ME5424 课程项目报告：老鹰-母鸡链条对抗的分阶段强化学习

项目参与人：项目组

指导教师：——

提交单位：——

提交日期：2025 年 11 月 20 日

目录

摘要

本报告围绕多智能体对抗强化学习在复杂物理场景的冷启动与收敛难题，提出基于课程学习的分阶段训练方案：阶段一训练母鸡对抗启发式老鹰，阶段二冻结母鸡策略训练老鹰。在 Box2D 物理仿真与 Gymnasium 接口下，采用 Stable-Baselines3 的 PPO 算法，并行采样与 TensorBoard 日志支持，实现可复现的工程流水线。主要工作涵盖物理环境建模（世界、链条与关节、边界与反弹）、观测与奖励设计、并行训练与评估回调、以及行为可视化与指标分析。关键成果显示阶段一母鸡防守能力提升（平均回报上升、解释方差接近 1、熵损失下降），阶段二老鹰逼近与抓捕策略有效（潜在奖励稳定为正、抓捕事件带来稀疏正回报），整体训练曲线平稳。项目局限在于奖励权重与物理参数仍需经验调参。后续将考虑联合微调、自博弈与更复杂场景扩展。

关键词：多智能体对抗；课程学习；PPO；Gymnasium；Box2D

1 引言

本项目聚焦在真实物理约束下的多智能体对抗强化学习，构建“母鸡护卫小鸡链条、老鹰进攻链尾”的二维仿真任务，以检验在稀疏目标与强耦合动力学条件下的稳定学习。训练初期策略弱与回报稀疏使对抗难以形成，链条由多个个体通过距离关节相连，动作与惯性沿链传递产生滞后与振荡，边界约束与反弹进一步加剧优化难度，直接联合训练易震荡或早收敛。为此我们采用分阶段课程学习：第一阶段固定老鹰为启发式追尾，仅训练母鸡习得“挡线与护卫”；第二阶段冻结母鸡策略，训练老鹰在非平凡对手下学会“逼近与抓捕”。该方案在 Gymnasium 接口与 Box2D 物理仿真框架下实现，并配合视角特定观测归一化、潜在奖励塑形及统一的边界与反弹建模，降低分布不稳定与探索失败带来的训练困难。

角色设定方面，老鹰每步根据场内未被捕获的小鸡位置与护卫、队友的观测组成局部状态，奖励强调快速抓捕、避免越界与重复围捕，并考虑被母鸡击退的惩罚；母鸡围绕安全半径持续构建观测，优先护住半径内的小鸡，并通过挡在老鹰与小鸡连线之间获取更高奖励，一旦有小鸡被捕会整体受罚以驱动贴身防御；小鸡持续感知最近的母鸡与老鹰，策略倾向靠近保护者、远离威胁，奖励鼓励保持在护卫范围内并惩罚被捕，促成协同躲避行为。该三方设定使任务目标、物理约束与可解释奖励形成一致闭环，便于后续分析与评估。核心目标在于通过物理真实的链条动力学与可解释奖励，验证分阶段训练在稀疏奖励的多智能体任务上的有效性，并形成一套可复现、可评估且具教学价值的工程实践流程。行为层面，母鸡在链条惯性作用下保持稳健居中、防守链尾并在老鹰与尾端连线之间形成有效“挡线”，老鹰学会从侧后绕行与加速逼近；指标层面，阶段一平均回报稳步提升、解释方差接近 1、熵损失逐步下降，阶段二抓捕成功率提升、逼近潜在奖励稳定为正且总体损失曲线平缓。我们同时强调训练过程的重复性与可验证性，通过固定随机种子、规范评估回调与可视化呈现，使结果具备教学与研究双重价值。

2 研究现状分析

多智能体强化学习在对抗任务中常采用联合训练或自博弈策略，但在奖励稀疏、物理耦合复杂的场景下容易出现训练震荡、梯度估计方差大、策略互相适应导致分布漂移问题。主流方案包括：

- 使用 PPO、SAC 等 on-policy/off-policy 算法配合并行采样与熵正则来缓解探索不足
- 采用 Curriculum Learning 逐步提升任务难度
- 引入潜在奖励、shaping 或模仿学习稳定训练

