ICS 35.240
CCS L 70

团 体 标 准

T/CCF 0002—2025

强化学习系统 第2部分：强化学习环境技术要求

Reinforcement learning system part 2: Technical requirements for reinforcement learning environment

2025－06－11 发布 2025－06－11 实施

中 国 计 算 机 学 会 发布

目　　次

前　　言 II
引　　言 III
强化学习系统 第2部分 强化学习环境技术要求 1
1　范围 1
2　规范性引用文件 1
3　术语和定义 1
4　符号和缩略语 3
5　功能架构 3
6　场景适配 4
6.1　概述 4
6.2　仿真器适配 4
6.2.1　功能描述 4
6.2.2　仿真器的场景适配接口规范 4
6.2.3　数据格式 5
6.3　源数据适配 7
6.3.1　功能描述 7
6.3.2　源数据的场景适配接口规范 7
6.3.3　数据格式 8
6.4　其他要求 9
附　录　A （资料性） 强化学习环境在训练和推理工作流的适配 10
A.1 强化学习环境在训练工作流的适配 10
A.2 强化学习环境在推理工作流的适配 11
附　录　B （资料性） 常见范式中的强化学习环境 11
B.1 在线强化学习 11
B.2 离线强化学习 11
B.3 面向多智能体的强化学习 12
B.4 面向大语言模型的强化学习 12
B.5 面向数学、物理、化学、生物等基础科学的强化学习 13
B.6 面向具身智能的强化学习 13
附　录　C （资料性） 面向围棋的强化学习环境示例 14
参考文献 16

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中国计算机学会标准工作委员会提出并归口。

本文件起草单位：腾讯科技（成都）有限公司、北京大学、中国科学技术大学、新华出版社、成都实娱商业管理有限公司、上海交通大学、清华大学、北京中关村人工智能研究院、西安交通大学、中科院自动化所、电子科技大学、浙江大学、西南交通大学、南京大学、中国电力科学研究院有限公司、四川大学、武汉大学、复旦大学、中山大学、杭州宇树科技有限公司、数字图书馆教育部工程研究中心、哈尔滨工业大学（深圳）、北京航空航天大学、华中科技大学、北京邮电大学、OPPO广东移动通信有限公司、燧原智能科技（成都）有限公司、摩尔线程智能科技(北京)股份有限公司、重庆邮电大学、西南民族大学、四川具身人形机器人有限公司。

本文件主要起草人：邓民文、叶振斌、汪永毅、杨耀东、李文新、刘林、张伟楠、周文罡、张海峰、杨巍、覃洪杨、赵鉴、许华哲、万里鹏、张寅、鲁云龙、李凌峰、匡乐成、王永霞、周丹丹、谢宁、邢焕来、季向阳、汪文俊、曹相成、陶吕方、黄蓝枭、林夏、姜羿、张浩哲、李厚强、高阳、兰旭光、吕建成、王新迎、余超、何召锋、徐建、吴秉昊、赵卫东、温颖、宋麟、梁超、章欣、杨丰、李梦露、汤臣薇、石荣晔、吴文峻、刘渝、周可、王进、蒋溢、蔡英、涂钥轩。

引　言

作为一种重要的人工智能方法，强化学习在具身智能、大语言模型、自然科学研究、游戏AI等众多领域得到了广泛应用。

强化学习是智能体通过和环境交互收集数据来优化决策的机器学习范式。智能体执行动作并从环境中接收观测、奖励等；环境则接收智能体的动作，以一定机制更新当前状态并向智能体反馈观测、奖励等。在不同的强化学习算法中，智能体利用环境数据学习的方式有诸多差异，因此强化学习算法的运行依赖框架的实现。综上，智能体、环境、框架为构成强化学习系统的三个基本要素。

尽管强化学习对既有数据依赖度低，在求解问题上普适性强，于当下有丰硕的应用成果，于未来有广阔的应用前景。然而在缺乏相关标准的情况下，强化学习算法开发和应用落地存在诸多挑战。一方面，场景、算法强耦合，工程可复用度低，重复开发浪费人力；另一方面，训练、部署难迁移，应用接入成本高，服务质量难以保障。

本系列标准将为强化学习系统设计和应用生态构建提供参考和依据，推动环境开发者、算法开发者、应用开发者、平台运营者在统一的框架下展开协作。标准规范的设计保证系统各功能模块即插即用，从而减少重复开发、降低应用成本、提升服务质量，进而实现技术生态协同，提高资源利用效率，使强化学习技术惠及更多领域发展，本系列标准由4个部分构成。

