PiVLA: Vision-Language-Action Based on Parallel Intelligence

Fei Lin, Tengchao Zhang, Jun Huang, Qinghua Ni, Jingwei Ge, and Yonglin Tian

***Abstract*—The vision-language-action (VLA) paradigm is gradually becoming the core path of embodied intelligence. However, its training and validation, which rely on simulation environments, face serious sim2real challenges, such as navigation deviations in drones caused by wind speed differences between simulation and real-world environments. Existing iterative methods based on digital twins can alleviate the problem of virtual-real alignment to some extent. However, their high dependence on twin consistency limits their adaptability and scalability in complex environments. To break through this bottleneck, the PiVLA framework is proposed in this letter to reconstruct the VLA paradigm with parallel intelligence. Furthermore, we introduce the parallel deep foundation model (PDFM) and, based on it, propose model parallel control (MPC) and the parallel interaction protocol (PIP), establishing a unified interaction mechanism for disembodied agents and embodied agents. This provides a scalable and robust solution for complex tasks involving embodied intelligence.**

I. Introduction

With the continuous evolution of large language models (LLMs) and multimodal large language models (MLLMs), vision-language-action (VLA) has gradually become the core topic of embodied intelligence research. The goal of VLA is to closely align visual perception with language reasoning and further transform them into executable action strategies, thereby forming an end-to-end "perception-reasoning-execution" closed loop. Constrained by cost and scalability, a large number of studies tend to obtain training data and complete inference verification in simulation environments [1]. However, solely relying on simulation brings significant sim2real challenges. First, the visual modality has differences with the real world in rendering quality and semantic distribution. Second, complex factors such as dynamics, energy consumption, and latency in the real physical world are complex to be accurately replicated on the simulation side, leading to a significant decline in transfer performance and robustness [2].

To alleviate this problem, researchers propose the "real-sim-real" iterative paradigm, achieving closed-loop optimization through back-and-forth correction between reality and simulation [3]. For example, works such as RialTo [4], ManiSkill3 [5], and DexSim2Real^2 [6] utilize digital twins (DTs) to establish one-to-one mappings between simulation and reality, improving the virtual-real alignment to a certain extent. However, such methods usually require the simulation environment to almost strictly replicate the real system, which not only has high modeling costs, but also is challenging to cover all the uncertainties in complex environments, limiting its adaptability and scalability in VLA scenarios.

Parallel intelligence (PI) and artificial systems, computational experiments, and parallel execution (ACP) method, which were proposed more than twenty years ago, provided a natural theoretical fit to address this problem [7]. The core idea lies in: not pursuing strict copying of the physical world, but flexibly constructing artificial systems according to the needs of the actual embodied systems; then systematically exploring the potential policy space through computational experiments; and finally achieving dynamic interaction and evolutionary iteration between virtual and real systems relying on parallel execution. In contrast, DT approaches emphasize high-fidelity replication and strict twin consistency, whereas PI focuses on key mapping and dynamic optimization, thereby demonstrating stronger robustness and generalization in complex and uncertain environments. Based on this idea, we propose PiVLA, aiming to reconstruct the VLA paradigm with parallel intelligence and establish a sustainable, co-evolving, and embodied intelligence framework across virtual and real domains.

The main contributions of this letter are as follows:

(1) We construct the PiVLA framework based on parallel intelligence, redefining the real-sim-real research paradigm of vision-language-action.

(2) We propose the interaction paradigm for disembodied agents and embodied agents for the first time, building a unified bridge between agentic systems in cyberspace and embodied agent systems in the physical world.

(3) With the support of the parallel deep foundation model (PDFM), we introduce model parallel control (MPC) and parallel interaction protocol (PIP), providing scalable mechanisms and theoretical support for control and interaction between virtual and real agents.

Manuscript received: 23 July 2025; revised: 2 August 2025; accepted: 4 August 2025. *(Corresponding author: Yonglin Tian.)*

Citation: F. Lin, T. Zhang, J. Huang, Q. Ni, J. Ge, and Y. Tian, PiVLA: Vision-language-action based on parallel intelligence, *Int. J. Intell. Control Syst.*, 2025, 30(3), 253–259.

Fei Lin, Tengchao Zhang, Jun Huang, and Qinghua Ni are with Faculty of Innovation Engineering, Macau University of Science and Technology, Macao 999078, China (e-mail: feilin@ieee.org; zhangtengchao@ieee.org; junhuang@ieee.org; qinghua.ni@ieee.org).

Jingwei Ge is with University Research and Innovation Center, Óbuda University, Budapest 1034, Hungary (e-mail: jingwei.ge@uniobuda.hu).