现有不足：

- 联合训练早期两个策略均弱导致对抗无效，产生近零回报和无意义梯度
- 单视角观测未能在自身体坐标系下稳定归一化，导致感知不一致
- 物理碰撞与边界约束未被合理编码到奖励，易出现“边界卡死”等策略

本项目的必要性与创新点：

- 视角特定观测，将观测统一到各自坐标与速度归一化以降低分布不稳定
- 分阶段冻结策略，将对手策略内联推理到环境，避免早期无效对抗
- 引入几何挡线得分与物理反弹机制，提升可解释性与物理一致性
- 并行采样与张量日志配合高批量训练，提升稳定性与吞吐

3 文献综述

3.1 研究现状与挑战

多智能体强化学习在对抗任务中常采用联合训练或自博弈策略，但在奖励稀疏、物理耦合复杂的场景下容易出现训练震荡、梯度估计方差大、策略互相适应导致分布漂移问题。主流方案包括：

- 使用 PPO、SAC 等 on-policy/off-policy 算法配合并行采样与熵正则来缓解探索不足
- 采用 Curriculum Learning 逐步提升任务难度
- 引入潜在奖励、shaping 或模仿学习稳定训练

现有不足:

- 联合训练早期两个策略均弱导致对抗无效，产生近零回报和无意义梯度
- 单视角观测未能在自身体坐标系下稳定归一化，导致感知不一致
- 物理碰撞与边界约束未被合理编码到奖励，易出现“边界卡死”等策略

3.2 多智能体强化学习与自博弈

多智能体对抗任务广泛存在于机器人协作与博弈环境中。自博弈（Self-Play）与对手建模（Opponent Modeling）是提升策略鲁棒性的常见路径，但在早期阶段经常遭遇稀疏奖励与不稳定分布的问题。策略间相互适应会导致训练数据分布漂移，加剧优化难度。

在系统性综述中，研究者总结了多智能体深度强化学习在通信、协作与对抗中的挑战与进展，指出分布非平稳、信用分配与探索不足是主要瓶颈，需通过稳定的算法与合理的任务设计加以缓解 [1]。在通信方面，基于端到端学习的可微通信机制可在部分可观测条件下提升协作与对抗效果，但也引入了新的优化不稳定性，需配合正则与课程策略进行控制 [2]。

3.3 近端策略优化（PPO）

PPO 通过截断的目标函数限制策略更新幅度，结合优势函数估计与熵正则，兼顾稳定性与样本效率，已成为连续动作场景的主流基线。在本项目中，我们采用 SB3 实现，并行环境与高批量配置以降低时间相关性，配合评估回调与 TensorBoard 曲线监控收敛行为。

与其他连续控制算法相比，DDPG 以确定性策略与目标网络稳定训练，适合低维连续动作但对探索与超参数较敏感 [3]；SAC 基于最大熵原理鼓励更均衡的探索，在复杂任务上常表现出更好的稳健性 [4]。PPO 以其更新简单、实现成熟、对超参数不敏感的优势在诸多工程场景成为首选，并有可靠的开源实现支撑复现实验 [5]。

3.4 课程学习（Curriculum Learning）

课程学习通过逐步提升任务难度缓解冷启动与探索问题。在本项目中，首先在阶段一固定老鹰为启发式策略，让母鸡在可控对手下学习“挡线与防守”，随后在阶段二冻结母鸡策略，使老鹰在非平凡对手下学习“逼近与抓捕”。这一方案避免了联合训练早期对抗无效导致的近零梯度。

从经典视角看，课程学习通过样本或目标的序列化组织，使学习过程由易到难，从而提升泛化与稳定性 [6]。在对抗性的多智能体场景中，分阶段或半静态训练能够在不

改变最终任务的前提下，缓解双边同时学习带来的震荡与劣适应问题，相关工作在群体对抗中也展现了有效性 [7]。

3.5 奖励塑形与潜在函数

奖励塑形通过引入潜在函数或设计几何得分，使稀疏目标转化为可优化的密集信号。本项目采用：(1) 几何挡线得分，衡量母鸡是否有效位于老鹰与链尾之间；(2) 逼近潜在奖励，衡量老鹰与链尾的距离缩短；(3) 拉伸、边界与反弹项，约束物理合理性与鼓励探索活跃度。

