Parallel Deep Foundation Model: A Co-Evolution Framework for Analogical Imagination and Embodied Cognition of Parallel Intelligence

Yonglin Tian, Fei Lin, Cong Wang, Jingwei Ge, Zhiyao Luo, and Fei-Yue Wang

Abstract—**The rise of foundation models has brought significant advances to artificial intelligence, especially in reasoning, commonsense understanding, and tool use. These capabilities, when integrated into agent systems, hold great promise for real-world applications such as vision-language navigation (VLN) and vision-language action (VLA). However, deploying such models in practice presents ongoing challenges, particularly in adapting and optimizing them across diverse and changing environments. This letter proposes a parallel deep foundation model (PDFM) framework to support continuous model evolution in cloud-edge-device systems. The framework establishes a co-evolution process between two complementary capabilities: embodied cognition, which reflects the model's grounded understanding and task adaptation in physical systems, and analogical imagination, which enables creative exploration and capacity expansion in virtual environments. Through three core processes, learning and training, experiment and evaluation, and management and control, the system supports iterative refinement and dynamic interaction between virtual and real spaces. This enables general-purpose models to gradually converge toward domain-specific intelligence, supporting long-term, adaptive deployment.**

I. Introduction

The rapid progress of large foundation models has provided new momentum for the development of general artificial intelligence (AI). By leveraging advanced capabilities in commonsense understanding and reasoning, these models enable AI systems to perform memory-based inference, make context-aware decisions, and interact with external tools. Such abilities mark a significant step toward mimicking key aspects of human cognition and behavior, and lay a solid groundwork for advancing AI into more complex, embodied, and interactive environments.

Despite their impressive potential, large foundation models still face several critical challenges in real-world applications. First, there remains a significant gap between general-purpose capabilities and the specialized needs of specific domains; models often struggle to transfer effectively without costly fine-tuning or adaptation. Second, ensuring continuous optimization throughout the model's lifecycle is difficult, especially in changing environments where data distributions, task requirements, or system constraints may shift over time. Third, the practical deployment of large models remains a major hurdle, as their massive parameter sizes and computational demands pose serious limitations for integration into edge devices or resource-constrained systems.

To address the challenges of deploying foundation models in real-world applications and to enable their long-term optimization, this letter proposes the parallel deep foundation model (PDFM), a framework that integrates cloud-edge-device systems with the principles of parallel intelligence [1]. The architecture centers on two key components: a front-end model for real-time applications and a shadow model for continuous improvement. It incorporates a unified process of pretraining, post-training, and deployment-stage distillation. Through training-time interaction (at the levels of data, features, and gradients) and inference-time interaction (enhanced by information and knowledge fusion), the two models form a virtual-physical collaborative learning mechanism. This mechanism supports the adaptive deployment and lifelong optimization of foundation models in complex and evolving environments, combining embodied cognition, real-time learning through physical interaction, with analogical imagination, the ability to reason about novel tasks by drawing on past experience and virtual simulation.

The main contributions of this letter are as follows:

(1) We propose the architecture of cloud-edge-end intelligence, abbreviated as Cloudedgend Intelligence that integrates cloud-edge-device systems, organizing models around three key roles: computation, coordination, and

Manuscript received: 10 December 2024; revised: 25 January 2025; accepted: 30 January 2025. *(Corresponding author: Fei-Yue Wang.)*

Citation: 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, *Int. J. Intell. Control Syst.*, 2025, 30(1), 83–90.

Yonglin Tian, Cong Wang, and Zhiyao Luo are with State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yonglin.tian@ia.ac.cn; wangcong2024@ia.ac.cn; luozhiyao2024@ia.ac.cn).

Fei Lin is with Faculty of Innovation Engineering, Macau University of Science and Technology, Macao 999078, China (e-mail: feilin@ieee.org).

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

Fei-Yue Wang is with John von Neumann Faculty of Informatics, Obuda University, Budapest H-1034, Hungary, with State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with Faculty of Innovation Engineering, Macau University of Science and Technology, Macao 999078, China (e-mail: feiyue.wang@ia.ac.cn).

Digital Object Identifier 10.62678/IJICS202503.10184

execution. This enables adaptive model routing and flexible task allocation based on scenario requirements and task complexity.

(2) We introduce a dual-model design that decouples real-time application and long-term optimization by separating the front-end model and the shadow model. This parallel structure supports the full lifecycle of model training, adaptation, and deployment.

(3) We develop a dual-channel interaction mechanism for parallel learning, involving inference-time interaction (at both the information and knowledge levels) and training-time interaction (across data, features, and gradient layers), enabling continuous and efficient lifelong model optimization.

This framework provides a unified solution that aligns model architecture, continuous optimization, and deployment. By decoupling real-time execution and long-term learning, and enabling interaction across training and inference stages, it offers a scalable approach to building foundation models that are not only adaptable and efficient, but also deployable across heterogeneous environments.

