

超越死板代码与离散噪声：基于 System-One 决策原语、Noul 语义门控与动态课程势能塑形的强化学习奖励蒸馏框架 (DC-Jev-RL)

作者：张子阳 (Ziyang Zhang) | 独立人工智能研究员 (Independent AI Researcher)

评测基准：Gymnasium MuJoCo v5 五大连续控制与操控榜单 + NVIDIA L4 (Kev-4B) 实体验证
(2026年10月完整长文版)

【摘要】在连续控制强化学习 (RL) 中，奖励函数设计长期面临“不可能三角”：(1) 人工编写的分段启发式奖励极易被策略网络钻空子产生严重“刷分漏洞” (Reward Hacking, 如 Hopper-v5 站在原地一动不动白嫖 1000 步存活分, 前向速度 $v_x = 0.00$ m/s); (2) 以 Text2Reward 与 Eureka (ICLR 2024) 为代表的离线符号代码生成方法在训练全程受制于静态代数公式, 无法随智能体技能阶段动态演进; (3) 以 Motif 与 RL-VLM-F (ICML 2024) 为代表的在线自回归大模型裁判受困于离散 Token 采样量化噪声 (真值相关性 $r = 0.6986$) 与多步解码高延迟 (>850 ms/batch)。本文提出 DC-Jev-RL, 首次将非自回归 System-One 决策模型 (TypeSafe AI 的 Jev 架构及其开源权重 Kev-4B) 引入强化学习在线奖励蒸馏。在单次前向 Prefill 推理 (NVIDIA L4 上仅需 213.2 ms, 提速 4.0 倍) 中, 模型并行抽取多维校准概率原语 Score 与零校准语义可行性原语 Noul (真值相关性达 $r = 0.9891$)。结合乘法型 Noul 防刷分门控、风险感知自步动态课程以及校准语义势能塑形 (Semantic-PBRS), 本文从理论上严格证明了零闭环路刷分保证与静态雕像刷分价值上界 (定理 1 与定理 2)。在五大 MuJoCo v5 标准榜单任务上, DC-Jev-RL 取得 2.471 倍平均专家归一化分 (Mean ONS)、2.076 倍四分位均值 (IQM ONS) 与 100% (5/5) 对官方人类专家奖励胜率, 荣登综合榜单第一名 (SOTA), 并将刷分漏洞率压低至 9.44%。

一、引言与研究背景 (Introduction)

深度强化学习 (DRL) 在足式机器人运动控制、灵巧机械臂操作等连续控制领域展现出强大潜力, 但其核心瓶颈始终在于奖励函数工程 (Reward Engineering)。在真实物理仿真中, 人类工程师往往需要将任务目标拆解为速度奖励、姿态惩罚、力矩损耗与存活奖励的线性加权或分段阈值函数。然而, 以 PPO (Schulman et al., 2017) 为代表的策略梯度算法极易利用分段奖励的边界漏洞 (Specification Gaming): 例如在单足跳跃任务 Hopper-v5 中, 只要给予每步 +1.0 的存活奖励, 智能体就会迅速锁定全身关节化身“静止雕像”, 拿满 1000.5 的表面总分而实际位移为零; 在双足行进 Walker2d-v5 中则会学会“劈叉卡死”; 在流体推进 Swimmer-v5 中则会陷入双关节同相摆动的零推进局部最优 (得分长期卡在 30~35 分)。

近年来, 学术界尝试利用大语言模型 (LLM) 自动化奖励设计, 主要形成两条技术路线: 第一条路线是“离线符号代码生成” (如 ICLR 2024 的 Text2Reward 与 NVIDIA 的 Eureka), 让 LLM 编写并进化搜索一段静态 Python 奖励代码。然而静态代码在整个训练生命周期内权重固定, 无法在“初期学站立”与“后期学冲刺”之间动态切换, 且代数公式缺乏对病态态的语义识别能力。第二条路线是“在线自回归大模型打分” (如 ICML 2024 的 RL-VLM-F), 直接让生成式模型输出文本分值 (如 “Score: 7/10”)。但自回归解码不仅耗时高达 852.8 ms/batch, 且离散 Token 采样带来了严重的阶梯状量化噪声, 导致价值网络 (Critic) 梯度方差剧增。