Yonglin Tian is with State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yonglin.tian@ia.ac.cn).

Digital Object Identifier 10.62678/IJICS202509.10242

II. PRELIMINARY

A. PI and the ACP Method

PI has emerged as a paradigm for addressing the limitations of traditional models and heuristic rules in managing the uncertainty and complexity of modern socio-technical systems [7]. Built upon the ACP framework, PI integrates virtual and real systems into a dynamic feedback loop: artificial systems serve as computational surrogates of reality, computational experiments enable systematic exploration and optimization of strategies, and parallel execution closes the loop through bidirectional interaction between virtual models and real-world operations. Unlike digital twins that primarily focus on high-fidelity replication, PI incorporates descriptive, predictive, and prescriptive intelligence into a unified mechanism, encapsulated by the principle of "learning, simulating, and optimizing virtually before applying to reality". Over the past two decades, PI has been extended to diverse domains, such as computer vision [8] and drug discovery [9], demonstrating its value as a reusable methodological framework for adaptive governance and sustained system evolution through iterative virtual-real system interaction.

B. Advance in Vision-Language-Action

VLA establishes a unified decision-making framework across vision, language, and action, enabling agents to comprehend multimodal inputs and accomplish complex tasks. Compared with the traditional perception-planning-control pipeline, VLA emphasizes end-to-end fusion and task-driven execution, facilitating the transition of robots from structured scenarios to general environments [10]. As a pivotal direction in embodied intelligence research, VLA has evolved from modular to end-to-end paradigms, with representative works such as RT-2 [11] and OpenVLA [12] advancing multimodal integration. Current researches focus on improving generalization in complex interactions. Models such as LoHoVLA [13] and TraceVLA [14] optimize performance through task decomposition, future prediction, and trajectory modeling. In contrast, SwitchVLA [15] and OTTER [16] enhance task switching and visual feature selection capabilities. Nevertheless, challenges remain in ensuring dynamic consistency and long-term reasoning. The integration of world models (WMs) provides predictive support for physical consistency, thereby reinforcing the interaction capabilities of VLA systems in real-world environments [17].

C. PDFM and World Model

PDFM [18] proposes a parallel intelligence framework that integrates analogical imagination and embodied cognition, establishing a co-evolution mechanism between the data world and the real world to enable adaptive deployment and continual optimization of foundation models. The architecture comprises a front-end execution model and a shadow model, encompassing the unified stages of pre-training, post-training, and deployment-stage distillation. It supports co-evolution through both training-time interaction and inference-time interaction mechanisms.

WMs simulate future states and behavioral outcomes to provide predictive support for path planning, policy rehearsal, and long-horizon task modeling. They are widely applied in navigation, interaction, and embodied intelligence [1, 19]. For instance, NWM [20] enhances path prediction generalization through conditional video generation. TWISTER [21] leverages Transformers and contrastive prediction to improve long-term modeling. EVA [22] introduces the reflection of generation (RoG) strategy to enhance robustness in video prediction. WMs provide the predictive foundation for PDFM by modeling environmental dynamics and supporting the handling of physical world constraints. When combined with PDFM's analogical reasoning and multimodal alignment mechanisms, these modeling capabilities significantly enhance the cognitive and adaptive performance of models in dynamic environments.

D. Agent Protocol

In recent years, several protocols have been introduced to support interoperability between intelligent agents and large models, marking a shift from general communication mechanisms to domain-specific safeguards [23]. The model context protocol (MCP), proposed by Anthropic, provides a JSON-RPC interface linking models with tools, databases, and application programming interfaces (APIs), enabling a context-execution-feedback loop similar to the language server protocol (LSP). MCP has been applied in retrieval-augmented generation (RAG), code execution, and database querying [24, 25]. Google's agent-to-agent (A2A) protocol [26] enables heterogeneous agents to discover capabilities and coordinate tasks through standardized "Agentcard", supporting applications such as multi-robot coordination and cross-platform assistants. The agent communication protocol (ACP) [27] focuses on asynchronous and multimodal messaging for real-time dialogue and collaboration.

III. PIVLA FRAMEWORK

The overall theoretical framework of PiVLA is shown in Fig. 1. For general or open-domain problems (large problems), it is necessary to rely on large models to handle the high degree of uncertainty. However, these large models are not optimized for specific VLA scenarios and often lack domain depth and adaptability. Therefore, under the support of parallel intelligence, it is necessary to transform external large models into PDFM that can support the cyclic interaction and continuous learning between virtual and real systems. In the context of VLA, this model framework is concretely embodied as PiVLA, which is used to address domain-specific problems and thus achieve effective adaptation to VLA tasks. Furthermore, at the task level, task-specific problems (small problems) often need to be grounded in concrete application scenarios. PiVLA can be continuously optimized and iterated, gradually evolving into edge/end models that meet the requirements of VLA tasks, thereby supporting the real-world deployment and application of deeply embodied intelligent agents.