II. Related Work

A. Foundation Model in Cloud-Edge System

Cloud-edge systems are a computing architecture aimed at real-time deployment response requirements through deployment at the network edge. Nowadays, large language models (LLMs) have been widely adopted in cloud-edge systems in various automated scenarios such as industrial Internet of Things (IoT), intelligent transportation, and telemedicine [2–9]. Transmitting all raw data back to the cloud for processing would overwhelm network bandwidth, while delays in critical commands could lead to safety incidents. To address this challenge, the cloud-edge computing paradigm has been proposed. Edge nodes typically execute high-frequency, low-latency core tasks and equipment status monitoring and transmit refined, high-value information or model update requests to the cloud. With enormous computing power, the cloud typically handles global optimization scheduling, long-term trend forecasting, and the training of advanced AI models, and then deploys optimized, lightweight versions of these models to the edge.

Collaboration between large language models and small language models (SLMs) is a key technical pathway for implementing cloud-edge intelligence, since it could address the conflict between the powerful capabilities of LLMs and the resource constraints of edge deployment. Chen et al. [6] summarized 5 collaboration modes for LLMs and SLMs, including pipeline collaboration, hybrid/routing collaboration, auxiliary/enhancement collaboration, knowledge distillation (KD) driven collaboration, and integration/fusion collaboration. Foundation models, pre-trained on massive datasets to capture general patterns, serve as more robust backbone systems. They enable efficient development of downstream AI applications, significantly reducing costs and resource requirements. For the cloud-based models, foundation models are also a common choice for their emergent general capabilities [3, 8, 9].

B. Continuous Optimization of Foundation Model

The continuous optimization of foundation models has evolved through three distinct yet interconnected paradigms: supervised fine-tuning (SFT) for knowledge injection, reinforcement learning from human feedback (RLHF) for alignment, and test-time adaptation for dynamic environment interaction. This progression reflects the systematic efforts of AI community to bridge the gap between static pretraining and real-world deployment challenges.

The foundational work on SFT established parameter-efficient adaptation as a cornerstone for model specialization. Subsequent innovations like token-based scaling [10] and fact-based scaling demonstrated how structured data generation enhances knowledge injection efficiency in LLMs, particularly for time-sensitive domains like sports analytics. However, the efficacy of SFT heavily depends on data quality, prompting solutions like RobustFT [11], which integrates multi-expert noise detection and entropy-based data selection to mitigate label noise. The emergence of parameter-efficient methods (e.g., LoRA [12] and prefix-tuning [13]) further revolutionized SFT by enabling low-rank adaptations that preserve base model capabilities while reducing catastrophic forgetting. These developments laid the groundwork for integrating SFT with reinforcement learning paradigms.

While SFT excelled at knowledge transfer, its inability to model human preferences led to the rise of RLHF frameworks. The OpenRLHF system [14] exemplifies this transition, coordinating four-model architectures for scalable preference learning. Theoretical breakthroughs like XPO [15] redefined exploration strategies through Q^*-approximation, achieving 7.5% improvement on AlpacaEval-2 benchmarks. However, heterogeneous human feedback introduced new challenges, addressed through personalized reward modeling [16] that combines representation learning with mechanism design for truthful preference aggregation. Crucially, Ref. [17] revealed the suboptimality of sequential SFT-RLHF training, proposing joint optimization to prevent knowledge forgetting, an insight bridging supervised and reinforcement paradigms.

The final frontier emerges in dynamic environments where models must adapt post-deployment. Reference [18] introduced closed-loop fine-tuning for traffic simulation, reducing covariate shift through trajectory-based reinforcement. Simultaneously, packing optimization [19] demonstrated how intelligent sequence combination enhances hardware utilization for models with 70 billion or more parameters. For persistent knowledge updates, Novel-WD [13] pioneered prefix-tuning architectures that encode new Wikidata facts without catastrophic forgetting, achieving 92.0% retention on temporal reasoning tasks. These innovations converge with contrastive meta-learning approaches that maintain differentiation capacity for out-of-distribution samples, completing the continuum from static pretraining to embodied cognition.

C. Embodied Intelligence

The vision-language navigation (VLN) [20] is one of the most popular embodied tasks which is designed to evaluate a robot's ability to navigate in unknown environments by following natural language instructions. Humans navigate efficiently in familiar environments primarily by constructing cognitive maps that integrate spatial information and visual cues such as landmarks [21, 22]. Similar to human behavior, robots perceive their surrounding environment through visual inputs (e.g., RGB-D) and move within novel environments based on linguistic instructions. Recent studies [23, 24] have demonstrated that leveraging foundation models [25], such as LLMs and vision-language models (VLMs), can enhance robot navigation performance in real-world environments. Specifically, LLMs can parse instructions into landmarks or executable code, capitalizing on their powerful language understanding capabilities, while VLMs are employed to process complex visual observations and ground the parsed linguistic instructions within the environment.

