摘要

随着视觉观测、语言反馈、动态环境关系和异质演示数据共同进入智能体学习过程，训练系统需要处理两类中介问题：信息进入学习回路时的形态差异，以及学习更新面对不同关系、反馈和样本时的责任差异。视觉输入具有连续、局部和多尺度结构，语言反馈包含任务阶段语义，动态关系会随干扰发生漂移，离线演示数据也存在来源、形态和执行结果不均衡。视觉单元划分不当会放大序列学习难度，反馈语义过粗会削弱策略修正方向，关系归因缺失会使局部环境变化干扰稳定更新，样本质量差异则会使不同价值的演示承担相近梯度权重。若缺少面向学习更新的组织方式，学习算子难以稳定吸收有效信息，梯度更新也难以区分不同对象的更新责任。本文以学习更新为中介，围绕表征适配与决策归因展开研究，分析从信息组织到更新分配的关键机制。

在输入侧表征适配方面，本文首先提出 Being-VL，研究视觉信息进入统一生成模型时的输入单元组织问题。Being-VL 通过基于字节对编码的视觉词元化方法、优先级引导编码机制、词表扩展和多阶段训练策略，将连续视觉内容整理为具有明确边界的视觉语义单元，并将二维局部结构保留到可学习词元序列中。该方法把视觉内容从高维连续特征整理为离散语义单元，使统一生成模型能够以接近文本序列的方式处理视觉输入，同时保留图像块之间的局部组合关系。理论分析和实验结果说明，保留局部结构的显式词元化能够降低视觉信息的学习难度，并改善训练稳定性与视觉理解性能。由此，视觉信息不再只是作为外部特征进入多模态模型，而是被组织为能够参与统一序列学习的表征单元。

在反馈侧表征适配方面，本文提出 AdaRefiner，研究语言反馈与策略更新之间的语义错配问题。长任务链决策中的原始反馈通常混合奖励变化、阶段性成就和未完成目标，直接用于训练时难以反映当前策略最需要修正的行为环节。AdaRefiner 由适配语言模型、决策语言模型和理解分数组成，将行为轨迹、奖励变化和阶段性成就整理为更贴近当前策略能力的目标语义，使反馈从原始评价转化为可用于策略修正的语义目标。其中，适配语言模型负责从交互轨迹中提炼子目标语义，决策语言模型将语义目标转化为策略可执行的训练信号，理解分数衡量轨迹与目标之间的匹配程度。该方法说明，表征适配可以覆盖输入端、反馈端和监督端。

在关系层面的决策归因方面，本文提出 COREP，研究动态环境中状态关系变化导致的更新依据不稳定问题。COREP 通过环境共享并图、因果起源表示和选择性关系更新，将策略学习建立在跨环境保留的关系骨架之上，降低局部关系漂移对更新的干扰。环境共享并图用于汇集不同环境中反复出现的状态依赖，因果起源表示刻画关系结构

对当前状态的贡献，选择性关系更新则在环境变化时保留稳定关系并调整受扰动关系。由此，策略更新能够从平均依赖所有状态关系，转向优先利用对任务迁移和动态适应更稳定的关系成分。该方法把更新责任从整体状态表示细化到关系结构内部，为动态环境下的稳定策略学习提供了关系归因机制。

在输入单元、反馈语义和关系结构三个层面之后，本文进一步提出 PTR，研究视觉语言动作模型离线后训练中的后验归因问题。面对多来源、多形态演示数据，动作监督的价值常常取决于执行结果是否清晰、动作片段是否可归因以及样本来源是否可靠。PTR 利用候选集识别构造动作执行结果的后验分数，并将其映射为保守式权重，从而控制不同演示样本在动作监督中的梯度预算；同时引入信念词元器（BeliefTokenizer）改善中间状态组织，使视觉、语言、动作和执行结果能够在同一后训练流程中协同参与更新。保守式权重降低结果含糊样本或低可归因样本的更新强度，使模型更集中地利用动作结果清晰、监督价值较高的演示片段。由此，PTR 将中间状态适配与后验样本归因结合起来，把表征适配和决策归因推进到异质演示数据下的视觉语言动作模型后训练场景。

本文在统一生成式多模态建模、长任务链决策、动态环境强化学习以及视觉语言动作模型后训练等场景中进行了系统验证。实验覆盖理论命题检验、训练动态诊断、行为案例分析、受控扰动评测、跨形态联合训练和真实机器人测试，既考察主任务性能，也考察训练稳定性、扰动退化、样本利用差异和真实执行可靠性。结果表明，Being-VL 能够改善视觉词元进入统一生成模型时的训练稳定性，AdaRefiner 能够提升语言反馈对长任务链策略学习的引导效果，COREP 能够在关系漂移条件下获得更稳定的策略更新，PTR 能够在异质演示数据和复杂机器人任务中更有区别地利用样本监督。消融实验和诊断分析显示，输入侧单元组织、反馈侧语义整理、关系层更新分配和后验样本加权分别对应智能体学习中不同类型的中介瓶颈。全文结果说明，智能体学习的有效更新依赖于合适的表征适配，也依赖于对关系结构、反馈语义和后训练样本的有原则归因；二者共同决定异质信息能否被稳定吸收，并决定学习责任能否被合理分配。