——第1部分：通用要求。目的在于确立强化学习系统的参考架构，规定通用技术要求。

——第2部分：强化学习环境技术要求。目的在于确立强化学习环境的参考架，规定其接口和数据格式要求。

——第3部分：强化学习智能体技术要求。目的在于确立强化学习智能体的参考架构，规定其接口和数据格式要求。

——第4部分：强化学习框架技术要求。目的在于确立强化学习框架的参考架构，规定其接口和数据格式要求。

强化学习系统　第2部分：强化学习环境技术要求

1　范围

本文件规范了强化学习系统中强化学习环境的技术要求，包括仿真器的接入、场景适配的开发实现流程、以及交互协议和数据格式等。

本文件适用于指导强化学习环境的开发、接入、适配、部署全流程设计与构建。

2　规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 43782-2024　人工智能 机器学习系统技术要求

GB/T 41867-2022　信息技术 人工智能 术语

ISO/IEC DIS 22989　Information technology — Artificial intelligence — Artificial intelligence concepts and terminology

T/CCF 0001-2025　强化学习系统 第1部分 通用要求

3　术语和定义

下列术语和定义适用于本文件。

3.1

模型 model

基于数据生成推理或预测的数学结构。

[来源：GB/T 41867-2022, 3.2.11]

3.2

机器学习 machine learning

按照给定标准，利用数据建立模型参数的一类算法。

示例：主成分分析（Principle Component Analysis，PCA）、支持向量机（Support Vector Machine，SVM）。

[来源：GB/T 41867-2022, 3.2.10]

3.3

环境 environment

基于自身状态以及动作输入，输出观测和奖励信息的问题模型，包括状态观测机制、状态转移机制、奖励机制。其中状态观测机制决定状态和观测的形式与二者间关系，状态转移机制决定输入动作导致环境状态变化的方式，奖励机制决定环境对动作的数值评价。

3.4

强化学习 reinforcement learning

一种通过与环境交互收集数据来优化决策的机器学习范式。

[来源：GB/T 41867-2022，3.2.25]

3.5

样本 sample

同环境交互的对象与环境在交互过程中产生的结构化的数据，可经一定处理，一般包括状态、观测、奖励、动作等信息。

3.6

奖励 reward

奖励是环境输出的标量值，用于评价同环境交互的对象与环境的交互过程。

3.7

状态 state

状态是某一时刻下环境蕴含的全部信息，与输入环境的动作共同决定环境输出的观测、奖励及后续时刻的状态。

3.8

动作 action

动作是智能体在特定状态下可执行的操作，动作作为环境的输入。

3.9

观测 observation

观测是环境的输出，为环境状态信息的一个映射。观测描述了智能体在环境的状态信息中可感知的部分。

3.10

智能体 agent

能够感知环境（输出的观测）并产生动作的实体。

[来源：ISO/IEC DIS 22989，3.1.1]

3.11

轨迹 trajectory

轨迹是智能体与环境在一次完整交互过程中产生的样本（状态、动作、奖励）序列的完整记录，是强化学习数据收集与算法优化的核心对象。

轨迹通常表示为：$\tau=(s_0, a_0, r_0, s_1, a_1, r_1, \cdots, s_T, a_T, r_T)$。

3.12

仿真器 simulator

一种通过模仿目标系统的外在表现和行为来实现其功能的工具或系统。它专注于复现目标系统的实际运行状态，而非抽象模型。

3.13

源数据 source data

源数据是由历史交互轨迹组成的集合，这些数据由某一（或多个）未知的行为策略生成，且在学习过程中不再与环境实时交互获取新数据。源数据也称为离线数据集或静态数据集。

4 符号和缩略语

下列符号和缩略语适用于本文件。

AI：人工智能（Artificial Intelligence）

RL：强化学习（Reinforcement Learning）

DRL：深度强化学习（Deep Reinforcement Learning）

RLHF：人类反馈强化学习（Reinforcement Learning from Human Feedback）

IQL：独立Q-Learning（Independent Q-Learning）

QMIX：Q值混合网络（Q-value Mixing Network）

COMA：反事实多智能体策略梯度（Counterfactual Multi-Agent Policy Gradients）

MADDPG：多智能体深度确定性策略梯度（Multi-Agent Deterministic Policy Gradient）

MAPPO：多智能体近端策略优化算法（Multi-Agent Proximal Policy Optimization）

SIL：自模仿学习（Self Imitation Learning）

BCQ：批数据约束的Q学习算法（Batch-Constrained Q-Learning）

BEAR：减少自举误差累积的算法（Bootstrapping Error Accumulation Reduction）

BRAC：行为规范的演员-评论家算法（Behavior Regularized Actor-Critic）

TD3-BC：基于行为克隆的双延迟深度确定性策略梯度算法（Twin Delayed Deterministic Policy Gradient with Behavior Cloning）