SayCan [26] demonstrated how LLMs extract and apply knowledge to accomplish physically-grounded tasks, including simple navigation and object manipulation. In VLN tasks, various approaches have attempted to employ LLMs as zero-shot, training-free navigators. For instance, NavGPT [27] pioneered the use of GPT-4 as a zero-shot navigator, converting visual observations of candidate viewpoints into textual descriptions that are subsequently processed by the LLM to determine the next action. Similarly, DiscussNav [28] introduced a specialized role division framework: ChatGPT handles instruction analysis, InstructBLIP [29] provides visual perception, and GPT-4 conducts completion assessment and decision testing. This collaborative framework achieves a modular and interpretable decision-making process. LLM-based methods also offer the additional advantage of transparency. Traditional VLN models are typically regarded as “black boxes”, making it difficult to understand the reasoning process behind an agent’s specific action predictions. In contrast, LLMs can clearly articulate their decision-making processes, providing researchers with deeper insights into the cognitive processes underlying navigation decisions.

Although the aforementioned approaches have demonstrated the feasibility of employing large language models as navigators, they still face numerous challenges in practical applications. Methods such as NavGPT and DiscussNav heavily rely on application programming interface (API) calls to advanced models like GPT-4, resulting in substantial operational costs. Meanwhile, using locally deployed large models demands high computational requirements from devices, imposing greater demands on edge devices and increasing equipment costs.

III. Parallel Deep Foundation Model

A. Cloudedgend Intelligence

To support the deployment, adaptation, and lifelong optimization of foundation models in real-world systems, we propose a structured architecture of Cloudedgend Intelligence. This architecture distributes models across three tiers, cloud, edge, and end, each fulfilling a distinct functional role: computation, coordination, and execution. By leveraging the complementary strengths of each layer, the system enables hybrid and adaptive intelligence.

Cloud system: Cloud systems focus on computation-centric tasks. They host resource-intensive models and form the backbone for large-scale processing and intensive training. This layer includes infrastructure models that enable distributed computation, data management, and secure communication; foundation models that provide general-purpose capabilities in understanding, reasoning, and generation; and field models that are developed to meet the specialized needs of specific domains or industries. Collectively, these models supply the core knowledge and computing power necessary for downstream applications.

Edge system: Edge systems act as the coordination layer, managing knowledge, resources, and processes across distributed environments. Organization models structure and integrate diverse resources, including knowledge bases, tool repositories, and service APIs, to support composable and modular intelligence. Coordination models handle collaboration, task assignment, conflict resolution, and incentive mechanisms, drawing inspiration from economic or organizational systems. Operation models manage and execute context-specific workflows, translating abstract tasks into actionable processes. Through these models, edge systems enable intelligent orchestration between the cloud and end layers.

End system: End systems provide execution-centric capacities at the frontlines of interaction with the physical world. Deployed on terminal devices, robots, or embedded systems, they operate under real-time constraints and limited resources. Interaction models support multimodal, real-time communication between humans and machines, as well as machine-to-machine communication. Optimization models capture feedback from users and the environment to support continuous learning and local adaptation. Execution models are responsible for enacting concrete actions with high efficiency and responsiveness in diverse scenarios.

These three layers operate in a tightly coupled and feedback-driven manner. Systems closer to the cloud emphasize intensive computation and model refinement, while systems near the edge and end focus on real-time perception, interaction, and task execution. When edge or end devices encounter complex, uncertain, or novel scenarios beyond their capacity, relevant data and context are transmitted back to the cloud for analysis and model enhancement. Cloud-side models then store these experiences and periodically perform retraining or adaptation, propagating the updated capabilities back to the edge and end layers to close the loop of continuous improvement.

B. Interaction of Front-End Model and Shadow Model

To achieve both real-time applicability and long-term optimization of foundation models, we introduce a parallel framework composed of front-end models and shadow models. These two models are structurally and functionally decoupled, yet closely linked through bidirectional interactions across both training and inference stages. This design embodies the principles of parallel intelligence, where the front-end model operates in the physical world, while the shadow model functions in a virtual space to explore and generalize. This collaboration is structured around three core processes inspired by parallel intelligence [1]: learning and training, experimental evaluation, and management and control. In the learning phase, large-scale data from both real and simulated environments are used to build robust general-purpose models. During experimental evaluation, domain-specific tasks are tested and refined in artificial societies to support specialization. Finally, in the control phase, deployment-focused optimization techniques are applied to ensure that models meet practical constraints across heterogeneous systems.

Front-end models are mainly deployed at the edge or end layer to handle real-time perception, interaction, and task execution. They are optimized for low-latency inference and situational awareness, enabling effective responses in dynamic and resource-constrained environments. Due to their direct engagement with physical systems and users, front-end models exhibit embodied cognition, a form of intelligence grounded in sensory-motor interaction with the physical world. Much like how a robot uses its sensors (e.g., cameras, LiDAR, and tactile input) to perceive its surroundings and adapt its behavior accordingly, embodied cognition allows AI systems to develop contextual understanding through continuous perception-action loops and real-time feedback.