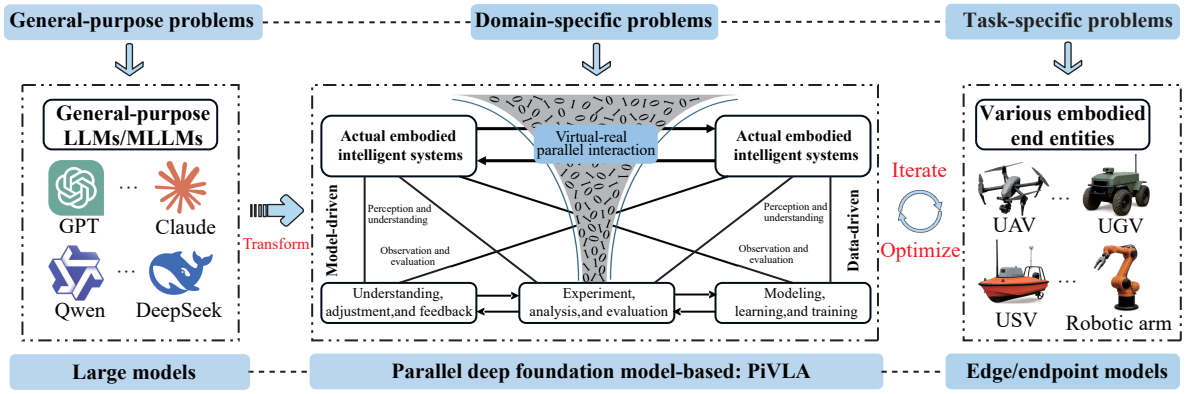

Fig. 1 Theoretical framework of PiVLA. UAV stands for unmanned aerial vehicle, UGV represents unmanned ground vehicle, and USV is unmanned surface vehicle.

Based on this, PiVLA takes parallel intelligence as its theoretical core and constructs corresponding artificial and embodied intelligent systems through the modeling and abstraction of the requirements of actual and embodied intelligent systems. With the mechanism of "human-in-the-loop" understanding-adjustment-feedback and modeling-learning-training, PiVLA organically integrates extensive computational experiments, analysis, and evaluation processes, and explores the potential solution space of complex problems through virtual-real parallel execution. Thus, this framework not only supports reliable management and control of complex systems, but also enables the continuous evolution of embodied agents in the real world during long-term operation.

The specific implementation framework of PiVLA is shown in Fig. 2, which focuses on the training and deployment process of VLA and can essentially be understood as the engineering expansion of the ACP method in the field of embodied intelligence. Specifically, modeling and representation (A) are first completed in artificial scenarios, followed by computational experiments generating and validating executable strategies (C). Finally, parallel execution is carried out in both the artificial and real scenarios, with feedback flowing back to the artificial scenarios (P), thereby forming a continuously iterative closed loop. Based on this cyclic mechanism, PiVLA can continuously optimize the performance and adaptability of VLA systems in virtual-real combined environments. The following sections further introduce its key components, interaction modes, and implementation processes.

A. Artificial System in PiVLA

The iterative mode of "real-simulation-real" adopted in VLA is essentially a concrete manifestation of virtual-real parallel interaction. In the PiVLA framework, artificial scenarios together with disembodied agents constitute artificial systems, while the corresponding real scenarios and embodied agents form actual systems. The construction of artificial scenarios needs to be based on the requirements of actual systems. Their forms can include DT, virtual simulation environments, and software-defined environments. In the context of VLA, artificial scenarios usually cover three aspects: (1) the three-dimensional representation of embodied entities and their action spaces, (2) the physical properties involved in interactions, and (3) the constraints of the relevant environments and behaviors. It should be particularly emphasized that artificial scenarios are not equivalent to DTs.

