摘要

近年来，以大型语言模型（LLMs）和视觉语言模型（VLMs）为代表的通用基础模型在数字域展现出较强的性能。然而，将基础模型的能力迁移至能感知物理环境、执行具身任务并与现实世界交互的机器人系统中，即跨越“数字智能”与“物理智能”之间的差距，仍是迈向通用人工智能（AGI）的关键挑战。当前具身模型在开放复杂场景中面临三方面核心瓶颈：精细化空间关系感知不足、长程时序建模能力薄弱、以及抽象推理与泛化困难，导致模型难以应对三维物理约束并完成长序列动态任务。

本文围绕上述难题展开了三方面工作。首先，本文给出了一种通用具身多模态认知大脑模型的构建路径。从开放物理世界的认知诉求出发，本文将具身认知划分为空间感知推理与时序因果建模两条线索，据此汇集了约 500 万条样本构建大规模混合微调数据集，并提出涵盖基础感知对齐、思维链注入和强化学习对齐的三阶段递进训练方案。十余项基准的评测结果显示，所提模型在大部分指标上跻身开源和闭源模型前列，为机器人系统奠定了兼备度量级空间理解与闭环时序规划的认知基础。

在此基础上，为检验该认知大脑在单台机器人上执行物理操作的实际效果，本文设计了一种端到端视觉语言动作（VLA）模型，其运作逻辑遵循“感知-思考-表征-执行”的思维链闭环。边界框、关键点、方向箭头等可视化基元被用来将大脑模型的高层空间推理意图显式表达出来，随后转换为几何约束信息，对底层流匹配动作专家的连续动作输出施加条件引导。实验证实，该设计使单体机器人在长程操作完成率、实时纠错和跨场景泛化方面都获得了可观改善。

最后，为考察认知大脑模型能否胜任更开阔的真实环境以及多台机器人的协同作业，本文构建了一套大小脑分层式多智能体协作框架。核心机制是时空本体统一记忆模块，负责将场景空间几何拓扑、历史执行记录与不同构型机器人的本体状态整合为持久化表征，让大脑模型的全局调度能力贯穿到小脑层面的技能执行中。本文在多类仿真平台和实体机器人上展开了充分实验，从终身适应性、协同可扩展性与系统鲁棒性三个切面，对大脑模型在开放条件下的高层任务编排与动态适应做出了系统评估。

总而言之，本文以具身认知大脑模型的构建为主线，沿“单本体端到端精准执行”至“多本体大小脑分层协同”的方向逐步推进，对具身大脑模型在开放物理世界中达成精准空间感知、长程任务规划及推理泛化执行的可行性与必要性进行了论证，期望为具身智能的后续探索提供参考。

关键词：具身智能，视觉语言模型，空间感知，长程时序建模，端到端控制，多机协同

Research Methodologies for Spatiotemporal Cognition in Embodied Foundation Models for Open-World Scenarios

Huajie Tan (Computer Application Technology)
Directed by: Prof. Huizhu Jia

ABSTRACT

Large Language Models (LLMs) and Vision-Language Models (VLMs), which are types of general foundation models, have made big progress in the digital domain. But transferring the knowledge they've built up into robotic systems that can sense physical surroundings, do embodied tasks, and interact with the real world is a big pressing bottleneck on the way to Artificial General Intelligence (AGI). You can see that existing embodied models meet with three main difficulties when working in open and complex environments. Firstly, their understanding of fine-grained spatial relationships is still very coarse. Temporal reasoning across long time horizons often gets fragmented, which was kind of surprising to us. And they are not good at abstract reasoning, and can't use what they know in unfamiliar situations. This holds them back a lot. All these limitations together stop such models from doing long-sequence dynamic tasks under three-dimensional physical constraints.

We use three contributions to tackle the above difficulties in this thesis. First, we put forward a construction pathway for a general-purpose embodied multimodal cognitive brain model. Embodied cognition is organized along two axes, spatial perception reasoning and temporal causal modeling. And we gather around 5 million samples to make a large-scale hybrid fine-tuning dataset. That's a huge amount of data, which was kind of surprising. Then we introduce a three-stage progressive training scheme. The first stage covers perceptual alignment. The second one focuses on chain-of-thought injection. The third uses reinforcement learning alignment. We did evaluations on more than ten spatial and temporal benchmarks. You can see that the proposed model ranks among the top open-source alternatives on most metrics. It establishes a cognitive substrate that plays an important role in integrating metric-level spatial understanding with closed-loop temporal planning for robotic systems.

And second, we want to check if the cognitive brain can bring tangible results when a single robot does physical execution tasks, so we design an end-to-end Vision-Language-Action

(VLA) model. This model follows a "See-Think-Sketch-Act" chain-of-thought closed loop. Visual primitives like bounding boxes, keypoints, and directional arrows externalize the brain model's high-level spatial reasoning intentions. And these primitives are converted into geometric constraints. You can see that these constraints play an important role in conditionally guiding a flow-matching action expert to generate continuous actions. We ran experiments to test this setup. The experiments show big gains in long-horizon task completion. That was kind of surprising. We also see big gains in real-time error correction. And cross-scenario generalization has big gains too.

Third, we want to find out if the cognitive brain can scale to wider real-world settings and coordinated multi-robot operation, so we built a hierarchical Brain-Cerebellum multi-agent collaboration framework. Its core is a unified Spatio-Temporal-Embodiment Memory module which plays an important role in consolidating scene geometry, execution history, and proprioceptive states of morphologically diverse robots into a persistent representation. And this allows the brain model's global scheduling capabilities to permeate down to cerebellum-level skill execution. We do comparative experiments across multiple simulation platforms and physical robots, they evaluate high-level task orchestration and dynamic adaptation from three perspectives. Firstly, we check lifelong adaptability. Secondly, we look at collaborative scalability. Thirdly, we test system robustness. I think these three points are important, important because they cover all key parts of multi-robot collaboration.

We can see that this thesis takes the construction of an embodied cognitive brain model as its guiding thread, it moves from "single-embodiment end-to-end precise execution" toward "multi-embodiment hierarchical Brain-Cerebellum collaboration." That's its core focus. I think it shows embodied brain models which are both feasible and necessary for accurate spatial perception in the open physical world. And they also play an important role in long-horizon task planning. You can notice they support generalized reasoning too. This provides important theoretical grounding and important practical ways for people to keep exploring embodied intelligence.

KEY WORDS: Embodied AI, Vision-Language Models, Spatial Perception, Long-Horizon Temporal Modeling, End-to-End Control, Multi-Agent Collaboration