In contrast, shadow models reside in the cloud and focus on analogical imagination and long-term learning. Built upon foundation models, they are continually enriched by data, episodic experiences, and performance feedback collected from deployed front-end models. Analogical imagination refers to the ability to reason about novel or uncertain situations by drawing parallels from past experience and prediction. Just as humans imagine new solutions by reconfiguring familiar knowledge, shadow models simulate virtual scenarios, explore counterfactuals, and abstract across domains to generate novel strategies, concepts, or configurations beyond the reach of the front-end alone.

Two types of interactions connect the front-end and shadow models: inference-time interaction and training-time interaction. These interactions form the foundation of a virtual-physical co-evolution mechanism, enabling real-time adaptability and long-term learning.

- **Inference-time interaction:** Inference-time interaction focuses on enhancing the model's input and reasoning process through both information-based and knowledge-based augmentation. Information-based interaction is exemplified by retrieval-augmented generation (RAG), where the front-end model formulates queries that are sent to the shadow model. The shadow model conducts large-scale retrieval and aggregation from structured databases or online sources, returning relevant information to improve the front-end model's inference. Knowledge-based interaction, on the other hand, leverages techniques such as mixture-of-experts (MoEs). The shadow model hosts a large pool of specialized experts and dynamically routes queries to the most relevant subsets based on the front-end task. This targeted expert activation allows the front-end model to benefit from rich, domain-specific knowledge without incurring high local computational cost.
- **Training-time interaction:** Training-time interaction supports model enhancement and continual optimization across three levels: data, feature, and gradient. Data-level interaction identifies hard or uncertain scenarios encountered by the front-end model and uses the generative capabilities of the shadow model to synthesize targeted examples, i.e., artificial intelligence generated content (AIGC). These challenging samples are then used to fine-tune the front-end model, strengthening its robustness. Feature-level interaction applies knowledge distillation techniques to transfer intermediate representations, such as reasoning chains or hidden features, from the shadow model to the front-end model, enriching its internal structure and inference capacity. Gradient-level interaction treats the shadow model as a supervisor: it evaluates the front-end model's outputs and provides feedback in the form of language-guided gradients, guiding the front-end model's update process even in settings where direct backpropagation is infeasible.

Together, these dual-channel interactions enable a tightly coupled loop between real-world embodiment and virtual abstraction. The front-end model is continually refined with targeted support from the shadow model, while the shadow model evolves through aggregated experiences from the field. This co-evolution supports scalable deployment and lifelong optimization of large foundation models across diverse applications.

C. Lifelong Learning in Parallel Mode

To enable continuous optimization and adaptation of foundation models in dynamic environments, we adopt the parallel intelligence framework, grounded in the ACP methodology: artificial societies, computational experiments, and parallel execution. This paradigm integrates three tightly coupled processes: learning and training, experimentation and evaluation, and management and control. Through iterative interaction between physical and virtual systems, models are developed, tested, and deployed in a closed loop, supporting scalable and sustainable intelligence. Within this framework, lifelong learning proceeds in three interconnected stages, each aligned with one phase of the ACP cycle.

The first stage is the general-purpose pre-training phase, corresponding to the learning and training process. In this stage, foundation models are trained on massive datasets generated from both real-world environments and high-fidelity virtual simulations. These diverse corpora provide rich knowledge for building commonsense reasoning and general understanding. Virtual environments enable the simulation of rare or complex scenarios that are difficult to capture in physical settings.

The second stage is the domain-specific fine-tuning phase, corresponding to experimentation and evaluation. Here, pretrained models are adapted to specialized tasks and domains by analyzing their performance in representative scenarios. Artificial societies provide a testbed for running controlled experiments, allowing the model to be selectively refined for domain-specific needs.

The third stage is the application-oriented post-tuning phase, corresponding to management and control. This stage focuses on practical deployment requirements. Techniques such as distillation, pruning, and quantization are applied to compress models into lightweight, deployable versions. These optimizations ensure that the model meets real-time constraints in latency, power, and compute resources without sacrificing reliability.

By structuring lifelong learning in parallel mode, foundation models evolve across generalization, specialization, and deployment stages. This approach ensures continuous adaptation and enables intelligent systems to operate effectively in diverse and changing real-world environments.

IV. Application

A. Foundation Low-Altitude System

In wide-area perception and multi-task scheduling scenarios involving low-altitude unmanned aerial vehicles (UAVs) [30], traditional LLM-based embodied UAV systems still face significant bottlenecks in real-time responsiveness, environmental adaptability, and task generalization. To address these challenges, as shown in Fig. 1, we introduce the PDFM architecture into an integrated “cloud-edge-end-aerial” low-altitude system, constructing a foundational intelligent framework that integrates perception, cognition, and control to support continuous operation and intelligent evolution in low-altitude environments.