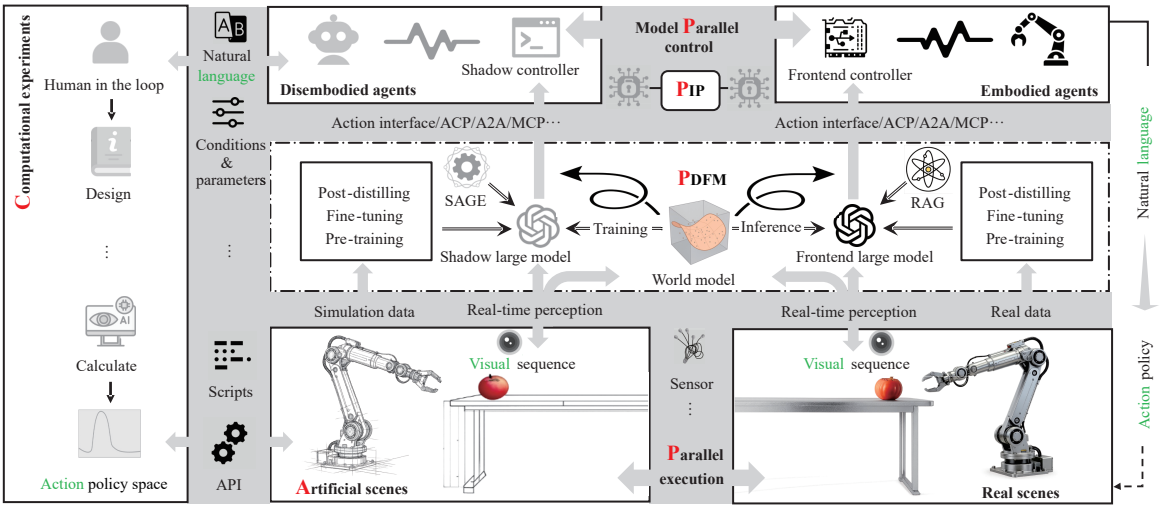

Fig. 2 Implementation framework of PiVLA.

Their goal is not to strictly replicate reality, but to ensure that the key representations within them have corresponding mappings in real systems, thereby serving as the fundamental condition for parallel execution and virtual-real interaction.

B. PDFM

PiVLA relies on PDFM as the "embodied brain" of the system, jointly supporting the operation of artificial systems and actual systems, and enabling parallel transmission and processing from scene-level visual perception to agent control. Specifically, PDFM consists of a frontend large model, a shadow large model, and a WM. The frontend large model is deployed at the edge or terminal, directly acting on the physical world to ensure the real-time and interactive nature of task execution. The shadow large model resides in the cloud, responsible for exploration, generalization, and inference in virtual space. During training, the frontend and shadow large models achieve capability enhancement and continuous optimization through multi-level interactions at the data layer, feature layer, and gradient layer. Simulation data and AI-generated content (AIGC) data generated in artificial scenarios are mainly used for the pre-training, fine-tuning, and distillation of the shadow large model. Environmental data and human-machine interaction data collected by sensors in real scenarios drive the updates of the frontend large model. During inference, visual sequence information from scenarios is simultaneously input into both the frontend and shadow models to form real-time perception and interaction. Among these, the shadow model further enhances the input quality and reasoning effectiveness of the frontend model through information augmentation and knowledge augmentation.

In this process, a dual-mode mechanism of retrieval-augmented generation [28] and search-augmented generation and extension (SAGE) [29] is introduced to enhance the capability of PDFM. The frontend large model integrates RAG, enhancing understanding and response to real-world scenarios through external knowledge retrieval and context augmentation, thereby ensuring the real-time and reliable nature of inference results. The shadow large model adopts SAGE, whose core mechanism goes beyond one-time retrieval, constructing a cyclic process of "search-generation-extension-feedback", supplemented by caching and multi-level knowledge management to support cross-task consistency and long-term optimization. In this way, SAGE enables deeper inference and information completion in artificial scenarios, effectively compensating for the shortcomings of PDFM in factual consistency and complex constraint modeling.

However, PDFM constructed based on LLM/MLLM is difficult to directly handle dynamic constraints in the physical world, such as dynamics, friction, energy consumption, latency, and uncertainty. To address this, PiVLA introduces the WM as a supplement. WM models environmental dynamics and predicts future states to provide dynamic consistency support for interactions with the physical world. Relying on real-time collected visual sequences, the system, supported by the video generation diffusion models, can predict the evolutionary trend of the environment at future moments, thereby correcting the planning and decision-making of LLM/MLLM. At the same time, diffusion models can implicitly embed physical laws during the generation process, thus maintaining dynamical consistency in the explicit prediction results.