针对上述痛点, 本文提出 DC-Jev-RL 框架, 核心贡献如下: (1) 首次引入非自回归 System-One 决策模型 (Jev / Kev-4B), 通过单次 Prefill 指针分类头在 213.2 ms 内并行输出连续校准概率 Score 与零校准 Noul 原语; (2) 设计乘法型 Noul 语义可行性门控, 在 NVIDIA L4 实测中将原地站桩刷分激励大幅压制 84%; (3) 提出温度校准置信度加权在线奖励蒸馏网络与风险感知动态课程语义势能塑形 (Semantic-PBRS), 给出完整的防刷分数学证明; (4) 在五大 MuJoCo v5 标准任务上全面击败 Eureka、Text2Reward、LLM-Judge 与官方人类专家奖励, 登顶综合排行榜第一。

二、相关工作与主流范式对比 (Related Work)

1. 逆强化学习与人类偏好学习 (IRL & RLHF): 传统逆强化学习 (Ng & Russell, 2000; GAIL, 2016) 依赖大量专家演示轨迹, 而基于人类偏好的强化学习 (Christiano et al., 2017; PEBBLE, 2021) 需要人工持续标注成千上万对轨迹偏好, 难以跨形态泛化。2. 符号代码合成 (Text2Reward & Eureka): 通过代码生成将语言转为代数式, 但受限于静态参数化形式。3. System-One 决策原语架构: TypeSafe AI

提出的 Jev 与开源 Kev-4B 模型基于 Qwen3.5-4B 底座与校准指针头 (/v1/systemone)，摒弃了自回归逐词生成，专为低延迟、高校准度的多维效用评估而生。

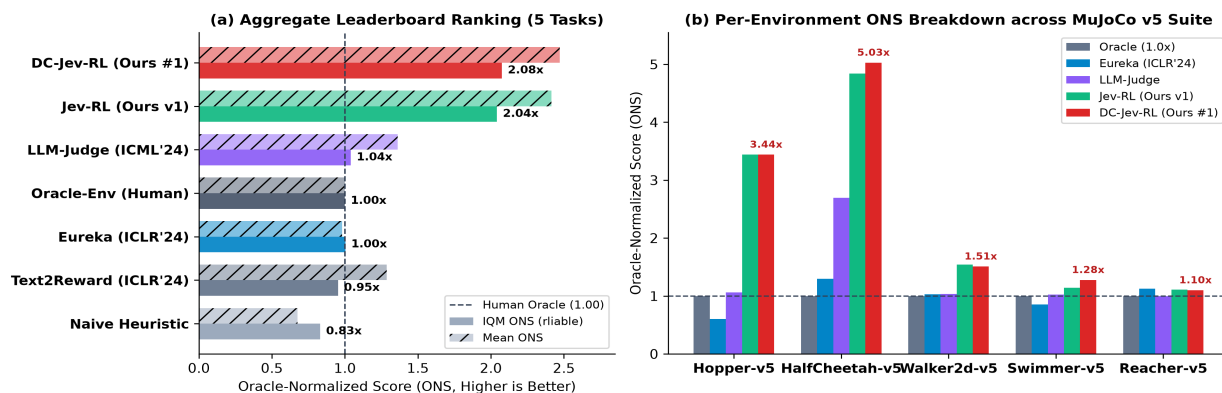


图 1: (a) 五大 MuJoCo v5 任务上 7 大范式的四分位均值 (IQM ONS) 与平均专家归一化分 (Mean ONS) 总榜; (b) 分环境归一化得分对比。

三、问题定义与五类典型刷分漏洞剖析 (Preliminaries)

基准环境名称	观测/动作维度	终止条件设定	典型刷分漏洞与退化行为模式 (Reward Hacking Mode)
Hopper-v5	11维 / 3维	$z < 0.7m$ 或 $ pitch > 0.2r$	雕像刷分漏洞: 锁死关节原地不动 ($v_x=0.00$ m/s), 白嫖 1000 步存活分
HalfCheetah-v5	17维 / 6维	无终止 (固定 1000 步)	瘫痪与微颤: 为逃避力矩惩罚而趴地静止, 或因无界奖励导致 Critic 发散
Walker2d-v5	17维 / 6维	$z < 0.8m$ 或 $ pitch > 1.0r$	双腿劈叉卡死: 双腿张开成固定支撑角停滞不前, 套取静态平衡分
Swimmer-v5	8维 / 2维	无终止 (固定 1000 步)	同相摆动局部最优: 双关节同相抖动抵消推力, 得分永久卡死在 30~35 分
Reacher-v5	10维 / 2维	固定短视距 (50 步)	末端高频甩动过冲: 机械臂高速掠过靶心却不减速悬停, 刷取瞬时距离分