RLAIF：基于AI反馈的强化学习（Reinforcement Learning from AI Feedback）

5 功能架构

强化学习环境是基于输入动作，输出观测、奖励等反馈的功能模块，用于表达强化学习算法所求解的问题场景，包括源数据、仿真器与场景适配等。本文件不对仿真器和源数据的实现进行规范，仅规范场景适配模块。场景适配对源数据、仿真器进行封装，将其特化的接口、协议转换为强化学习环境统一的接口和协议，与强化学习智能体交互，如图 1 所示。

强化学习环境与强化学习智能体在训练、推理等过程中交互操作，包括：

a) 强化学习环境输出观测、奖励信息，可用于强化学习推理、训练、评估等工作流。其中推理依赖观测信息，训练和评估依赖观测和奖励信息。

b) 强化学习环境可输出观测、奖励之外的其他信息供强化学习系统相关组件使用以实现特定功能。其他信息可包括可视化数据、日志数据等，实现的功能包括环境可视化、运行状况监测等。

c) 强化学习环境接收智能体生成的动作，完成状态转移并产生观测和奖励。

d) 强化学习环境可接收配置信息，用于指定自身初始化方式。

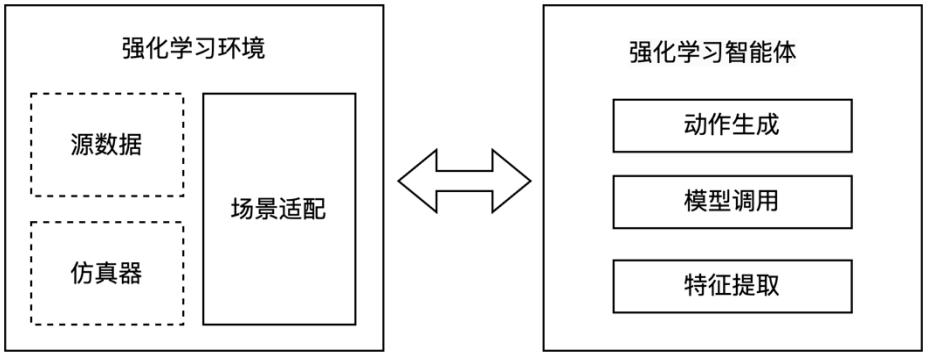

图1：强化学习环境与强化学习智能交互的功能架构

注：强化学习环境在训练和推理工作流中的适配见附录A。不同强化学习范式中的强化学习环境描述见附录B。

6 场景适配

6.1 概述

场景适配模块为仿真器或源数据所表达的问题场景提供统一封装的的接口和协议，方便与强化学习智能体交互。

6.2 仿真器适配

6.2.1 功能描述

仿真器是一种通过模仿目标系统的外在表现和行为来实现其功能的工具或系统，它专注于复现目标系统的实际运行状态，而非抽象模型。仿真器可以是软件程序，也可以是硬件设备。在软件层面，仿真器运行于计算机之上，通过建立被测系统的模型，包括其输入、输出、状态和行为等方面的描述，来模拟系统的运行过程。在硬件层面，仿真器作为专用硬件设备，凭借高速处理器和大容量存储器，实现对目标系统行为的模拟，并且能够达到实时或接近实时的模拟效果。不同仿真器一般在架构、流程、数据协议等方面存在差异，有特有的协议、数据格式、时序模型等。

场景适配模块是介于强化学习智能体与仿真器之间的软件中间件，通过实现标准化的接口、数据协议和交互流程，将仿真器封装为符合强化学习范式的训练环境，消除仿真器差异对强化学习智能体的影响。场景适配将来自强化学习智能体的输入动作封装为符合仿真器定义的合法动作，将仿真器的输出封装成符合强化学习范式的观测和奖励等信息，对于不包含奖励信号的输出数据，需要人为指定奖励信号。对于来自强化学习智能体的非法输入动作，场景适配将不执行该非法动作，并返回相关信息。

6.2.2 仿真器的场景适配接口规范

6.2.2.1 状态重置接口

接口名：reset。

接口功能：调用仿真器的接口重置环境到初始状态，返回初始观测。

接口参数列表 ：

options：dict，可选参数， 默认值为{}，用于传入环境特有的可调参数，例如 seed 参数用于固定环境的初始状态的随机性。

接口返回值：环境观测 observation，见章节 6.2.3 的表 1 观测数据表示。

6.2.2.2 状态转移接口

接口名：step。

接口功能：调用仿真器的接口执行输入动作，返回奖励、新的观测。

接口参数列表：

action：dict，必要参数，用于传入环境所需的智能体动作信息以进行状态转移。

extra_info：dict， 可选参数，默认值为{}，复杂的环境可能需要额外的信息来执行动作，如在模拟机器人运动的环境中，可能需要传递机器人的当前电量、传感器误差等额外信息。