As shown in Fig. 2, the front-end model is deployed on the UAV (edge-end) platform, which is responsible for executing high-frequency tasks such as VLN, target search, dynamic tracking, and multi-agent cooperative response, with an emphasis on real-time performance and environmental adaptability. Meanwhile, the shadow model resides in the cloud, continuously absorbing data, logs, and task feedback from UAVs to perform high-dimensional scene modeling, causal structure extraction, and knowledge transfer learning, enabling virtual evolution and strategy reconstruction for low-altitude tasks.

During the execution phase, the system relies on inference-time interaction mechanisms, leveraging techniques such as RAG and MoE to provide UAVs with instant contextual information and cross-domain knowledge support, enhancing task decision-making in emergencies and unfamiliar areas. During task intervals, the system activates training-time interaction mechanisms, with the shadow model leading the generation of complex examples, flight path optimization, and structural distillation, thereby promoting continuous optimization and generalization of the front-end model. This enables a cyber-physical closed loop from local task execution to cloud-based knowledge evolution, advancing the perception-cognition-action capabilities of low-altitude UAV systems along a trajectory that mimics a humanoid-like evolution.

B. VLN with Foundation Robot

Existing VLN systems face two primary challenges: (1) a lack of environmental modeling capabilities, which hinders effective understanding and adaptation to complex, dynamic environments; (2) insufficient long-term reasoning abilities, limiting performance in extended tasks. To address these issues, we propose an integration of world model-driven

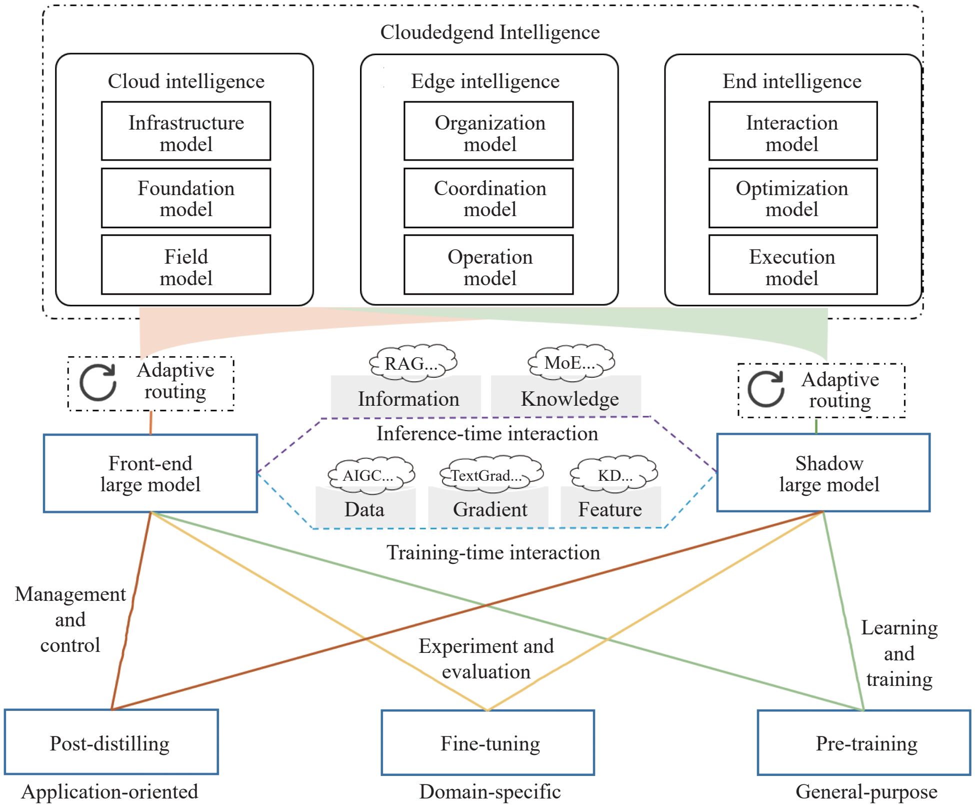


Figure 1 Framework of the proposed PDFM method.

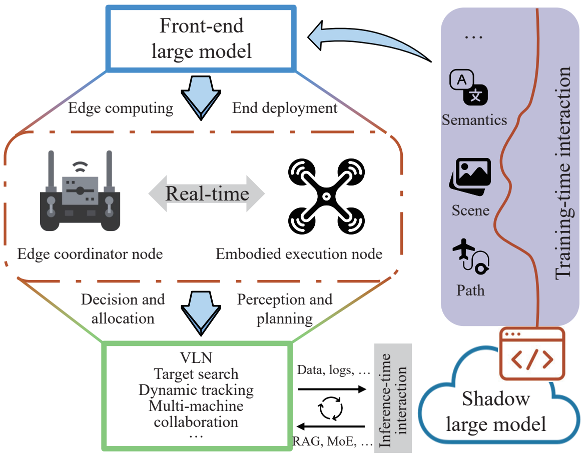


Figure 2 Foundation low-altitude system framework.

continuous learning to enhance the environmental understanding and decision-making capabilities of VLN systems, as shown in Fig. 3.