C. Agentic System, MPC, and PIP

Since PDFM itself still mainly functions as a passive model, typically relying on "human-in-the-loop" prompts and supervision, VLA, in specific scenarios, requires agents to possess active interaction and exploration capabilities. To this end, PiVLA introduces agentic systems to support more autonomous decision-making and collaboration. To connect PDFM with both disembodied agents and embodied agents, it is necessary to establish standardized interaction and collaboration mechanisms at the upper level. First, action interfaces are needed to transform high-level semantic intentions into executable plans or instructions, which can be realized through tool/function calls. Second, transmission protocols are required to ensure reliable interaction of actions and feedback, such as model context protocol for tool invocation, agent-to-agent for multi-agent collaboration, and ACP (referring to the agent communication protocol) for connecting agents, applications, and human users. However, these protocols alone are insufficient to support parallel operation across both virtual and real domains, and a control mechanism is required to ensure dynamic consistency and coordination between the two.

Based on this, we propose the MPC as a novel control mechanism. Unlike traditional model predictive control, which performs finite-horizon prediction and optimization only in real systems, MPC emphasizes bidirectional parallel regulation between virtual artificial systems and real physical systems. The frontend model undertakes real-time control under low-latency and safety constraints in real-world environments. In contrast, the shadow model carries out counterfactual reasoning and multi-branch policy evaluation in artificial systems. The two remain coordinated through consistency constraints and risk gating. In this way, the controller can continuously optimize through computational experiments in the virtual domain while ensuring stable execution and constraint satisfaction in the real domain, thereby achieving parallel operation of disembodied agents and embodied agents.

Within the PiVLA framework, we propose the PIP to regulate the collaboration between disembodied agents and embodied agents across virtual and real domains. The core of PIP lies in establishing standardized channels between the frontend controller and the shadow controller, unifying data flow (observations and uncertainties), knowledge flow (augmentation and inference results), and control flow (parameter scheduling and risk gating). The frontend controller is deployed in real-world scenarios, relying on the frontend large model to achieve real-time execution under low latency and safety constraints. The shadow controller operates in artificial scenarios, driven by shadow large models to perform counterfactual reasoning, multi-branch evaluation, and long-term policy learning. Through PIP, the frontend

controller can transmit state and uncertainty information to the shadow controller during inference and receive suggestions for knowledge augmentation and parameter scheduling in return, thus enabling dynamic correction. Meanwhile, trajectories and failure cases after execution are fed back to the shadow controller, supporting relearning and distillation. In this way, PIP not only achieves real-time linkage and dynamic consistency constraints between the frontend and shadow controllers, but also provides a general communication and interaction mechanism for the parallel operation of disembodied agents and embodied agents, thereby promoting long-term optimization and cross-task generalization of complex embodied tasks while ensuring low latency and safety.

D. Computational Experiment and Parallel Execution

The stage of computational experiments is mainly oriented toward disembodied agents. At this stage, researchers design experimental scenarios according to the requirements of actual systems, and through human-in-the-loop interactions, specify initial conditions such as the number of agents, resource distribution, and network topology using natural language. Subsequently, the experiments are executed in artificial scenarios via APIs or scripts, and their results not only provide the basis for strategy generation and evaluation but also guide new experimental designs in turn. In this way, dynamic simulation, closed-loop feedback, and continuous evolution are achieved, gradually approaching the action policy space of agents. For example, in an embodied grasping VLA task, disembodied agents can perform large-scale simulations of grasping actions in artificial scenarios, exploring optimal grasping trajectories and action sequences through continuous trial-and-error and policy updates. These candidate strategies obtained on the virtual side are then filtered through consistency constraints and mapped to the operations of real robotic arms, thereby effectively shortening the real training cycle and improving the safety and robustness of policies.

In the PiVLA framework, parallel execution is manifested in multiple ways. The most essential form is that the computational experiment results of disembodied agents in artificial scenarios are synchronously mapped to the real systems through MPC. Meanwhile, artificial and real scenarios achieve real-time perception and data exchange through sensors and the Internet of Things (IoT) technologies, ensuring alignment and updating of states across the virtual and real domains. Furthermore, the frontend large model and the shadow large model that support the agents also maintain continuous co-evolution during parallel execution: the frontend model ensures low-latency execution and satisfaction of physical constraints, while the shadow model undertakes counterfactual inference and long-term optimization. The two iteratively update within the virtual-real closed loop, thereby achieving cross-scenario adaptability and lifelong learning.

IV. Application

This section presents the potential applications of PiVLA in typical scenarios of embodied intelligence, including autonomous UAV navigation, robotic manipulation in autonomous self-driving laboratories (SDLs), and intervention tasks in surgical robots. These scenarios correspond to the construction and execution of navigation, grasping, and intervention policy spaces, respectively. It should be noted that these cases only serve as illustrative applications of the framework, intended to demonstrate the adaptability and scalability of PiVLA, rather than experimental evaluations on specific systems.