接口返回值：环境观测observation,见章节6.2.3的表2动作数据表示；环境奖励reward，见章节6.2.3的表3奖励数据表示。

6.2.2.3 状态可视化接口

接口名：render（可选）。

接口功能：将环境当前状态，智能体的观测和其他额外信息，可视化展示在监视设备上。

接口参数列表：

options：dict， 可选参数，默认值为{}，对可视化形式进行描述。

接口返回值：None。

注：render接口不参与环境运算或模拟，仅将已有的环境状态信息尽量真实的可视化展示在监视设备上。状态可视化接口可根据不同的在线仿真环境进行不同的实现，如通过render()展示环境运行状态；前期将环境历史状态储存，后期通过render()接口溯源历史环境与智能体轨迹等。

6.2.3 数据格式[1]

6.2.3.1 观测数据

观测数据格式定义见表1。

表 1 观测数据表示

数据	类型	必选或可选	注释
env_id	string	必选	当前环境实例的唯一 id（取值范围 1 到 36 个字符）
frame_no	int32	必选	当前环境实例运行时的帧号（取值范围 -2^{31} 到 $2^{31}-1$）
observation	Structure	必选	当前帧的观测信息
extra_info	Structure	可选	当前帧的可选额外信息
terminated	int32	必选	当前环境实例是否结束（取值为 0 或 1）
truncated	int32	必选	当前环境实例是否异常或中断（取值为 0 或 1）

1) [1] 附录 C 提供了面向围棋场景的强化学习环境示例，包括观测、动作、奖励等典型数据描述。

其中，observation（观测）为环境在交互中对智能体可见的信息，区别于环境内部状态信息。观测描述了智能体能够从环境中感知到的信息，通常为环境状态的一部分或对环境状态的一个映射下的像。

extra_info（额外信息）为环境对训练框架和工作流提供的额外信息，作为提供给智能体的观测信息的补充。在支持多智能体的强化学习环境中，额外信息应包含整体观测信息到每个智能体各自观测信息的映射，以支持环境对多个智能体提供不同观测信息。额外信息还可以包含环境内部状态信息等智能体观测不到的信息，以辅助“中心化训练、去中心化执行（CTDE）”等训练工作流的需求，但此类信息在推理时应当无法被智能体获取。额外信息还可以包含智能体动作执行信息，若智能体执行非法动作，将包含错误提示信息。

6.2.3.2 动作数据

动作数据格式定义见表2。

表 2 动作数据表示

数据	类型	必选或可选	注释
env_id	string	必选	当前环境实例的唯一 id（取值范围 1 到 36 个字符）
frame_no	int32	必选	当前环境实例运行时的帧号（取值范围-2^{31}到$2^{31}-1$）
action	Structure	必选	当前帧智能体所执行的动作

其中，action（动作）是强化学习智能体在特定状态下可执行的操作，动作作为环境的输入，既可以是单个智能体的单个动作，也可以是多个智能体的联合动作。

6.2.3.3 奖励数据

奖励数据格式定义见表3。

表 3 奖励数据表示

数据	类型	必选或可选	注释
env_id	string	必选	当前环境实例的唯一 id（取值范围 1 到 36 个字符）
frame_no	int32	必选	当前环境实例运行时的帧号（取值范围-2^{31}到$2^{31}-1$）
reward	Structure	必选	当前帧环境给智能体的奖励

其中，reward（奖励）是环境当前帧对智能体给出的及时奖励，通常为标量。在支持多智能体的强化学习环境中，奖励应以字典的形式包含每个智能体到其奖励的映射，即环境可能对每个决策的智能体分别提供不同的奖励信号。

6.2.3.4 场景可视化数据

场景可视化数据格式定义见表4。

表 4 场景可视化数据表示

数据	类型	必选或可选	注释
env_id	string	必选	当前环境实例的唯一 id（取值范围 1 到 36 个字符）
frame_no	int32	必选	当前环境实例运行时的帧号（取值范围-2^{31}到$2^{31}-1$）
frame_data	Structure	必选	当前帧的可视化数据

6.3 源数据适配

6.3.1 功能描述

源数据是由历史交互轨迹组成的集合，这些数据由某一（或多个）行为策略生成，可表现为结构化形式（如表格、数据库）或非结构化形式（如文本、图像、音频、视频），用于分析、强化学习模型训练或其他数据处理任务。