Dynamic prediction with cloud-based world model. By leveraging the representational learning capabilities of large-scale foundational models combined with temporal modeling, we construct a world model capable of understanding physical laws, spatial relationships, and causal logic. This world model takes current observations and trajectory as inputs to output probabilistic predictions of future environmental states, providing a forward-looking foundation for navigation decisions. Through this approach, the system can perform counterfactual reasoning and hypothesis verification, enhancing its ability to adapt to environmental changes.

End-side self-supervised continuous learning. Utilizing the dynamic predictions provided by the cloud-based world model, end-side devices can engage in self-supervised continuous learning. This mechanism allows the system to continuously optimize and refine its environmental model during navigation. Specifically, by comparing predicted states with actual observations, the system can identify areas of uncertainty within the model and conduct targeted exploration. This exploration not only helps improve the system's understanding of the environment but also enhances model accuracy and reliability without compromising navigation efficiency. Through ongoing learning and improvement, edge devices can better adapt to complex and changing environments, ensuring the successful execution of navigation tasks.

C. *Vision Language Action (VLA) with Foundation Robot*

The widespread application of existing VLA models is limited by two key factors: (1) the lack of generality in current VLA models, which hinders their ability to be transferred and deployed in various applications; (2) the current VLA models cannot be easily deployed on robotic terminals, as they still require significant computational power. Therefore, we propose the development of a universal cloud-edge-end collaborative VLA framework based on a parallel foundational model structure to address the issues of compatibility and computational demands for terminal robots.

Cross-platform general representation learning. Traditional VLA models typically output low-level control commands (such as joint angles or end-effector poses), resulting in strong coupling with the kinematics of specific robots. To enhance cross-platform generality, we propose a task-level action abstraction method, decomposing robotic actions into high-level semantic instructions (such as "grab object A") and platform-independent motion primitives. Specifically, the cloud-based VLA model generates semantic actions that are independent of the robot's configuration, while the edge device uses a lightweight adapter network to transform these abstract instructions into low-level control signals for the target robot in real-time. This adapter network employs a dynamically reconfigurable architecture that automatically adjusts parameters based on the kinematic characteristics of different robots, enabling zero-shot transfer of the same model across various robots.

Dynamic computation offloading and lightweight deployment. To achieve real-time inference on terminal devices, this framework adopts a stratified computation offloading strategy. The cloud hosts a high-performance VLA foundational model responsible for complex scene understanding and task planning, while the end-side device runs a lightweight student model focused on generating real-time control commands. The two models dynamically collaborate through the cooperation of edge-side: when the end model encounters unfamiliar scenarios (such as new objects or complex instructions), the edge-side model automatically triggers the involvement of the cloud model to ensure the reliability of task execution.

V. Conclusion

This letter presents the parallel deep foundation model, a framework designed to support the scalable deployment and continuous optimization of foundation models in real-world intelligent systems. By integrating cloud-edge-end

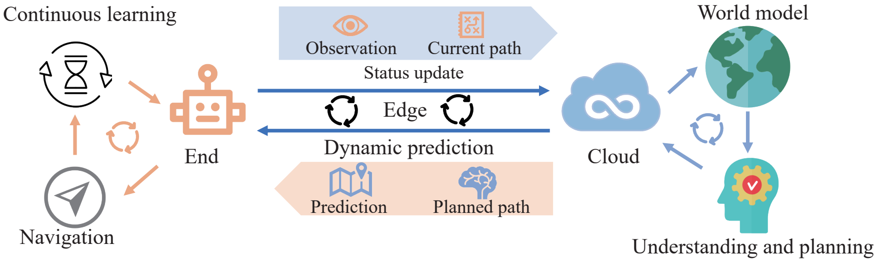


Figure 3 Cloud-edge-end collaborative navigation framework.

architecture with parallel intelligence principles, the PDFM framework enables a functional division of computation, coordination, and execution across system layers. The dual-model design, featuring a front-end model for real-time interaction and a shadow model for long-term improvement, establishes a virtual-real co-evolution mechanism through both inference-time and training-time interactions. Furthermore, a lifelong learning paradigm is established based on the ACP methodology, aligning general-purpose pre-training, domain-specific fine-tuning, and application-oriented post-tuning into a coherent process. This architecture offers a promising foundation for building adaptive, embodied, and continuously evolving intelligent agents in complex environments.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 62303460), the Science and Technology Development Fund of Macao SAR (Nos. 0145/2023/RIA3 and 0093/2023/RIA2), and the Young Elite Scientists Sponsorship Program of China Association for Science and Technology (No. YESS20220372).

References

[1] J. Yang, X. Wang, Y. Tian, X. Wang, and F.-Y. Wang, Parallel intelligence in CPSSs: Being, becoming, and believing, *IEEE Intell. Syst.*, 2023, 38(6), 75–80.

[2] J. Chen, S. Dai, F. Chen, Z. Lv, J. Tang, and L. Han, Edge-cloud collaborative motion planning for autonomous driving with large language models, in *Proc. 24th International Conference on Communication Technology (ICCT)*, Chengdu, China, 2024, 185–190.