A. Vision-Language-Navigation (VLN) in Autonomous UAV System

In autonomous UAV systems, VLN is one of the most representative tasks of embodied intelligence. PiVLA constructs the navigation policy space through computational experiments to explore path planning and instruction parsing strategies in complex environments. As shown in Fig. 3(a), disembodied UAV agents conduct large-scale policy simulation and robustness evaluation in artificial scenarios. Subsequently, through the parallel control mechanisms of MPC and PIP, the selected trajectory templates, sub-goals, and constraint parameters are transmitted to the embodied UAV agents, forming a virtual-real closed loop together with sensor feedback. Throughout this process, PDFM provides consistency support and online correction for perception, reasoning, and world modeling, and is expected to enhance the reliability of low-altitude navigation.

B. Autonomous Robotic Manipulation for SDL

In self-driving laboratory scenarios, robotic manipulation tasks impose extremely high requirements on precision, repeatability, and dynamic adaptability. PiVLA generates the grasping policy space during computational experiments, encompassing grasping and manipulation schemes for multiple objects, poses, and constraints. As shown in Fig. 3(b), disembodied robotic arm agents complete large-scale action inference and policy optimization in artificial scenarios. Subsequently, grasping poses, trajectory/velocity profiles, and force-control parameters are delivered to embodied robotic arms through MPC and PIP. At the same time, real execution data and failure cases are fed back to support relearning and policy refinement.

C. Vision-Language-Guided (VLG) Intervention in Robotic Surgery

In surgical robot applications, the VLG intervention tasks impose stringent requirements on safety, dynamic responsiveness, and real-time decision-making. PiVLA constructs the intervention policy space through computational experiments, covering key aspects such as intervention path planning, instrument coordination, and safety boundary settings. As shown in Fig. 3(c), disembodied surgical robot agents perform counterfactual inference and multi-path risk evaluation in artificial scenarios. After undergoing risk gating and consistency constraints in the PIP channel via MPC, the selected intervention strategies and control parameters are delivered in parallel to embodied surgical robots, with intraoperative feedback flowing back to support continuous correction.

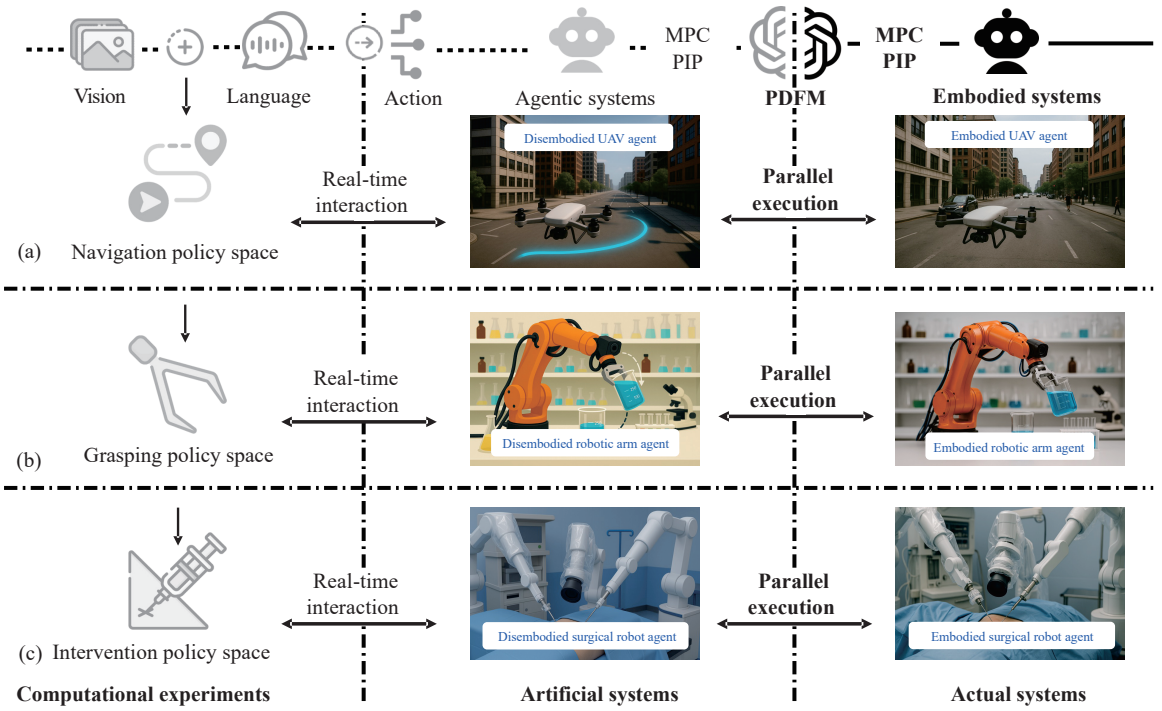

Fig. 3 PiVLA application framework for representative embodied tasks.