面向源数据的场景适配模块通过实现标准化的接口、数据协议和交互流程，将源数据处理成符合强化学习范式的轨迹，处理过程可包括为源数据补充缺失的奖励信号。

6.3.2 源数据的场景适配接口规范

6.3.2.1 初始化接口

接口名：init。

接口功能：设置数据集到初始状态或特定状态。

接口参数列表：

options：dict, 可选参数， 默认值为{}，用于传入源数据特有的可调参数，比如用于固定环境的初始状态随机性的随机种子等。

接口返回值：无。

6.3.2.2 单步数据采样接口

接口名：sample_steps。

接口功能： 返回批量的单步转移训练数据和额外信息。

接口参数列表：

batch_size：int，必选参数，批训练数据的数量。

extra_info：dict, 可选参数，默认值为{}，可提供按优先级采样、按累计回报阈值采样等额外信息。

接口返回值：批量单步转移数据,见章节6.3.3.1表5单步数据表示。

6.3.2.3 轨迹数据采样接口

接口名：sample_trajectories。
接口功能：返回批量的完整轨迹和其他额外信息。
接口参数列表：
trajectory_size：本次采样的轨迹数。
extra_info：dict，可选参数，默认值为{}，可提供按优先级采样、按累计回报阈值采样等额外信息。
接口返回值：批量轨迹数据,见章节6.3.3.2表6轨迹数据表示。

6.3.2.4 动作接口

接口名：get_all_actions。
接口功能：返回数据集中当前状态下所有出现过的动作（用于约束策略）。
接口参数列：
extra_info：dict，可选参数，默认值为{}，用于传入源数据特有的可调参数。
接口返回值： 动作数据,见章节6.2.3.2表2动作数据表示。

6.3.2.5 统计接口

接口名：statistics。
接口功能： 返回数据集统计信息（如状态、动作分布）。
接口参数列表：
extra_info：dict，可选参数，默认值为{}，用于传入统计方式等参数。
接口返回值： 轨迹数据,见章节6.3.3.2表6轨迹数据表示；或者是统计的标量数据。

6.3.3 数据格式

6.3.3.1 单步转移数据

单步转移数据格式定义见表5。

表 5 单步转移数据表示

数据	数据类型	必选或可选	注释
env_id	string	必选	当前环境实例的唯一 id（取值范围 1 到 36 个字符）
frame_no	int32	必选	当前环境实例运行时的帧号（取值范围-2^{31}到$2^{31}-1$）
observation	Structure	必选	当前帧的观测信息，类型参考 6.2.3.1
action	Structure	必选	当前帧的动作，类型参考 6.2.3.2
reward	Structure	必选	当前帧的奖励，类型参考 6.2.3.3
next_observation	Structure	可选	下一帧的观测信息，类型参考 6.2.3.1

done	int32	可选	终止标志，表示轨迹是否结束（取值为 0 或 1）

6.3.3.2 轨迹数据

轨迹数据格式定义见表6。

表 6 轨迹数据表示

数据	数据类型	必选或可选	注释
env_id	string	必选	当前环境实例的唯一 id（取值范围 1 到 36 个字符）
trajectory_id	int32	必选	当前轨迹的唯一 id（取值范围-2^{31}到$2^{31}-1$）
steps_set	Set(Structure)	必选	当前轨迹的单步转移集合

6.4 其他要求

为保证强化学习智能体开发者正确使用环境，场景适配的其他要求如下：

a) 应提供场景适配后的接口调用方法说明。

b) 应提供场景适配后的协议说明，包括输入输出数据类型、取值范围、含义描述等。

c) 宜支持基于场景适配协议的测试用例自动化生成。

附 录 A
（资料性）
强化学习环境在训练和推理工作流的适配

A.1 强化学习环境在训练工作流的适配

强化学习环境提供统一的与强化学习智能体交互的接口和协议，为强化学习训练工作流提供了标准的接入和使用方式。依据标准实现的强化学习环境作为软件实体，接口和数据协议满足6.2.3、6.3.3节规范，在强化学习训练工作流中，强化学习智能体调用环境接口可以与环境进行交互，发送并且接收协议化数据，完成适配。在基于仿真器的强化学习环境下（图2），强化学习智能体通过调用强化学习环境接口将动作发送到环境推动环境执行并接收观测和奖励；在基于源数据的强化学习环境下（图3），强化学习智能体通过调用强化学习环境接口获得环境的轨迹数据。