[3] Y. Pan, Z. Su, Y. Wang, S. Guo, H. Liu, R. Li, and Y. Wu, Cloud-edge collaborative large model services: Challenges and solutions, *IEEE Netw.*, 2025, 39(4), 182–191.

[4] Y. Wang, C. Yang, S. Lan, L. Zhu, and Y. Zhang, End-edge-cloud collaborative computing for deep learning: A comprehensive survey, *IEEE Commun. Surv. Tutorials*, 2024, 26(4), 2647–2683.

[5] G. Wang, J. Liu, C. Li, Y. Zhang, J. Ma, X. Wei, K. Zhang, M. Chong, R. Zhang, Y. Liu et al., Cloud-device collaborative learning for multimodal large language models, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2024, 12646–12655.

[6] Y. Chen, J. Zhao, and H. Han, A survey on collaborative mechanisms between large and small language models, arXiv preprint arXiv: 2505.07460, 2025.

[7] P. Zhu and T. Yang, CE-LSLM: Efficient large-small language model inference and communication via cloud-edge collaboration, arXiv preprint arXiv: 2505.14085, 2025.

[8] X. Chen, Z. Guo, X. Wang, H. Yang, C. Feng, J. Su, S. Zheng, and T. Q. S. Quek, Foundation model based native AI framework in 6G with cloud-edge-end collaboration, arXiv preprint arXiv: 2310.17471, 2023.

[9] L. He, L. Fan, X. Lei, P. Fan, A. Nallanathan, and G. K. Karagiannidis, The road toward general edge intelligence: Standing on the shoulders of foundation models, *IEEE Commun. Mag.*, to be published.

[10] N. Mecklenburg, Y. Lin, X. Li, D. Holstein, L. Nunes, S. Malvar, B. Silva, R. Chandra, V. Aski, P. K. R. Yannam et al., Injecting new knowledge into large language models via supervised fine-tuning, arXiv preprint arXiv: 2404.00213, 2024.

[11] J. Luo, X. Luo, K. Ding, J. Yuan, Z. Xiao, and M. Zhang, RobustFT: Robust supervised fine-tuning for large language models under noisy response, arXiv preprint arXiv: 2412.14922, 2024.

[12] D. Biderman, J. Portes, J. J. G. Ortiz, M. Paul, P. Greengard, C. Jennings, D. King, S. Havens, V. Chiley, J. Frankle et al., LoRA learns less and forgets less, arXiv preprint arXiv: 2405.09673, 2024.

[13] M. Méloux and C. Cerisara, Novel-WD: Exploring acquisition of novel world knowledge in LLMs using prefix-tuning, arXiv preprint arXiv: 2408.17070, 2024.

[14] J. Hu, X. Wu, W. Shen, J. Liu, Z. Zhu, W. Wang, S. Jiang, H. Wang, H. Chen, B. Chen et al., OpenRLHF: An easy-to-use, scalable and high-performance RLHF framework, arXiv preprint arXiv: 2405.11143, 2024.

[15] T. Xie, D. J. Foster, A. Krishnamurthy, C. Rosset, A. H. Awadallah, and A. Rakhlin, Exploratory preference optimization: Harnessing implicit Q^*-approximation for sample-efficient RLHF, in *Proc. 13th International Conference on Learning Representations*, Singapore, Singapore, 2025.

[16] C. Park, M. Liu, D. Kong, K. Zhang, and A. Ozdaglar, RLHF from heterogeneous feedback via personalization and preference aggregation, arXiv preprint arXiv: 2405.00254, 2024.

[17] H. Fernando, H. Shen, P. Ram, Y. Zhou, H. Samulowitz, N. Baracaldo, and T. Chen, Mitigating forgetting in LLM supervised fine-tuning and preference learning, arXiv preprint arXiv: 2410.15483, 2024.

[18] Z. Zhang, P. Karkus, M. Igl, W. Ding, Y. Chen, B. Ivanovic, and M. Pavone, Closed-loop supervised fine-tuning of tokenized traffic models, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2025, 5422–5432.

[19] S. Wang, G. Wang, Y. Wang, J. Li, E. H. Hovy, and C. Guo, Packing analysis: Packing is more appropriate for large models or datasets in supervised fine-tuning, in *Proc. Findings of the Association for Computational Linguistics*, Vienna, Austria, 2025, 4953–4967.

[20] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, 3674–3683.

[21] M. M. Chun and Y. Jiang, Contextual cueing: Implicit learning and memory of visual context guides spatial attention, *Cogn. Psychol.*, 1998, 36(1), 28–71.

[22] R. A. Epstein, E. Z. Patai, J. B. Julian, and H. J. Spiers, The cognitive map in humans: Spatial navigation and beyond, *Nat. Neurosci.*, 2017, 20(11), 1504–1513.

[23] C. Huang, O. Mees, A. Zeng, and W. Burgard, Visual language maps for robot navigation, in *Proc. 2023 IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023, 10608–10615.