在最大熵框架下，通过熵正则鼓励多样化策略，可在局部最优附近维持足够探索，结合潜在奖励能进一步提升训练的稳定性与样本效率 [4]。在多智能体设置中，奖励通常同时依赖个体与队伍整体状态，因此塑形需与物理约束一致，以避免产生不合理的策略偏好 [1]。

3.6 物理仿真与可解释性

基于 Box2D 的距离关节与刚体碰撞为链条与主体提供可解释的动力学。通过统一边界约束与反弹规则，学习到的策略在可视化层面符合直觉（如母鸡张开“翅膀”形成阻挡带、老鹰绕后与加速逼近），增强了实验结果的物理一致性与教学可读性。

在控制任务的标准化评估方面，DeepMind Control Suite 提供了可复现实验的任务集合与评测指标，为连续控制算法的比较与复现实验奠定了基础 [8]。本项目遵循相同的可复现原则：固定随机种子、规范日志与评估回调、统一物理参数与接口定义，以提升结果的可信度与可重复性 [5]。

3.7 并行采样与稳定训练

并行环境（`SubprocVecEnv`）打破单环境时间相关性，提升采样吞吐与数据独立性；高批量配置与固定随机种子提高训练稳定性与可重复性。评估回调按固定频率保存最优模型，便于阶段性复盘与结果展示。

工程实践表明，稳定的实现、明确的日志与评测协议、以及与社区基线的一致性对复现实验至关重要 [5]。我们在实现与评估上与主流基线保持一致的接口与设置，以尽可能减少实现差异带来的偏差。

4 项目设计与实现

4.1 需求分析

系统需支持在二维世界边界内的稳定对抗：

- 母鸡防守链尾并保持链条运动稳定
- 老鹰在有限步数内高效逼近与抓捕
- 观测需在各自坐标系下归一化
- 奖励需兼顾目标达成与物理合理性
- 训练需具备并行采样与可重复性，提供评估与可视化

4.2 方案设计（技术路线与选择依据）

算法选择为 Stable-Baselines3 的 PPO，考虑其在连续动作空间的鲁棒性与实现成熟度。物理仿真采用 Box2D，通过距离关节模拟链条的弹性与阻尼。环境设计为共享物理核心 `BasePhysicsEnv`，在其上派生阶段一母鸡训练环境与阶段二老鹰训练环境。训练采用并行子进程环境与 TensorBoard 日志，保证采样独立性与可观测性。

4.3 实现过程（流程与关键步骤）

阶段一

- 初始化世界与链条
- 老鹰使用启发式追尾策略
- 母鸡以 PPO 学习防守与位置控制
- 评估回调保存最优模型

阶段二

- 加载并冻结母鸡策略
- 老鹰以 PPO 学习逼近与抓捕
- 同样使用评估回调保存最优模型
- 提供可视化脚本用于行为检查

4.4 关键问题与解决方法

针对“开局贴脸”与“边界卡死”，通过扩大初始距离、增加边界惩罚与统一反弹机制，并设计退避—绕圈状态机缓解；对奖励塑形进行消融与灵敏度分析，平衡挡线权重、潜在逼近系数与活跃度项，提升收敛稳定性与行为合理性。对象为两个策略网络（母鸡与老鹰），变量包括物理参数（世界大小、速度上限、链条长度等）与 PPO 超参数（学习

率、批大小、折扣因子等)。控制条件包括固定仿真步长与最大步数、固定随机种子以可重复复现实验。可重复性设计：所有参数均在源码中明确，训练脚本输出模型与事件日志，图像与曲线来自同一目录，确保复现路径可见。

4.5 物理环境实现

物理参数由统一配置管理,包含 world_size、dt、max_steps、chain_links、chain_spacing、hen_max_speed、eagle_max_speed、catch_radius、block_margin 等。世界构建、链条与关节由专用构建函数实现，并在动作应用时进行力作用与速度裁剪以保持仿真稳定。