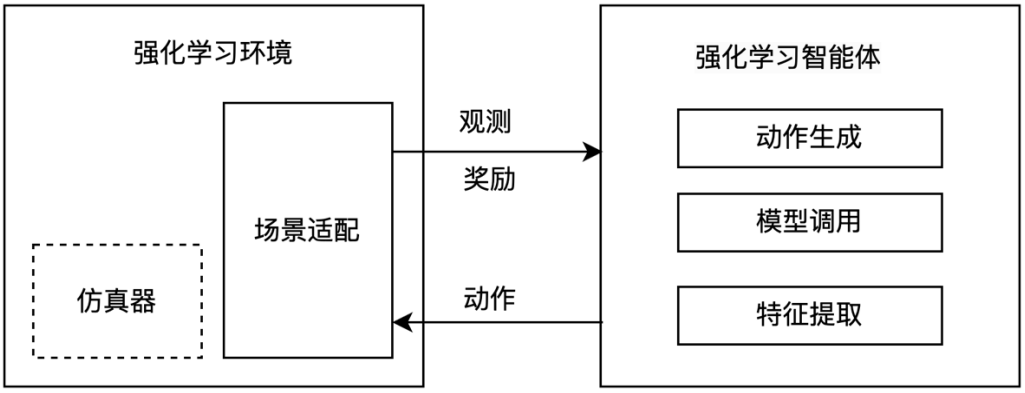

图2：训练工作流中面向仿真器的强化学习环境适配

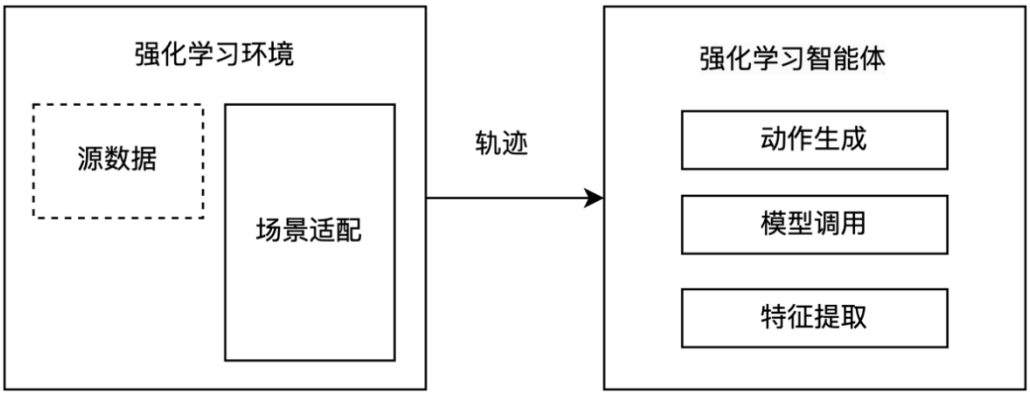

图3：训练工作流中面向源数据的强化学习环境适配

A.2 强化学习环境在推理工作流的适配

对于基于仿真器的强化学习环境，其推理工作流的适配可参考A.1中仿真器环境的适配方式；对于基于源数据的强化学习环境,其推理工作流既可直接使用源数据(参考A.1 中源数据环境的适配方式)，也可在需要时与仿真器交互（参考A.1中仿真器环境的适配方式）。

附 录 B
（资料性）
常见范式中的强化学习环境

B.1 在线强化学习

B.1.1 构建在线强化学习的环境

在线强化学习一般拥有可供智能体交互的仿真环境或真实环境，按照仿真器适配方式即可完成强化学习环境的适配工作。

B.2 离线强化学习

B.2.1 概述

离线强化学习是利用已有的交互记录——即源数据——的一种强化学习范式。

常见的离线强化学习算法包括:BCQ(Batch-Constrained Q-Learning)、BEAR(Bootstrapping Error Accumulation Reduction）、BRAC（Behavior Regularized Actor-Critic）、TD3-BC（Twin Delayed Deterministic Policy Gradient with Behavior Cloning）等。

B.2.2 构建离线强化学习环境

离线强化学习求解的问题场景表达形式为源数据，场景适配可按照源数据适配规范实现标准化的数据获取方式，满足不同类型的离线强化学习算法。

B.2.2.1 自模仿学习

自模仿学习的训练一般使用的是经过筛选的高回报轨迹数据（一般包含状态、动作、奖励、终止标志、下一帧状态等）。

自模仿学习可通过轨迹数据采样接口实现按轨迹采样训练数据用于模仿高回报序列；可通过单步数据采样接口实现按单步采样训练数据用于策略梯度更新；可通过单步数据采样接口或轨迹数据采样接口按优先级采样基于回报或优势值赋予采样权重高的样本，优先学习高价值样本。其中，轨迹数据采样接口应具备筛选高回报轨迹，提供每条轨迹的累计回报值，或支持按回报阈值过滤的功能。若数据集中包含专家演示，轨迹数据采样接口可以采样专家与非专家轨迹。自模仿学习可通过数据统计接口实现轨迹统计、优势值统计、离线策略评估等等。

