# 摘要

随着人工智能的发展，其需要解决的任务逐渐从一些静态的感知生成走向真实世界的动态交互，构建能够在交互环境中理解环境、做出决策并执行行动的多模态智能体，已成为实现下一代人工智能的关键路径之一。多模态智能体需要在动态变化的交互环境中同时处理视觉、语言等多种模态的信息，并基于对环境的理解做出决策和执行动作。然而，从多模态理解到交互环境中的有效行动，智能体在“感知表征—规划决策—物理执行”的能力链条上面临三个层面的核心挑战：在表征层面，现有多模态智能体的视觉表征存在语义纠缠和组合性不足的问题，难以在动态环境中实现可泛化的理解；在决策层面，现有的智能体决策模型主要依赖静态先验知识进行推理，缺乏通过环境交互获取新经验并持续适应的机制；在对齐层面，基于视觉-语言-动作模型的智能体在二维视觉观测与三维物理空间之间缺乏有效对齐，导致动作执行精度不足。

本文围绕上述挑战，沿着交互环境复杂度递增的路线，系统性地提出了相应的解决方案。本文的主要贡献如下：

（1）针对表征层面的挑战，本文在通信涌现的合成交互环境中系统研究了可泛化结构化表征的形成机制。一方面，提出符号映射架构，在环境输入层引入显式的符号化映射层将输入特征解耦为独立的属性符号表征，并结合任务迁移课程学习方法实现渐进式训练，使智能体在复杂多轮对话游戏中的组合泛化成功率大幅提升至 85%，揭示了输入侧语义化编码对组合泛化的关键促进作用。另一方面，将泛化研究从单物体属性扩展到多物体间的抽象空间关系，通过设计不同差异程度的视觉环境输入，系统性地揭示了交互环境中功能性压力对抽象关系概念表征形成的驱动机制，并验证了涌现语言作为离散环境表征在下游多步马尔可夫决策过程任务中的迁移优势。

（2）在上述表征形成机制研究的基础上，本文将解耦语义化表征的思想从合成环境推进到真实世界视频场景，提出 VideoOrion 视频大语言模型。该模型采用双分支架构，其中对象中心分支通过“检测-分割-追踪”流水线协同多个专家视觉模型，从视频输入中显式提取各物体的时空动态特征并编码为紧凑的对象词元，实现了视频大语言模型中基于显式物体提取的对象级特征解耦表征方案。VideoOrion 在多个通用视频问答基准上相比同等基线模型取得了 10%–15% 的显著提升，在视频指称理解基准上大幅超越已有方法。

（3）针对决策层面的挑战，本文提出 LLaMA-Rider 探索式决策学习框架。该框架的核心创新在于构建了一条从主动探索到经验内化的完整路径：探索阶段通过多轮“反馈-修正”机制利用环境反馈引导大语言模型智能体逐步修正决策，学习阶段提出子任务重标注方法，利用任务的组合性结构从探索轨迹中挖掘出结构化训练数据，再通过

参数高效微调将探索经验内化到模型参数中。仅需约 1300 条探索数据的学习即可使智能体在复杂的 Minecraft 开放世界中的任务完成数量从 16 个提升至 25 个（共 30 个任务），在多数任务上超越了拥有更丰富环境知识的 ChatGPT 规划器，并能泛化到从未探索过的更困难的长程任务。

（4）针对对齐层面的挑战，本文提出 VIPA 视觉-物理对齐空间感知预训练范式。该范式的核心创新在于在视觉-语言预训练与机器人数据微调之间插入显式的视觉-物理对齐阶段。该阶段利用大规模人类演示视频提取二维视觉和三维空间对齐的监督信号，无需昂贵的机器人操作数据即可建立视觉空间与物理空间的对应关系。同时，提出双编码器架构，在语义视觉编码器之外引入三维视觉编码器并通过交叉注意力融合层实现语义特征与空间特征的有效融合。VIPA-VLA 在仿真基准上在不使用任何机器人预训练数据的条件下达到了与强基线可比的性能；在真实机器人操作任务中取得了最佳的整体表现，并在未见环境中展现出显著的泛化优势。

总而言之，本文的研究从多智能体通信环境出发，依次扩展到真实世界视频、虚拟开放世界和真实物理世界，完成了从结构化表征形成、对象级视频理解、开放世界决策到具身行动执行的完整研究链条，为构建能够在交互环境中有效理解、决策和行动的多模态智能体提供了方法支撑。

关键词：多模态智能体，交互环境，多模态理解，决策与探索，具身智能