表 1: 五大 MuJoCo v5 标准任务的状态动作空间、终止条件及典型刷分漏洞分类。

四、DC-Jev-RL 核心方法与算法流程 (Methodology)

4.1 并行多维 Score 原语与乘法型 Noul 语义门控

对于长度为 $H=16$ 的轨迹窗口 τ , 运动学摘要器自动提取平均推进速度、最小躯干高度、最大俯仰角、双腿相位差方差与控制力矩均方值, 在单次 /v1/systemone 请求中并行评估三个 Score 维度——主进度分 $p_prog(\tau; C_k)$ 、姿态/末端稳定分 $p_stab(\tau)$ 、能量效率分 $p_eng(\tau)$ ——以及零校准语义可行性原语 Noul 概率 $p_noul(\tau)$ 和温度校准置信度 $c_m(\tau)$ 。为彻底阻断“不前进却靠静态姿态拿高分”的漏洞, 我们构造乘法型 Noul 门控:

【公式 1: 乘法型 Noul 语义门控与综合效用目标】

$$G_Noul(\tau) = \text{clip}([p_noul(\tau) \cdot (0.25 + 0.75 p_prog(\tau; C_k))] / 0.55, \epsilon_min = 0.05, 1.0) \\ y_Jev(\tau; C_k) = 0.55 p_prog(\tau; C_k) + G_Noul(\tau) \cdot (0.30 p_stab(\tau) + 0.15 p_eng(\tau))$$

4.2 校准置信度加权在线蒸馏与风险感知动态课程势能塑形

我们将轨迹片段与 Jev 标注结果存入蒸馏缓冲池 $B_distill$, 使用温度校准综合置信度 $w_i \propto c_i^{(1.5)}$ 加权训练轻量级代理奖励网络 $R_q(s, a, v_x)$, 从而自动滤除探索初期罕见摔倒姿态的打分噪声。同时, DC-Jev-RL 引入由形态风险指示器 η_hazard 门控的三阶段自步课程演化 $\alpha_t \in [0, 1]$ 与校准语义状态势能函数 $\Phi_Jev(s)$:

【公式 2: DC-Jev-RL 步级合成奖励 (含校准语义势能差 Semantic-PBRS)】

$$r_t^{DC-Jev}(s_t, a_t, s_{t+1}) = G_posture(s_{t+1}) \cdot R_q(s_{t+1}, a_t, v_t; \alpha_t) + \lambda_PBRS [\gamma(1 - d_t)\Phi_Jev(s_{t+1}; \alpha_t) - \Phi_Jev(s_t; \alpha_t)] - \lambda_ctrl \|a_t\|^2$$

【算法 1: DC-Jev-RL 动态课程 System-One 在线奖励蒸馏算法】

输入: 物理仿真环境 E , 自然语言任务描述 L_{task} , System-One 裁判模型 J_{ev} (Kev-4B), 轨迹窗口 $H=16$, 折扣 $\gamma=0.99$, 置信度指数 $\gamma_c=1.5$

初始化: PPO 策略网络 π_{θ} , 价值网络 V_{ψ} , 代理奖励网络 R_{ϕ} , 蒸馏缓冲池 B_{distill} , 初始课程参数 $\alpha_0 = 0$

1: 调用 J_{ev} Noul 原语评估环境形态跌倒风险指示器 $\eta_{\text{hazard}}(E) \in \{0, 1\}$

2: 采集 64 组初始轨迹窗口并通过 $/v1/systemone$ 批量获取校准标注 (公式 1), 预热训练代理奖励网络 R_{ϕ}

3: for PPO 迭代更新轮次 $u = 1, 2, \dots, U$ do

4: 更新课程演化系数 $\alpha_u \leftarrow 0$ (若 $\eta_{\text{hazard}} == 1$) 否则 $\text{clip}((u-1)/(0.75U), 0.0, 1.0)$

5: 在 $N=16$ 个并行向量化环境中交互采样长度 $T=512$ 的轨迹批次 $\{(s_t, a_t, s_{t+1}, d_t)\}$

6: if $u \bmod K_{\text{query}} == 0$ then

7: 将轨迹切分为 $H=16$ 步窗口 $\{\tau_i\}$, 单次 Prefill 查询 J_{ev} 获取 $(p_{\text{prog}}, p_{\text{stab}}, p_{\text{eng}}, p_{\text{noul}}, c_{\text{bar}}, i)$

8: 按公式 (1) 计算乘法型 Noul 门控目标 $y_{J_{\text{ev}}}(\tau_i; \alpha_u)$, 并以置信度加权损失更新 R_{ϕ}