4.6 物理参数与配置

物理参数	数值
世界半边长 world_size	20.0
步长 dt	$\frac{1}{30}$ 秒
单回合最大步数 max_steps	600
链条节数 chain_links	7
链条间距 chain_spacing	1.0
母鸡最大速率 hen_max_speed	9.0
老鹰最大速率 eagle_max_speed	36.0
捕获半径 catch_radius	0.7
阻挡宽度 block_margin	2.0

表 1: 物理环境关键参数（项目设计与实现）

4.7 训练超参数与配置

4.8 链条动力学与小鸡队伍协同

小鸡通过距离关节串联形成柔性链条，关节的目标长度、频率与阻尼决定了链条的刚柔程度与振荡幅度。链条结构对队伍协同行为有直接影响:

- **耦合与滞后**：任意一节小鸡的加速或转向会沿关节传递到邻近节段，产生时间上的滞后与相位差，要求母鸡在防守时保持平稳引导，避免过度拉伸与摆动。
- **形态与队形**：链条在运动中自然形成“头部—中段—尾部”的速度与惯性梯度，尾端最易被老鹰逼近，因此防守策略需优先保护尾端并在老鹰与尾端连线之间构建“挡线”。

训练超参数	数值
并行环境 `n_envs`	16
每环境步数 `n_steps`	256
批大小 `batch_size`	512
学习率 `learning_rate`	3×10^{-4}
折扣因子 `gamma`	0.995
总步数 `total_steps`	1,000,000
设备 `device`	auto/cpu

表 2: PPO 训练关键超参数（项目设计与实现）

- **约束与协同**：链条的长度与张力对队伍整体机动性构成约束，小鸡个体策略倾向于靠近母鸡的安全半径并保持与队友的距离，从而降低拉伸惩罚并提高整体稳定性。
- **奖励耦合**：拉伸、边界与反弹等物理项会同时影响多个节段的奖励，使个体行为与队伍表现耦合在一起，推动形成协同性的躲避与护卫行为。

4.9 奖励与碰撞机制

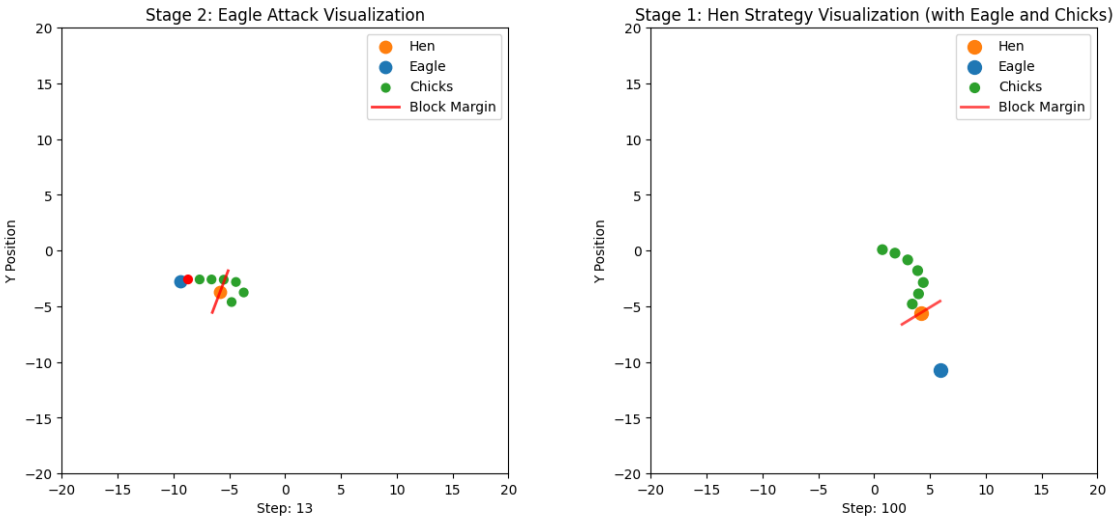
母鸡阶段奖励由几何挡线得分、避捕得分、链条拉伸与尾距惩罚、边界惩罚与生存奖励组成。挡线得分计算如下：设老鹰为 E、尾端为 T、母鸡为 H，当母鸡投影落在 $\overline{ET}$ 上且垂距不超过臂展 a 时，得分为 $s = \max(0, 1 - d/a)$，其中 d 为垂距。老鹰阶段奖励由抓捕稀疏奖励、逼近潜在奖励、链条拉伸奖励、反弹与边界惩罚、绕后与速度奖励、距离压力与活跃度、时间饥饿项构成。物理反弹在两阶段统一处理，包含翅膀带与实体碰撞的速度反射与位移弹开。