V. Conclusion

This letter proposes the PiVLA framework based on parallel intelligence to address the core challenges of VLA in sim2real transfer. By introducing PDFM and constructing MPC and PIP on the basis it, we establish a unified interaction mechanism for non-embodied agents and embodied agents, providing a robust and scalable solution for complex embodied tasks. Future research will progressively enable the practical application analysis and experimental evaluation of PiVLA in the aforementioned illustrative scenarios, and further extend its potential to cross-scenario, multi-task, and collective intelligence settings, driving the continuous evolution of embodied intelligence.

Acknowledgment

This work was supported by the Science and Technology Development Fund, Macao Special Administrative Region (Nos. 0157/2024/RIA2, 0145/2023/RIA3, and 0093/2023/RIA2). The authors gratefully acknowledge Prof. Fei-Yue Wang, the principal investigator of these projects, for initiating and supporting the research program under which part of this work was carried out.

References

[1] X. Long, Q. Zhao, K. Zhang, Z. Zhang, D. Wang, Y. Liu, Z. Shu, Y. Lu, S. Wang, X. Wei et al., A survey: Learning embodied intelligence from physical simulators and world models, arXiv preprint arXiv: 2507.00917, 2025.

[2] J. Duan, S. Yu, H. Tan, H. Zhu, and C. Tan, A survey of embodied AI: From simulators to research tasks, *IEEE Trans. Emerg. Top. Comput. Intell.*, 2022, 6(2), 230–244.

[3] Y. Liu, W. Chen, Y. Bai, X. Liang, G. Li, W. Gao, and L. Lin, Aligning cyber space with physical world: A comprehensive survey on embodied AI, *IEEE/ASME Trans. Mechatron.*, to be published.

[4] M. Torne, A. Simeonov, Z. Li, A. Chan, T. Chen, A. Gupta, and P. Agrawal, Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation, arXiv preprint arXiv: 2403.03949, 2024.

[5] S. Tao, F. Xiang, A. Shukla, Y. Qin, X. Hinrichsen, X. Yuan, C. Bao, X. Lin, Y. Liu, T.-K. Chan et al., ManiSkill3: GPU parallelized robotics simulation and rendering for generalizable embodied AI, arXiv preprint arXiv: 2410.00425, 2024.

[6] T. Jiang, Y. Guan, L. Ma, J. Xu, J. Meng, W. Chen, Z. Zeng, L. Li, D. Wu, and R. Chen, DexSim2Real^2: Building explicit world model for precise articulated object dexterous manipulation, arXiv preprint arXiv: 2409.08750, 2024.

[7] F.-Y. Wang, Parallel system methods for management and control of complex systems, *Control Decis.*, 2004, 19(5), 485–489, (in Chinese).

[8] K. Wang, C. Gou, N. Zheng, J. M. Rehg, and F.-Y. Wang, Parallel vision for perception and understanding of complex scenes: Methods, framework, and perspectives, *Artif. Intell. Rev.*, 2017, 48(3), 299–329.

[9] F. Lin, J. Yang, D. Sun, L. Kovács, and F.-Y. Wang, Autonomous drug discovery with parallel intelligence, *IEEE/CAA J. Autom. Sin.*, 2025, 12(8), 1742–1744.

[10] H. Li, Y. Chen, W. Cui, W. Liu, K. Liu, M. Zhou, Z. Zhang, and D. Zhao, Survey of vision-language-action models for embodied manipulation, arXiv preprint arXiv: 2508.15201, 2025.

[11] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid et al., RT-2: Vision-language-action models transfer web knowledge to robotic control, in *Proc. 7th Conference on Robot Learning*, Atlanta, GA, USA, 2023, 2165–2183.

[12] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong et al., OpenVLA: An open-source vision-language-action model, in *Proc. 8th Conference on Robot Learning*, Munich, Germany, 2024.

[13] Y. Yang, J. Sun, S. Kou, Y. Wang, and Z. Deng, LoHoVLA: A unified vision-language-action model for long-horizon embodied tasks, arXiv preprint arXiv: 2506.00411, 2025.

[14] R. Zheng, Y. Liang, S. Huang, J. Gao, H. Daumé III, A. Kolobov, F. Huang, and J. Yang, TraceVLA: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies, in *Proc. 13th International Conference on Learning Representations*, Singapore, Singapore, 2025.