B.2.2.2 免模型的离线强化学习

免模型的离线强化学习的训练一般使用有起始和终止索引的轨迹数据，支持按轨迹采样（例如筛选高回报轨迹）。

免模型的离线强化学习可通过轨迹数据采样接口实现按轨迹采样或实现按批次采样，用于模仿高回报序列，还可以通过累计回报值、每条轨迹的总奖励值，进行高回报轨迹的筛选；可通过单步数据采样接口实现随机采样单步转移用于策略梯度更新，也可实现按照优先级采样、基于优势值或 TD-error 权重高的样本，优先学习高价值样本。

B.2.2.3 基于模型的离线强化学习

基于模型的离线强化学习的训练一般使用有起始和终止索引的轨迹数据，支持按轨迹采样，用于多步规划或模型预测。轨迹可能包含行为策略生成动作的概率分布，用于重要性采样或约束策略分布偏移。轨迹应有数据来源标签如专家数据、随机策略或混合策略等，便于模型区分数据分布。

可通过轨迹数据采样接口获取完整轨迹，用于训练动态模型的多步预测能力。可通过单步数据采样接口实现随机采样单步转移数据，用于训练环境模型或策略优化；也可通过单步数据采样接口筛选特定状态-动作对的转移数据，用于分析局部动态特性。可通过数据统计接口获得离线模型评估数据计算动态模型在验证集上的预测误差或者在动态模型生成的仿真环境中运行策略，评估长期回报。

B.3 面向多智能体的强化学习

B.3.1 概述

多智能体强化学习研究环境中包含多个智能体的决策学习。环境中各个智能体目标可能有所不同，观测可能存在差异，动作可能相互影响。此外，各智能体之间可能存在合作、竞争等关系，并且这些关系可能随着环境状态的转移而发生变化。

常见的多智能体强化学习算法包括IQL（Independent Q-Learning）、QMIX（Q-value Mixing Network）、COMA（Counterfactual Multi-Agent Policy Gradients）、MADDPG（Multi-Agent Deterministic Policy Gradient）、MAPPO（Multi-Agent Proximal Policy Optimization）等。

B.3.2 构建多智能体强化学习环境

场景适配可将多智能体问题场景封装为统一接口协议的强化学习环境。

局部观测支持：在环境观测信息中extra_info字段里提供各智能体的局部观测信息与observation的映射关系，各智能体根据该映射关系获得自己的局部观测。

同时决策支持：环境提供的状态转移接口与单智能体情况相同，既可每次仅接收单个智能体的动作输入，输出观测、奖励等反馈，由场景适配转为多个智能体逐次决策；也可每次接收多个智能体的联合动作输入，同时输出多个观测、奖励等反馈。

智能体间通信支持：信息可作为信源智能体输出动作的一部分提供给强化学习环境，由场景适配转发给目标智能体，作为信宿智能体观测的一部分体现在观测数据的extra_info字段。

B.4 面向大语言模型的强化学习

B.4.1 概述

大语言模型中的强化学习主要包括人类反馈的强化学习（Reinforcement Learning from Human Feedback, RLHF），以及及其变种如AI反馈的强化学习（Reinforcement Learning from AI Feedback，RLAIF）等。RLHF通过人类反馈对大语言模型进行微调，使其输出更符合用户的期望。

目前，主流的RLHF实现方式包括以下步骤：

收集人类反馈：大量收集人类对问答的偏好数据。这些数据通常包括人类对不同输入-输出对的评分或排序。

训练奖励模型：利用收集到的问答-偏好数据集训练奖励模型。该模型能够根据问答的质量给出相应的奖励分数。

强化学习调优：使用奖励模型对大语言模型的问答进行强化学习调优。具体来说，通过PPO等强化学习算法，优化语言模型的参数，使其生成的回答能够获得更高的奖励分数。

通过这种方式，RLHF能够显著提升大语言模型的输出质量，使其更贴近用户的期望。

B.4.2 构建面向大语言模型的强化学习环境

收集人类反馈与训练奖励模型不属于强化学习，但训练得到的奖励模型可使用适配器封装为强化学习环境，用于后续强化学习调优过程。

人类反馈的强化学习的轨迹数据一般包含输入文本如用户指令或问题、模型输出如LLM生成的响应文本、奖励信号如基于人类反馈的数值化评分（偏好对比得分或人工打分）、终止标志如标记对话是否结束（达到最大生成长度或用户终止对话）。可通过采样接口实现数据获取，包括文本交互轨迹数据、人类偏好数据（如成对偏好标签）、 多维度评估标签、行为策略元数据、动作概率分布等。轨迹采样接口可实现按轨迹采样，返回完整对话轨迹（如多轮问答），用于训练长序列依赖模型；可实现按偏好对采样，返回成对响应及人类偏好标签，用于训练奖励模型；可实现按优先级采样，优先采样高奖励或高争议性样本，提升训练效率。

