint arXiv:2503.20384*, 2025.

[10] M. Ji, X. Peng, Liu, *et al.*, "Exbody2: Advanced expressive humanoid whole-body control," *arXiv preprint arXiv:2412.13196*, 2024.

[11] Z. Wu, J. Lu, Chen, *et al.*, "Stride: Automating reward design, deep reinforcement learning training and feedback optimization in humanoid robotics locomotion," *arXiv preprint arXiv:2502.04692*, 2025.

[12] R. Sapkota, Y. Cao, Roumeliotis, *et al.*, "Vision-language-action (vla) models: Concepts, progress, applications and challenges," *arXiv preprint arXiv:2505.04769*, 2025.

[13] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

[14] P. Intelligence, K. Black, Brown, *et al.*, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025.

[15] K. Black, N. Brown, Driess, *et al.*, "pi_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[16] M. J. Kim, C. Finn, and P. Liang, "Fine-tuning vision-language-action models: Optimizing speed and success," *arXiv preprint arXiv:2502.19645*, 2025.

[17] J. Zhang, Y. Guo, Hu, *et al.*, "Up-vla: A unified understanding and prediction model for embodied agent," *arXiv preprint arXiv:2501.18867*, 2025.

[18] H. Zhen, X. Qiu, Chen, *et al.*, "3d-vla: A 3d vision-language-action generative world model," *arXiv preprint arXiv:2403.09631*, 2024.

[19] T. Jiang, T. Yuan, Liu, *et al.*, "Galaxea open-world dataset and g0 dual-system vla model," *arXiv preprint arXiv:2509.00576*, 2025.

[20] H. Xue, X. Huang, Niu, *et al.*, "Leverb: Humanoid whole-body control with latent vision-language instruction," *arXiv preprint arXiv:2506.13751*, 2025.

[21] W. Song, J. Chen, Li, *et al.*, "Rationalvla: A rational vision-language-action model with dual system," *arXiv preprint arXiv:2506.10826*, 2025.

[22] C. Cui, P. Ding, Song, *et al.*, "Openhelix: A short survey, empirical analysis, and open-source dual-system vla model for robotic manipulation," *arXiv preprint arXiv:2505.03912*, 2025.

[23] Y. Ze, Z. Chen, Araújo, *et al.*, "Twist: Teleoperated whole-body imitation system," *arXiv preprint arXiv:2505.02833*, 2025.

[24] T. He, J. Gao, Xiao, *et al.*, "Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills," *arXiv preprint arXiv:2502.01143*, 2025.

[25] H. Deng, Z. Wu, Liu, *et al.*, "A survey on reinforcement learning of vision-language-action models for robotic manipulation," *Authorea Preprints*, 2025.

[26] G. Team, T. Mesnard, Hardin, *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.

[27] M. Mittal, P. Roth, Tigue, *et al.*, "Isaac Lab - A GPU-Accelerated Simulation Framework for Multi-Modal Robot Learning," *arXiv preprint arXiv:2511.04831*, 2025.

[28] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on November 20,2026 at 04:37:04 UTC from IEEE Xplore. Restrictions apply.