[24] D. Shah, B. Osinski, B. Ichter, and S. Levine, LM-Nav: Robotic navigation with large pre-trained models of language, vision, and action, in *Proc. Conference on Robot Learning*, Auckland, New Zealand, 2023, 492–504.

[25] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He et al., A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT, *Int. J. Mach. Learn. Cybern.*, to be published.

[26] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman et al., Do as I can, not as I say: Grounding language in robotic affordances, in *Proc. 6th Conference on Robot Learning*, Auckland, New Zealand, 2023, 287–318.

[27] G. Zhou, Y. Hong, and Q. Wu, NavGPT: Explicit reasoning in vision-and-language navigation with large language models, in *Proc. 38th AAAI Conference on Artificial Intelligence*, Vancouver, Canada, 2024, 7641–7649.

[28] Y. Long, X. Li, W. Cai, and H. Dong, Discuss before moving: Visual language navigation via multi-expert discussions, in *Proc. 2024 IEEE International Conference on Robotics and Automation (ICRA)*, Yokohama, Japan, 2024, 17380–17387.

[29] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, InstructBLIP: Towards general-purpose vision-language models with instruction tuning, in *Proc. 37th International Conference*

on Neural Information Processing Systems, New Orleans, LA, USA, 2023, 2142.

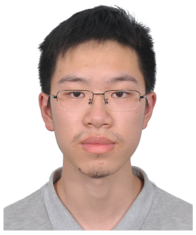
[30] Y. Tian, F. Lin, Y. Li, T. Zhang, Q. Zhang, X. Fu, J. Huang, X. Dai, Y. Wang, C. Tian et al., UAVs meet LLMs: Overviews and perspectives towards agentic low-altitude mobility, *Inf. Fusion*, 2025, 122, 103158.

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 Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, China. His research interests include parallel intelligence, autonomous driving, and intelligent transportation systems.

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, Macao, China. His research interests include parallel intelligence, large language models, and embodied agents.

Cong Wang received the BS degree from School of Automotive Studies, Tongji University, China, in 2024. He is currently pursuing the PhD degree at Institute of Automation, Chinese Academy of Sciences, China. His research interests include world models and 3D/4D generation and reconstruction.

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, Obuda University, Hungary. His research interests include intelligence testing, scenario generation, intelligent vehicles, and digital twins.



Zhiyao Luo received the BS degree in artificial intelligence from University of Electronic Science and Technology of China, China, in 2024. He is currently pursuing the MS degree at Institute of Automation, Chinese Academy of Sciences, China. His research interests include autonomous driving safety testing and intelligent transportation systems.

Fei-Yue Wang received the PhD degree in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990. He joined The University of Arizona in 1990 and became a professor and the director of the Robotics and Automation Laboratory and the Program in Advanced Research for Complex Systems. In 1999, he founded the Intelligent Control and Systems Engineering Center at Institute of Automation, Chinese Academy of Sciences, Beijing, China, under the support of the Outstanding Chinese Talents Program from the State Planning Council, and in 2002, was appointed as the director of the Key Laboratory of Complex Systems and Intelligence Science, Chinese Academy of Sciences, China. In 2011, he became the state specially appointed expert and the director of the State Key Laboratory for Management and Control of Complex Systems. His research interests include methods and applications for parallel intelligence, social computing, and knowledge automation. He is a fellow of INCOSE, IFAC, ASME, and AAAS. In 2007, he received the National Prize in Natural Sciences of China and became an Outstanding Scientist of ACM for his work in intelligent control and social computing. He received the IEEE ITS Outstanding Application and Research Awards in 2009 and 2011, respectively. In 2014, he received the IEEE SMC Society Norbert Wiener Award. Since 1997, he has been serving as the general or program chair of over 30 IEEE, INFORMS, IFAC, ACM, and ASME conferences. He was the president of the IEEE ITS Society from 2005 to 2007, the Chinese Association for Science and Technology, in 2005, the American Zhu Kezhen Education Foundation from 2007 to 2008, the vice president of the ACM China Council from 2010 to 2011, and the vice president and the secretary general of the Chinese Association of Automation (CAA) from 2008–2018. He was the founding editor-in-chief (EiC) of *The International Journal of Intelligent Control and Systems* from 1995 to 2000, the *IEEE ITS Magazine* from 2006 to 2007, the *IEEE/CAA Journal of Automatica Sinica* from 2014–2017, and the *China's Journal of Command and Control* from 2015–2020. He was the EiC of the *IEEE Intelligent Systems* from 2009 to 2012, the *IEEE Transactions on Intelligent Transportation Systems* from 2009 to 2016, and is the EiC of the *IEEE Transactions on Computational Social Systems* since 2017, and the founding EiC of *China's Journal of Intelligent Science and Technology* since 2019. Currently, he is the president of CAA's Supervision Council, IEEE Council on RFID, and vice president of IEEE Systems, Man, and Cybernetics Society.