

王培东

EMBODIED AI · VLA · LLM REINFORCEMENT LEARNING

☎ (+86) 156-7039-2798 | ✉ pdongwang@163.com | 📺 Control-derek | 🏠 Peidong Wang

Profile

东北大学计算机科学直博生（2025.09 - 至今，导师：冯时副教授），研究聚焦具身智能、视觉语言动作模型（VLA）与大语言模型强化学习；曾于 **微软亚洲研究院（MSRA）** 开展 Embodied Agent 与 LLM RL 研究，现于 **行云至理研究院（XYZ AI Lab）** 担任基模实习生，参与 Code Agent 预训练、RL 与数据闭环；发表 **4 篇 CCF-A**、**3 篇 CCF-B** 论文，其中一作 **3 篇 CCF-A**、**1 篇 CCF-B**，发表于 ACL、EMNLP、ACM MM、COLING 等；具备从算法研究、规模化训练到真实机器人部署的完整经验。

Embodied & Robotics Experience

SHAPER | Self-Evolving Embodied Agents via Skill-Harness Evolution

Embodied Agent / arXiv

一作 | 微软亚洲研究院（MSRA） | EMBODIED AGENT、VLA、AGENT EVOLUTION

2026.08

- **方法**：冻结模型参数与底层执行器，通过目标环境 Rollout 自进化 Skill 与 Harness，使具身智能体无需更新权重即可适配目标环境。
- **结果与开源**：VLABench 成功率 **28.25% → 34.50% (+6.25 pp)**，ESI-Bench 准确率 **32.5% → 49.8% (+17.3 pp)**；代码被合入 Agent Lightning 官方仓库并作为官方 recipe 开源；该仓库累计 **17.8k+ stars**。

SO-ARM101 VLA 行为克隆与在线强化学习真机系统

BC / Online RL

独立项目 | LeROBOT、PyTorch、ACT、SmolVLA、 $\pi_{0.5}$ 、OPENPI、RESIDUAL RL、RLT

2026.07 - 2026.08

- **系统与 BC**：搭建 6 轴主从机械臂、双相机 30 Hz 遥操作与远端推理系统；采集 120 条（25,316 帧）药盒堆叠示范，构建 LeRobot 数据审计、静止帧裁剪与训练集归一化流水线；完成 $\pi_{0.5}$ 全参 BC 微调，在 13 种摆位首批真机测试中完成 **39/39 (100.0%)**。
- **在线 RL**：针对 BC 在“抓取玩具剑并精确放入剑架槽位”任务中易偏放、滑落的问题，基于冻结 $\pi_{0.5}$ 与离线训练的 RL Token，实现抓取/放置两阶段 residual Actor-Critic (TD3) 在线学习，将真机成功率从 **16.7% (2/12)** 提升至 **75.0% (9/12)**。

全国大学生机器人大赛 ROBOCON 2022 | 同创辉煌 / 机器马术

全国一等奖（亚军） / 三等奖

运动控制（大一） | STM32、PID、MPC、KALMAN FILTER

2021.10 - 2022.07

- **目标击打**：融合 Brown 三次指数平滑与 Kalman Filter 预测轨迹、补偿时延，并以三次 B-spline/minimum-snap 规划双机轨迹。
- **四足平衡**：设计动量轮 MPC 与 PID 控制器，推导三状态、三步预测 MPC，以 Eigen/OSQP 构造约束 QP，完成 5 ms Linux-STM32F407 闭环联调；参与 STM32F407 嵌入式控制开发、足端轨迹规划开发。