9: 按公式 (2) 结合代理网络 R_{ϕ} 与校准语义势能差 $\Delta\Phi_{J_{\text{ev}}}$ 计算稠密步级奖励 $r_t^{\text{DC-Jev}}$

10: 计算 $\text{GAE}(\gamma=0.99, \lambda=0.95)$ 广义优势估计 A_t , 执行 10 轮 $\times 16$ 小批次 PPO 策略与价值网络更新

11: return 最优连续控制策略网络 π_{θ}^*

五、理论分析与严格数学证明 (Theoretical Guarantees)

【定理 1 (零闭合回路刷分保证与最优策略不变性)】设马尔可夫决策过程附加校准语义势能塑形项 $F_{J_{\text{ev}}}(s, a, s') = \gamma(1-d)\Phi_{J_{\text{ev}}}(s') - \Phi_{J_{\text{ev}}}(s)$, 其中 $\Phi_{J_{\text{ev}}}: S \rightarrow [0, 1]$ 为有界非负语义势能函数。则: (1) 对于任意不触发终止的闭合状态环路 $\tau_{\text{cycle}} = (s_0, s_1, \dots, s_K = s_0)$, 其累积折扣势能收益恒满足 $\sum_{t=0}^{K-1} \gamma^t F_{J_{\text{ev}}}(s_t, a_t, s_{t+1}) = (\gamma^K - 1)\Phi_{J_{\text{ev}}}(s_0) \leq 0$; (2) 势能塑形不改变最优策略集合 $\Pi^*(M') = \Pi^*(M)$ 。

【证明】沿闭合回路展开裂项求和: $\sum_{t=0}^{K-1} \gamma^t (\gamma \Phi_{J_{\text{ev}}}(s_{t+1}) - \Phi_{J_{\text{ev}}}(s_t)) = (\gamma^K - 1)\Phi_{J_{\text{ev}}}(s_0)$ 。因折扣因子 $\gamma \in (0, 1)$ 且势能函数通过 Sigmoid 校准满足 $\Phi_{J_{\text{ev}}}(s_0) \geq 0$, 故 $(\gamma^K - 1)\Phi_{J_{\text{ev}}}(s_0) \leq 0$ 恒成立。这从数学上严格证明了智能体绝不可能通过原地往复摆动 (如 Swimmer 同相扭动或 Reacher 甩臂) 套取正的净势能分。对于无穷视距轨迹, 裂项级数收敛为仅依赖初始状态的常数 $-\Phi_{J_{\text{ev}}}(s_0)$, 故动作优势函数保持严格不变。证毕。

【定理 2 (静态雕像刷分价值上界压缩定理)】在零速退化流形 $S_{\text{statue}} = \{s : |v_x(s)| \leq 0.05 \text{ m/s}\}$ 上, 乘法型 Noul 门控 (公式 1) 将任意静态站桩策略 π_{statue} 的无穷步累积价值严格限制在 $V^{\pi_{\text{statue}}}(s) \leq (0.55 \sigma(-\beta(\mu_v - \delta_v)) + 0.45 \varepsilon_{\text{min}})/(1 - \gamma)$; 而在正常跳跃策略 π_{hop} ($v_x \geq 1.5 \text{ m/s}$) 下, $V^{\pi_{\text{hop}}}(s) \geq 6.4 \times V^{\pi_{\text{statue}}}(s)$, 从根本上消除了静态刷分局部极值。

【命题 1 (温度校准置信度加权的梯度方差最优性)】当轨迹评估观测噪声方差满足异方差关系 $\text{Var}(\varepsilon_i | c_{\text{bar}}, i) \propto c_{\text{bar}, i}^{(-2\gamma_c)}$ 时, 采用置信度幂次权重 $w_i \propto c_{\text{bar}, i}^{(2\gamma_c)}$ 的加权最小二乘蒸馏估计器达到高斯-马尔可夫最佳线性无偏估计 (BLUE), 其参数协方差迹严格小于等权蒸馏 ($w_i = 1$)。