4.10 训练与评估

阶段一训练采用并行环境配置（如 $n\_envs = 16$、每环境步数 $n\_steps = 256$、批大小 512、学习率 3×10^{-4}、折扣因子 $\gamma = 0.995$）。阶段二在冻结母鸡策略的条件下进行老鹰训练，参数设置与阶段一一致。评估回定期保存最优模型至指定输出目录。

4.11 可视化窗口

策略在物理环境中的行为通过可视化脚本以 Matplotlib 渲染。角色观测与奖励在引言中已概述。



(a) 阶段一 gui (b) 阶段二行 gui

图 1: 可视化窗口示例

4.12 阶段性成果

阶段性成果包括:

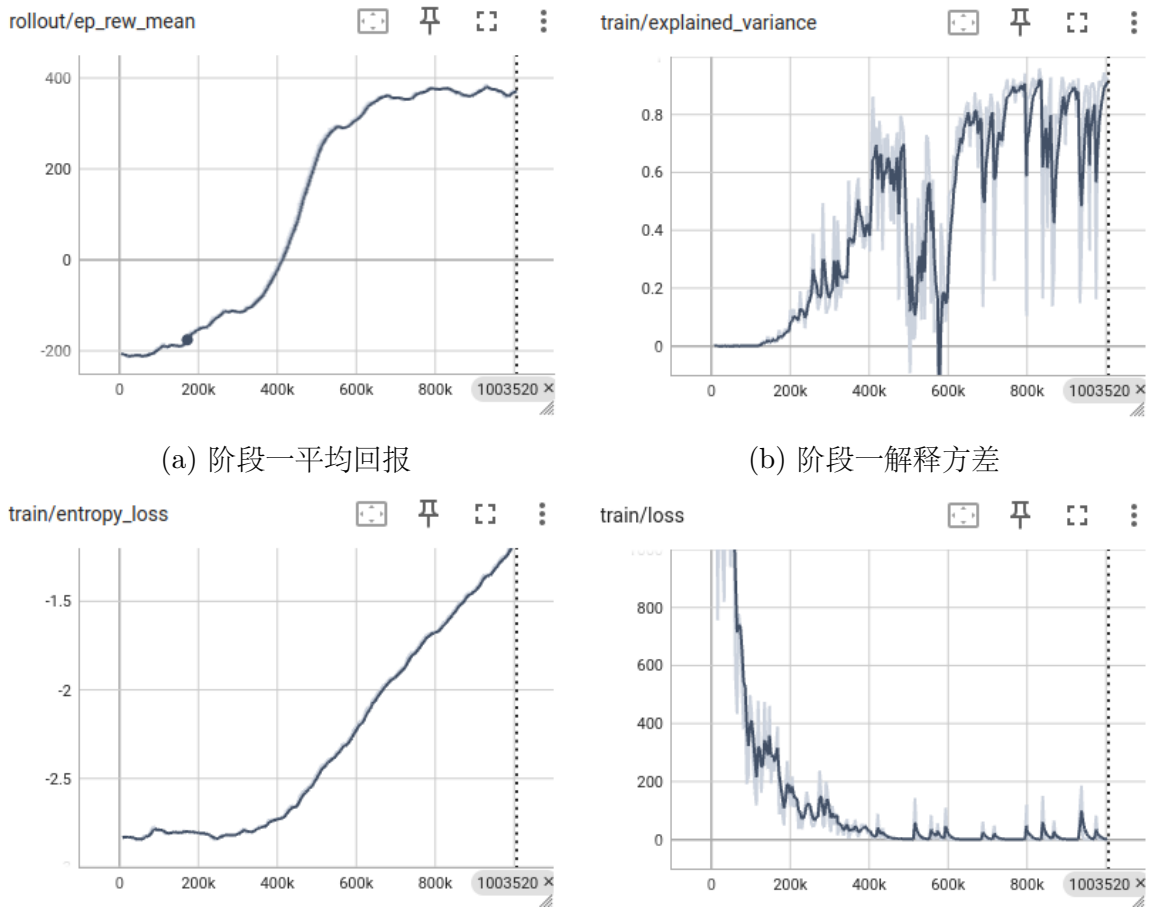
- 阶段一最优母鸡策略文件 `hen_stage_1.zip` 与评估最优 `best_hen/best_model.zip`
- 阶段二最优老鹰策略 `eagle_stage_1.zip`
- TensorBoard 日志与训练曲线
- 针对“开局贴脸”与“边界卡死”，我们扩大初始距离、增加边界惩罚并引入反弹与状态机退避绕圈，显著改善训练稳定性