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

[16] H. Huang, F. Liu, L. Fu, T. Wu, M. Mukadam, J. Malik, K. Goldberg,

and P. Abbeel, OTTER: A vision-language-action model with text-aware visual feature extraction, arXiv preprint arXiv: 2503.03734, 2025.

[17] D. Ha and J. Schmidhuber, World models, arXiv preprint arXiv: 1803.10122, 2018.

[18] Y. Tian, F. Lin, C. Wang, J. Ge, Z. Luo, and F.-Y. Wang, Parallel deep foundation model: A co-evolution framework for analogical imagination and embodied cognition of parallel intelligence, 2025, 30, 83–90.

[19] P. Wu, A. Escontrela, D. Hafner, P. Abbeel, and K. Goldberg, DayDreamer: World models for physical robot learning, in *Proc. 6th Conference on Robot Learning*, Auckland, New Zealand, 2023, 2226–2240.

[20] A. Bar, G. Zhou, D. Tran, T. Darrell, and Y. LeCun, Navigation world models, in *Proc. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2025, 15791–15801.

[21] M. Burchi and R. Timofte, Learning transformer-based world models with contrastive predictive coding, in *Proc. 13th International Conference on Learning Representations*, Singapore, Singapore, 2025.

[22] X. Chi, C.-K. Fan, H. Zhang, X. Qi, R. Zhang, A. Chen, C.-M. Chan, W. Xue, Q. Liu, S. Zhang et al., Empowering world models with reflection for embodied video prediction, in *Proc. 42nd International Conference on Machine Learning*, Vienna, Austria, 2025.

[23] A. Ehtesham, A. Singh, G. K. Gupta, and S. Kumar, A survey of agent interoperability protocols: Model context protocol (MCP), agent communication protocol (ACP), agent-to-agent protocol (A2A), and agent network protocol (ANP), arXiv preprint arXiv: 2505.02279, 2025.

[24] Anthropic, What is the model context protocol (MCP)? [Online], https://modelcontextprotocol.io/docs/getting-started/intro, 17 July 2025.

[25] X. Hou, Y. Zhao, S. Wang, and H. Wang, Model context protocol (MCP): Landscape, security threats, and future research directions, arXiv preprint arXiv: 2503.23278, 2025.

[26] Google Cloud, Announcing the Agent2Agent protocol (A2A) [Online], https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/, 17 July 2025.

[27] IBM Research, Agent communication protocol (ACP) [Online], https://research.ibm.com/projects/agent-communication-protocol, 17 July 2025.

[28] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel et al., Retrieval-augmented generation for knowledge-intensive NLP tasks, in *Proc. 34th International Conference on Neural Information Processing Systems*, Vancouver, Canada, 2020, 793.

[29] J. Zhang, G. Li, and J. Su, SAGE: A framework of precise retrieval for RAG, in *Proc. 2025 IEEE 41st International Conference on Data Engineering*, Hong Kong, China, 2025.

Fei Lin received the MS degree from Macau University of Science and Technology, China, in 2023. He is currently pursuing the PhD degree at Department of Engineering Science, Faculty of Innovation Engineering, Macau University of Science and Technology, China. His research interests include parallel intelligence, large language models, and embodied agents.

Tengchao Zhang received the MS degree from Macau University of Science and Technology, China, in 2024. He is currently pursuing the PhD degree at Department of Engineering Science, Faculty of Innovation Engineering, Macau University of Science and Technology, China. His research interests include parallel intelligence, large language models, and parallel tourism.

Jun Huang received the MS degree from Faculty of Science and Engineering, University of Manchester, UK, in 2021. He is currently pursuing the PhD degree at Department of Engineering Science, Faculty of Innovation Engineering, Macau University of Science and Technology, China. His research interests include parallel intelligence, multi modal large language models, and intelligent vehicles.

Qinghua Ni received the MS degree from King's College London, UK, in 2022. She is currently pursuing the PhD degree in intelligent science and systems study at Faculty of Innovation Engineering, Macau University of Sciences and Technology, China. Her research interests include parallel intelligence, decision intelligence, blockchain and DAO, and parallel theater.

Jingwei Ge received the PhD degree from Department of Automation, Tsinghua University, China, in 2024. He is currently a researcher at University Research and Innovation Center, Óbuda University, Hungary. His research interests include intelligence testing, scenario generation, intelligent vehicles, and digital twins.

Yonglin Tian received the PhD degree in control science and engineering from University of Science and Technology of China, China, in 2022. He is currently an assistant researcher at State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, China. His research interests include parallel intelligence, autonomous driving, and intelligent transportation systems.