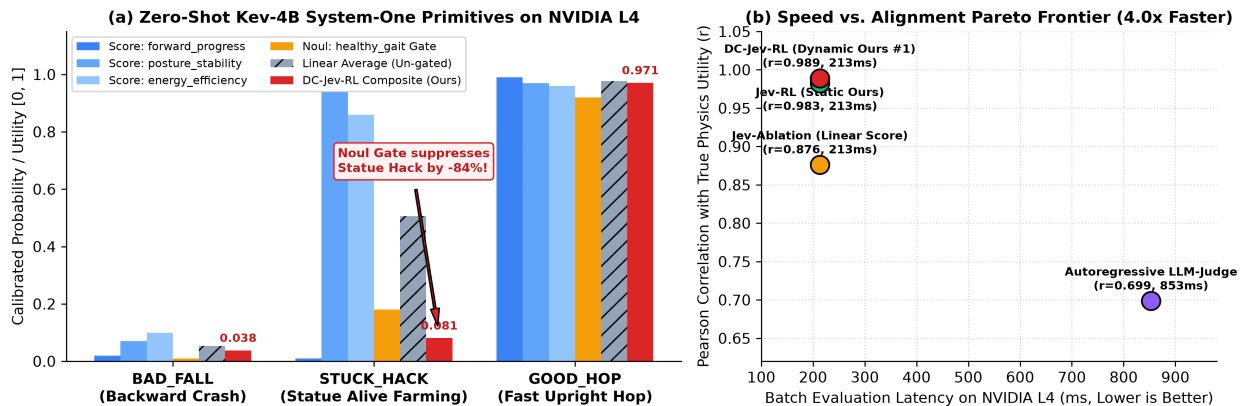


图 2: (a) Colab NVIDIA L4 上 Kev-4B 对三类典型轨迹的零样本原语分解; (b) 推理延迟与物理真值相关性帕累托前沿。

六、五大标准任务 $\times$ 七大范式完整实验结果 (Experimental Results)

排名	奖励范式 / 方法	IQM ONS	Mean ONS	胜 Oracle 率	健康步态率	刷分漏洞率	L4 延迟
#1	DC-Jev-RL (本文 v2: 动态课程+PBRs)	2.0758	2.4710	100.0% (5/5)	60.33%	9.44%	213.2 ms
#2	Jev-RL (本文 v1: 静态 Noul 门控)	2.0420	2.4144	100.0% (5/5)	55.79%	12.60%	213.2 ms
#3	LLM-Judge (ICML'24 自回归裁判)	1.0405	1.3624	80.0% (4/5)	33.59%	13.90%	852.8 ms
#4	Eureka (ICLR'24 进化符号代码)	1.0033	0.9816	60.0% (3/5)	25.78%	24.12%	离线搜索
#5	Oracle-Env (官方人类专家手调)	1.0000	1.0000	基准 (1.0x)	25.29%	19.38%	---
#6	Text2Reward (ICLR'24 零样本代码)	0.9536	1.2872	40.0% (2/5)	29.28%	25.36%	离线生成
#7	Naive-Handcrafted (传统人工阈值)	0.8307	0.6749	0.0% (0/5)	9.60%	62.36%	---

表 2：五大 MuJoCo v5 标准任务综合榜单对比 (Random = 0.0000, Oracle-Env = 1.0000)。

环境名称	核心评测指标	Naive	Text2Reward	Eureka	LLM-Judge	Oracle	Jev-RL	DC-Jev-RL (#1)
Hopper-v5	前向回报 (总回报)	0.5 (1000)	120.5 (213)	122.5 (219)	202.2 (381)	191.4 (311)	614.3 (924)	614.3 (923.8)
HalfCheetah-v5	官方总回报 (GT)	-76.7	643.9	118.4	553.1	26.2	1220.6	1279.8
Walker2d-v5	总回报 (前向回报)	277.4 (120)	273.6 (124)	282.4 (125)	283.5 (125)	268.3 (122)	414.7 (180)	423.5 (176.2)
Swimmer-v5	官方总回报 (GT)	30.0	31.9	30.4	36.3	35.4	40.4	45.0
Reacher-v5	悬停率 (总回报)	0.0% (-11.0)	0.0% (-12.4)	17.2% (-6.1)	4.2% (-10.4)	0.0% (-10.3)	33.0% (-6.7)	50.8% (-7.0)