LLM & RL Research Experience

微软亚洲研究院（MSRA）

上海

研究实习生 | LLM REINFORCEMENT LEARNING、RLVR/GRPO | MENTOR: XUFANG LUO

2025.06 - 2026.05

- 从“更新高熵 token 提升推理性能”与“低概率 token 主导梯度导致训练不稳”的矛盾发现出发，识别不同 rollout 更新次数造成的 off-policy 程度差异；推导 IS-ratio 方差与 token 概率的关系及 clipped GRPO 条件梯度期望，揭示低/高概率 token 的梯度主导反转。
- 据此提出 ACPO：按行为策略下的 token 概率分桶，依据桶内 IS-ratio 标准差自适应设定 clipping 边界，缓解固定全局边界引起的不均匀裁剪与梯度偏置；在 **3B/7B** 模型的数学、表格问答与逻辑推理任务及 near/off-policy 设置下，整体优于 DAPO、CISPO 等基线。

行云至理研究院（XYZ AI Lab）

上海

基模实习生 | CODE AGENT RL、数据与训练闭环 | MENTOR: XUFANG LUO

2026.05 - 至今

- 参与 Code Agent RL baseline 与训练闭环建设，负责四类 agentic 代码数据的 filter/render/dedup/score；完成 **2.32 亿 commit** 样本打分，构建 relaxed/strict 分层质量门控，筛除 **62.9%-94.4%** 未满足质量规则的任务候选，支撑 pre-train/mid-train 数据筛选。
- 作为作者参与 AI4AI at Scale 技术报告，总结 bounded-exploration、verification-gated 的 agentic system optimization 全链路；团队发布 XYZ-Aquila-mini/pro 开源 Deep Search Agent 及训练、评测与回放记录。

SAFE-QAQ | End-to-End Slow-Thinking Audio-Text Fraud Detection via RL

ACL 2026

一作 | 东北大学 × 中国移动互联网 | MULTIMODAL RL、AUDIO REASONING、GRPO

2025.03 - 2025.04

- 构建端到端音频-文本慢思考反诈框架，直接建模原始语音中的语调、情绪与环境声线索；基于 GRPO 设计 accuracy/format/deep-reasoning 规则奖励，并结合 rejection sampling 与 length-constrained RL 优化推理效率。
- 与中国移动合作将 SAFE-QAQ 部署至反诈业务生产链路，日处理超过 **7 万通** 电话；基于音频分块与阶段奖励实现实时风险判定，平均检测耗时 **8.98 s**；TeleAntiFraud-Bench 平均分类 F1 达 **87.49**，推理链缩短 **48.87%**。

Selected Research & Publications

2026.08	Self-Evolving Embodied Agents via Skill-Harness Evolution	Embodied Agent · VLA · RSI	AAAI 审稿中
2026.07	AI4AI at Scale: A Full-Pipeline System for Enhancing LLM Agentic Capabilities	RSI · Foundation Model	Tech Report
2026.04	SAFE-QAQ: End-to-End Slow-Thinking Audio-Text Fraud Detection via Reinforcement Learning	Multimodal RL	ACL 2026
2026.02	What Are Key Factors for Updates in RL for LLM Reasoning?	LLM RL · Off-Policy Optimization · Theoretical Analysis	NeurIPS 审稿中
2025.08	Language Models as Continuous Self-Evolving Data Engineers	Recursive Self-Improvement (RSI) · Data Agent	EMNLP 2025
2025.07	TeleAntiFraud-28k: An Audio-Text Slow-Thinking Dataset for Telecom Fraud Detection	Dataset · Benchmark	ACM MM 2025
2024.05	STICKERCONV: Generating Multimodal Empathetic Responses from Scratch	MLLM · Multi-Agent · Empathetic Dialogue	ACL 2024

Skills & Additional Awards

具身/机器人 LeRobot、OpenPI、Embodied Agent、ACT、SmolVLA、 $\pi_{0.5}$ 、Residual RL、RLT、SO-ARM101、STM32、PID、MPC

LLM/RL PyTorch、Transformers、RLVR/GRPO、Code Agent、Training Data Curation

其他奖项 CAIL 2024 司法主观题赛道二等奖；2023 中国大学生计算机博弈大赛一等奖