关键词：智能体学习，表征适配，决策归因，视觉词元化，后验归因

Representation Adaptation and Decision Attribution for Agent Learning

Wanpeng Zhang (Computer Applied Technology)
Directed by Prof. Zongqing Lu

ABSTRACT

As visual observations, language feedback, dynamic relational structures, and heterogeneous demonstrations enter agent learning, a training system must handle two intermediary issues: representational differences when information enters the learning loop, and responsibility differences when updates are assigned to relations, feedback, and samples. Visual inputs are continuous, local, and multi-scale; language feedback carries task-stage semantics; dynamic relations may drift under perturbations; and offline demonstrations vary in source, embodiment, and execution outcome. Improper visual unitization increases the difficulty of sequence learning, coarse feedback weakens the direction of policy correction, missing relation attribution lets local environment changes disturb stable updates, and uneven demonstration quality can assign similar gradient weights to samples with different values. Without organization oriented toward learning updates, the learning operator can hardly absorb reliable information, and gradient updates can hardly distinguish the responsibility carried by different objects. Taking learning updates as the intermediary concept, this dissertation studies the mechanisms from representation adaptation to decision attribution.

For input-side representation adaptation, this dissertation first proposes Being-VL to study how visual information is organized before entering a unified generative model. Being-VL introduces BPE-based visual tokenization, priority-guided encoding, vocabulary expansion, and multi-stage training to convert continuous visual content into visual semantic units with explicit boundaries, while preserving two-dimensional local structure in a learnable token sequence. The method converts high-dimensional continuous visual features into discrete semantic units, allowing a unified generative model to process visual inputs in a sequence form closer to text while retaining local compositional relations among image patches. Theoretical analysis and experiments show that locality-preserving explicit tokenization reduces the learning difficulty of visual information and improves training stability and visual understanding. Thus, visual in-

formation is organized as representation units that can participate in unified sequence learning, instead of entering the multimodal model only as external features.

For feedback-side representation adaptation, this dissertation proposes AdaRefiner to address the semantic mismatch between language feedback and policy updates. Raw feedback in long-horizon decision tasks often mixes reward changes, intermediate achievements, and unfinished goals, and therefore may not reveal which behavior should be corrected at the current learning stage. AdaRefiner uses an adapter language model, a decision language model, and a comprehension score to transform trajectories, reward changes, and achievements into target semantics aligned with the current policy capability, turning raw evaluation into semantic objectives for policy correction. The adapter language model extracts subgoal semantics from interaction trajectories, the decision language model turns such semantics into trainable policy signals, and the comprehension score measures the semantic match between a trajectory and the target. The method shows that representation adaptation covers the input, feedback, and supervision sides.

For relation-level decision attribution, this dissertation proposes COREP to address unstable update evidence caused by changing state relations in dynamic environments. COREP constructs an environment-shared union graph, learns causal-origin representations, and performs selective relation updates so that policy learning can rely on relation structures preserved across environments and reduce the disturbance caused by local relational drift. The union graph gathers state dependencies that repeatedly appear across environments, causal-origin representations describe how relation structures contribute to the current state, and selective updates preserve stable relations while adjusting perturbed ones. Thus, policy updates prioritize relational components that are more stable for transfer and adaptation, instead of treating all state relations uniformly. The method refines update responsibility from the whole state representation into internal relational structures, providing a relation-attribution mechanism for stable policy learning in dynamic environments.

After input unit organization, feedback semantic refinement, and relation attribution have been established, this dissertation further proposes PTR to study posterior attribution in offline post-training of vision-language-action models. With multi-source and cross-embodiment demonstrations, the value of action supervision often depends on whether the execution outcome is clear, whether the action segment is attributable, and whether the data source is reliable. PTR identifies action execution outcomes through a candidate set, converts the recognition posterior into a PTR score, and maps the score to conservative weights that control the

gradient budget assigned to different demonstrations. PTR also introduces a belief tokenizer to improve intermediate-state organization, enabling visual inputs, language instructions, actions, and execution outcomes to participate in the same post-training loop. Conservative weights reduce the update strength of ambiguous or weakly attributable samples, so the model can focus more on demonstrations with clearer execution outcomes and higher supervision value. In this way, PTR combines intermediate-state adaptation with posterior sample attribution and extends representation adaptation and decision attribution to VLA post-training with heterogeneous demonstrations.

The proposed methods are evaluated in unified multimodal generation, long-horizon decision tasks, non-stationary reinforcement learning, and VLA post-training. The evaluations cover theoretical verification, training diagnostics, case analysis, perturbation tests, cross-embodiment training, and robot assessment. They examine task performance, training stability, perturbation degradation, sample-use differences, and execution reliability. The results show that Being-VL improves the training stability of visual tokens in unified generative modeling, AdaRefiner improves language-feedback guidance in long-horizon policy learning, COREP yields more stable policy updates under relational drift, and PTR uses heterogeneous demonstrations more selectively in complex robot tasks. Ablations and diagnostics further show that input-side unit organization, feedback-side semantic refinement, relation-level update allocation, and posterior sample weighting correspond to distinct intermediary bottlenecks in agent learning. Overall, effective agent learning requires both representation adaptation and principled attribution over relational structures, feedback semantics, and post-training samples. Together, they determine whether heterogeneous information can be absorbed stably and whether learning responsibility can be allocated properly.

KEYWORDS: agent learning, representation adaptation, decision attribution, visual tokenization, posterior attribution