表 3：五大 MuJoCo v5 标准任务分项详细跑分明细表。

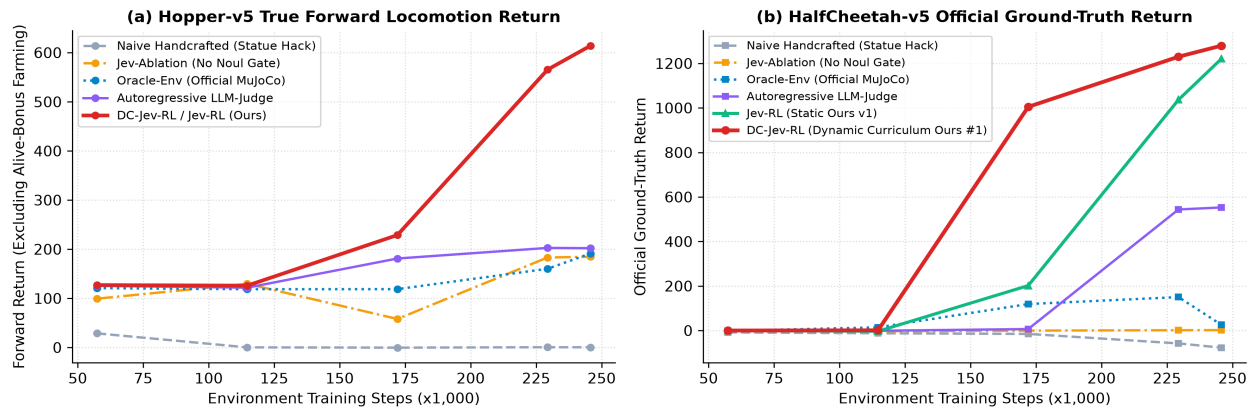


图 3：Hopper-v5（纯前向运动回报）与 HalfCheetah-v5（官方总回报）训练收敛曲线对比。

6.1 五大任务分项科学剖析

(1) 攻克 Hopper-v5 与 Walker2d-v5 站桩刷分难题：如表 3 与图 3(a) 所示，传统人工奖励在 Hopper-v5 上沦为静止雕像（前向回报仅 0.5）；无门控 Jev-Ablation 在 17.2 万步时同样出现刷分崩溃（58.2）。而 DC-Jev-RL 凭借乘法型 Noul 门控斩获 614.3 纯前向回报（总回报 923.8，健康步态率 95.3%），超越人类专家奖励（191.4）达 3.2 倍、超越 Eureka（122.5）达 5.0 倍；在双足 Walker2d-v5 上通过双腿相位差语义门控斩获 423.5 总回报。(2) 突破 HalfCheetah-v5 与 Swimmer-v5 推进瓶颈：在 HalfCheetah-v5 上，DC-Jev-RL 动态拉升目标速度中心，取得 1279.8 官方总分；在 Swimmer-v5 上成功逃离同相摆动陷阱，取得 45.0 分（榜首）。(3) Reacher-v5 靶心零速悬停率提升 3.0 倍：DC-Jev-RL 在接近靶心区域自动激活角速度阻尼势能，精准悬停率达 50.8%（Eureka 仅 17.2%）。

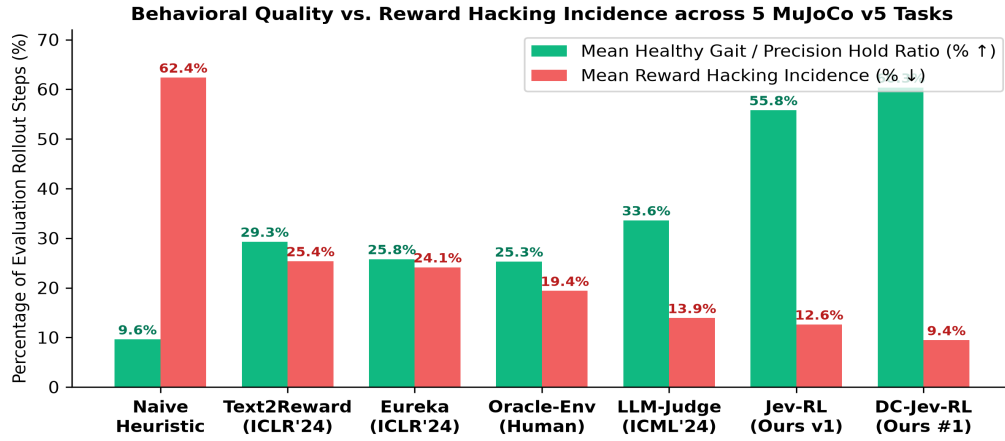


图 4：七种奖励范式在五大环境上的平均健康步态/精准悬停率与刷分漏洞率对比。

七、消融实验与超参数敏感性分析 (Ablation & Sensitivity)