5 成果与分析

5.1 成果呈现

5.2 数据采集与预处理

数据由训练过程自动记录到 TensorBoard 事件文件，包含每回合平均回报、损失与熵、解释方差等指标。图像与可视化快照用于直观展示训练趋势与行为。



(a) 阶段一平均回报

(b) 阶段一解释方差

(c) 阶段一熵损失

(d) 阶段一总体损失

图 2: 阶段一训练指标曲线

5.3 训练曲线与行为可视化

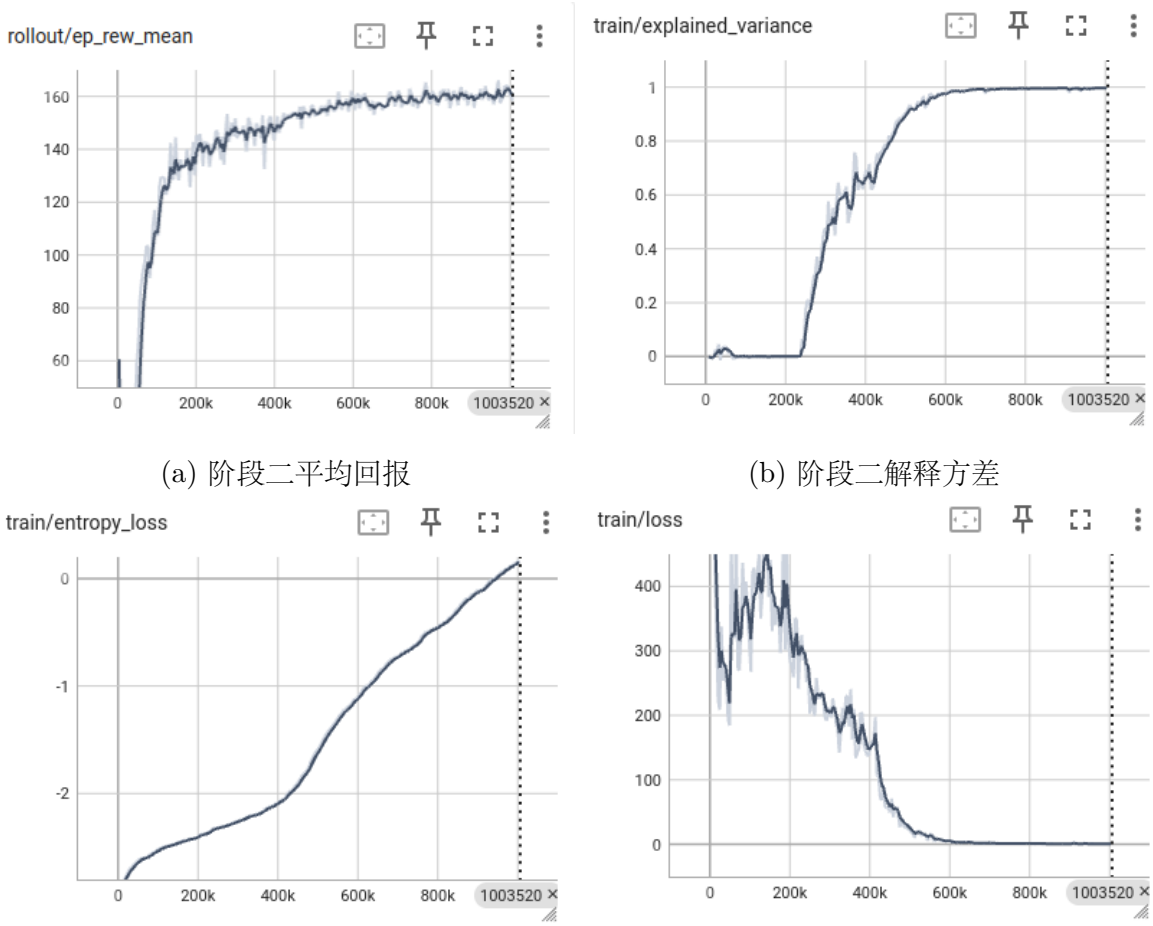
5.4 成果分析

阶段一中，平均回报随训练迭代稳步提升，解释方差接近 1，熵损失逐步降低，显示策略从探索向确定性过渡且价值函数拟合良好。阶段二中，潜在逼近奖励推动老鹰持续缩短与链尾距离，抓捕事件带来稀疏正回报，整体损失曲线平稳。行为可视化显示母鸡在链条惯性作用下保持居中防守并通过翅膀阻挡老鹰，老鹰在冻结母鸡下学会从侧后绕行与加速逼近。与预期目标对比，两个阶段目标均达成。

6 结论与展望

6.1 结论

总结：本项目以课程学习拆解多智能体对抗训练的冷启动难题，通过视角规范化、物理反弹与几何得分设计、并行采样与评估回调实现稳定训练与可视化验证，输出可重



(a) 阶段二平均回报

(b) 阶段二解释方差

(c) 阶段二熵损失

(d) 阶段二总体损失

图 3: 阶段二训练指标曲线

复的实验管线与模型成果。创新点包括冻结对手内联推理、挡线几何评分与统一反弹机制。局限性包括奖励设计仍需人工调参、策略在极端参数下可能过拟合物理边界。展望：引入联合微调阶段与自博弈，使用更丰富的潜在奖励或逆强化学习提高泛化；扩展到更复杂的障碍物与地图；将评估指标系统化并加入统计显著性检验。

7 附录：运行与复现实验

依赖安装使用 `pip` 与 `conda`，训练脚本分别为 `python src/train_hen.py` 与 `python src/train_eagle.py`，可视化脚本为 `python src/visualize_hen_stage1.py` 与 `python src/visualize_eagle_stage2.py`。日志查看使用 `tensorboard --logdir results/curriculum`。相关源码位置在文中已注明，以确保复现路径可见与实验可重复。

代码入口与结构：物理环境与课程学习逻辑位于 `src/curriculum_env.py`；训练脚本为 `src/train_hen.py` 与 `src/train_eagle.py`；可视化工具见 `src/visualize_hen_stage1.py` 与 `src/visualize_eagle_stage2.py`。

参考文献

[1] Thanh Thi Nguyen, Ngoc Duy Nguyen, and Saeid Nahavandi. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE Transactions on Cybernetics*, 50(9):3826–3839, 2020.

[2] Jakob Foerster, Yannis M Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2137–2145, 2016.

[3] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2016.

[4] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, pages 1861–1870. PMLR, 2018.

[5] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 23(268):1–8, 2022.

[6] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 41–48, 2009.

[7] Hongliang Cai, Yugang Luo, Hongli Gao, et al. A multiphase semistatic training method for swarm confrontation using multiagent deep reinforcement learning. *Computational Intelligence and Neuroscience*, 2023:2955442, 2023.

[8] Yuval Tassa, Yotam Doron, Alistair Muldal, Joel Z Leibo, Tom Erez, and Yuval Li. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

[9] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.