B.5 面向数学、物理、化学、生物等基础科学的强化学习

B.5.1 概述

随着强化学习不断普及，数学、物理、化学、生物等基础科学研究者开始逐步应用强化学习技术辅助求解本领域内问题。强化学习在自动定理证明、核聚变等离子体控制、分子合成路线设计、蛋白质结构设计等基础科学问题上都展现出了广阔的应用前景。

B.5.2 构建面向基础学科的强化学习环境

为基础科学领域的问题搭建强化学习环境时，通常需要调用领域特定的软件工具，因此场景适配应当为常用的此类软件工具提供必要的支持。该软件工具一般满足仿真器的特性，是一种可以复现基础学科各种功能的工具，它通过对该学科系统的输入和输出进行模拟，实现了类似该系统的运行效果，如果存在此类仿真器，应使用仿真器适配实现强化学习环境的构建。如果存在该领域特定的数据集合，应使用源数据适配实现强化学习环境构建。

B.6 面向具身智能的强化学习

B.6.1 概述

具身智能是指物理实体（如机器人、机动车、无人机等）通过和物理环境交互来表现和发展的智能。在具身智能的研究中，强化学习是一种重要的方法。

具身智能中应用强化学习的方式主要有模拟到现实迁移强化学习、真机强化学习。其中，模拟到现实迁移强化学习是在模拟环境中运行强化学习训练；而真机强化学习则是在物理环境中运行强化学习训练。两种方法所获智能体都将作为软件组件用于控制物理实体同物理环境交互。

B.6.2 构建面向具身智能的强化学习环境

基于模拟环境的具身智能强化学习：按照仿真器适配要求对模拟器进行环境适配。

基于真机历史数据的强化学习：按照源数据适配要求对历史数据进行环境适配，构建离线强化学习环境，根据不同的应用场景和算法，选择不同的构建方式（自模仿学习、基于模型的强化学习或免模型的强化学习）以满足要求。

基于实体真机的强化学习：按照仿真器适配要求对真机环境进行适配，其中状态重置接口需向人类或其他辅助工具发送信号来完成；状态转移接口需访问机器人传感器获得观测数据；状态可视化接口记录真机的行动数据以便进行可视化展示。

附 录 C
（资料性）
面向围棋的强化学习环境示例

面向围棋的观测数据示例见表7。

表 7 面向围棋的观测数据

数据	数据类型	必选或可选	注释
env_id	string	必选	当前环境实例的唯一 id
frame_no	int32	必选	当前手数

observation	{
	int32[19][19],
	int32;
	}
		必选	● int32[19][19]：表示棋盘每个格子的状态，值为-1 表示黑，值为 0 表示空，值为 1 表示白
			● int32：下一手棋子颜色，值为-1 表示黑，值为 1 表示白

extra_info	-	可选	无
terminated	int32	必选	胜负，值为 1 表示对局正常结束
truncated	int32	必选	终止，值为 1 表示对局异常中止

面向围棋的动作数据示例见表8。

表 8 面向围棋的动作数据

数据	类型	必选或可选	注释

env_id	string	必选	当前环境实例的唯一 id
frame_no	int32	必选	当前手数

action	{
	int32
	}
		必选	int32: 取值范围[0, 19*19+1), 含义是落子位置或停一手

面向围棋的奖励数据示例见表9。

表 8 面向围棋的奖励数据

数据	类型	必选或可选	注释
env_id	string	必选	当前环境实例的唯一 id
frame_no	int32	必选	当前手数

reward	{
	int32
	}
		必选	int32:
			● 棋局为未结束时, 值为 0
			● 棋局分出胜负时, 值为-1 表示失败, 值为 1 表示胜利

参 考 文 献

[1] Levine, Sergey, et al. "Offline reinforcement learning: Tutorial, review, and perspectives on open problems." arXiv preprint arXiv:2005.01643 (2020).

[2] Oh, J., Guo, Y., Singh, S. and Lee, H., 2018. Self-Imitation Learning. Proceedings of the 35th International Conference on Machine Learning (ICML), 80, pp.3878－3887.

[3] Levine, S., Kumar, A., Tucker, G. and Fu, J., 2020.Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. arXiv preprint arXiv:2005.01643.

[4] Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C. and Ma, T., 2021.MOPO: Model-based Offline Policy Optimization. Advances in Neural Information Processing Systems (NeurIPS), 34, pp.14129－14142.

[5] Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S. and Amodei, D., 2017. Deep Reinforcement Learning from Human Preferences. Advances in Neural Information Processing Systems (NeurIPS), 30, pp.4299－4307.

[6] Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J. and Quillen, D., 2018. Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection. The International Journal of Robotics Research (IJRR), 37(4-5), pp.421－436.