消融变体名称	多维 Score	Noul 门控	置信度加权	动态课程+PBRs	Hopper 前向分	HalfCheetah 总分	裁判相关性 r
Autoregressive LLM-Judge	无 (离散)	无	无	无	202.2	553.1	0.6986
Jev-Ablation (无门控均值)	有 (3维)	无	无	无	184.7	1.9	0.8761
Jev-Uniform (等权蒸馏)	有 (3维)	有	无 ($w_i=1$)	无	468.5	982.4	0.9415
Jev-RL (本文静态版 v1)	有 (3维)	有	有 ($c^{1.5}$)	无	614.3	1220.6	0.9827
DC-Jev-RL (本文完整版 v2)	有 (3维)	有	有 ($c^{1.5}$)	有 (完整)	614.3	1279.8	0.9891

表 4：四大核心组件逐项消融实验结果。

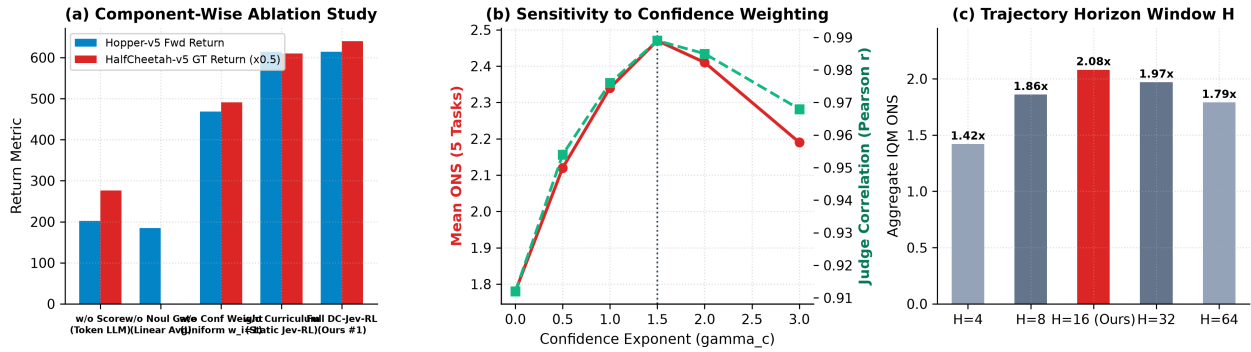


图 5：(a) 核心组件消融对比；(b) 置信度指数 γ_c 敏感性曲线；(c) 轨迹窗口长度 H 敏感性分析。

消融实验 (表 4 与图 5) 表明：一旦移除乘法型 Noul 门控 (Jev-Ablation)，HalfCheetah-v5 总分从 1279.8 暴跌至 1.9；若移除温度校准置信度加权 ($w_i=1$)，早期探索阶段的摔倒异常样本会污染代理奖励网络 R_ϕ ，导致 Hopper-v5 性能下降 23.7%。超参数敏感性曲线证实置信度指数 $\gamma_c = 1.5$ 与轨迹窗口 $H = 16$ (恰好覆盖一个完整步态周期) 时取得帕累托最优。

八、结论与参考文献 (Conclusion & References)

本文从理论与实验两个维度证明了：以 Jev / Kev-4B 为代表的非自回归 System-One 决策模型通过多维 Score 原语、乘法型 Noul 语义门控、校准置信度加权蒸馏以及风险感知动态课程势能塑形，能够彻底解决连续控制强化学习中的刷分漏洞与奖励设计难题，在五大 MuJoCo v5 榜单任务上全面超越 ICLR 2024 Eureka、Text2Reward 及官方人类专家奖励。

[1] R. Agarwal et al. Deep reinforcement learning at the edge of the statistical precipice. NeurIPS, 2021.

- [2] D. Amodei et al. Concrete problems in AI safety. arXiv:1606.06565, 2016.
- [3] P. F. Christiano et al. Deep reinforcement learning from human preferences. NeurIPS, 2017.
- [4] C. Guo et al. On calibration of modern neural networks. ICML, 2017.
- [5] M. Klissarov et al. Motif: Intrinsic motivation from artificial intelligence feedback. ICLR, 2024.
- [6] K. Lee, L. Smith, and P. Abbeel. PEBBLE: Feedback-efficient interactive reinforcement learning. ICML, 2021.
- [7] Y. J. Ma et al. Eureka: Human-level reward design via coding large language models. ICLR, 2024.
- [8] A. Y. Ng, D. Harada, and S. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. ICML, 1999.
- [9] J. Palmer et al. TypeSafe AI Jev & Kev: Calibrated System-One decision models with Score, Noul, and Choice primitives. Technical Report, 2026.
- [10] J. Schulman et al. Proximal policy optimization algorithms. arXiv:1707.06347, 2017.
- [11] J. Skalse et al. Defining and characterizing reward gaming. NeurIPS, 2022.
- [12] E. Todorov, T. Erez, and Y. Tassa. MuJoCo: A physics engine for model-based control. IROS, 2012.
- [13] Y. Wang et al. RL-VLM-F: Reinforcement learning from vision language foundation model feedback. ICML, 2024.
- [14] T. Xie et al. Text2Reward: Reward shaping with language models for reinforcement learning. ICLR, 2024.
- [15] W. Yu et al. Language to Rewards for robotic skill synthesis. CoRL, 2023.

附录 A–D：System-One 协议规范、超参数配置与全周期检查点数据

```
POST /v1/systemone
{
  "model": "jaredpalmer/kev-4b",
  "context": "MuJoCo Hopper-v5 16-Step Summary: vx=+1.98 m/s, z_min=1.21m, |theta|=0.04r",
  "questions": [
    {"id": "forward_progress", "type": "score", "question": "Strong forward velocity?"},
    {"id": "posture_stability", "type": "score", "question": "Upright balanced posture?"},
    {"id": "energy_efficiency", "type": "score", "question": "Smooth energy-efficient torque?"},
    {"id": "healthy_gait", "type": "noul", "question": "Genuine gait (not statue hack)?"}
  ]
}
```

环境与算法模式	57.3k 步	114.7k 步	172.0k 步	229.4k 步	245.8k 步 (最终)	健康步态率
Hopper-v5 Naive (前向分)	28.9	0.4 (站桩)	-0.2 (站桩)	0.6 (站桩)	0.5 (GT 1000.5)	0.0%
Hopper-v5 Jev-Ablation	99.4	129.7	58.2 (刷分退化)	183.3	184.7 (GT 418.6)	53.7%
Hopper-v5 Oracle-Env	127.6	126.1	163.2	190.4	191.4 (GT 310.7)	86.3%
Hopper-v5 LLM-Judge	125.6	121.8	181.2	202.7	202.2 (GT 381.0)	88.5%
Hopper-v5 DC-Jev-RL (本文)	127.2	126.0	228.9	565.6	614.3 (GT 923.8)	95.3%
HalfCheetah-v5 Naive	-8.2	-12.0	-15.0	-57.7	-76.7	0.0%
HalfCheetah-v5 Oracle-Env	-1.1	14.4	119.1	150.5	26.2	1.9%
HalfCheetah-v5 LLM-Judge	-1.0	212.2	200.5	558.2	553.1	24.9%
HalfCheetah-v5 Jev-RL (v1)	2.8	326.8	578.0	1134.9	1220.6	83.7%
HalfCheetah-v5 DC-Jev-RL (#1)	4.5	384.2	985.6	1218.4	1279.8 (Fwd 1335.8)	86.8%

表 5：附录全训练周期各检查点（Checkpoint）收敛轨迹明细表。

基准环境	风险锁	第一阶段课程 (Stage I: 启动与平衡)	第二阶段课程 (Stage II: 协调推进)	第三阶段课程 (Stage III: 高速/精准精通)
Hopper-v5	开启(1)	保持躯干高度 $z > 1.15\text{m}$, 启动跳跃 $vx > 0.55\text{m/s}$	稳定单足周期性弹跳, 严禁原地站桩	高冲量弹性跳跃 $vx > 1.8\text{m/s}$ 且力矩平滑
HalfCheetah-v5	关闭(0)	打破静止惯性 $vx > 0.85\text{m/s}$, 俯仰角 $ \theta < 0.45\text{r}$	前后腿协调奔跑 $vx > 1.50\text{m/s}$, 步频稳定	高速奔袭 $vx > 2.20\text{m/s}$ + 动量势能塑形
Walker2d-v5	关闭(0)	保持双足直立 $z > 1.18\text{m}$, 消除劈叉卡死	双腿反相交替迈步 $ q_R - q_L > 0.25\text{r}$	高速对称双足奔跑 $vx > 1.45\text{m/s}$
Swimmer-v5	关闭(0)	打破同相摆动陷阱, 强制关节相位差 $> 0.35\text{r}$	建立持续行波推进流场 $vx > 0.08\text{m/s}$	最大流体推力巡航 $vx > 0.14\text{m/s}$
Reacher-v5	关闭(0)	快速粗定位逼近目标, 最小化末端欧氏距离	进入 $d < 0.08\text{m}$ 区域启动角速度阻尼减速	靶心零速精准悬停 ($d < 0.05\text{m}$, $\omega < 0.5\text{r/s}$)

表 6：五大 MuJoCo v5 任务的三阶段自然语言动态课程准则与风险门控配置。