SuS：用于内在探索的策略感知惊奇

Mark Kashirskiy[1,2] Ilya Makarov[1]

[1]国立研究型大学高等经济学院，俄罗斯莫斯科[2]圣彼得堡国立信息技术、机械与光学大学 AI人才中心，俄罗斯圣彼得堡

353056@niuitmo.ru iamakarov@hse.ru

arXiv:2601.10349v1 [cs.LG] 15 Jan 2026

摘要

我们提出了策略感知惊讶度（SuS），这是一种新颖的内在动机框架，将前后预测不匹配用作强化学习探索中的新颖性信号。不同于完全依赖状态预测误差的传统好奇心驱动方法，SuS 引入了两个互补组件：策略稳定性（SS）和策略惊奇度（SuS）。SS 衡量跨时间步行为策略的一致性，而 SuS 则捕捉相对于智能体当前策略表征的意外结果。我们的组合奖励公式通过学习得到的加权系数利用这两种信号。我们使用大语言模型在数学推理任务上评估了 SuS，结果表明，其在准确率和解答多样性方面均取得了显著提升。消融研究证实，移除任一组件都会导致至少 10% 的性能下降，从而验证了我们方法的协同作用。与基线方法相比，SuS 的 Pass@1 提升了 17.4%，Pass@5 提升了 26.4%，同时在整个训练过程中保持了更高的策略多样性。

关键词：内在动机、强化学习、对比学习、探索

代码：https://github.com/mariklolik/ sus

1 引言

探索仍然是强化学习中的一项根本挑战，尤其是在奖励稀疏或延迟的环境中。当智能体必须在已知任务上保持表现的同时发现新颖行为时，探索-利用权衡便至关重要。在复杂领域中，随机探索策略往往不足以应对，因为要取得有意义的进展，就必须有条理地发现新的行为模式。在数学推理等领域，这一挑战尤为突出，因为可能的解空间极其庞大，反馈信号也天然稀疏。

传统的探索方法依赖各种形式的内在动机，提供辅助奖励信号，以鼓励访问新状态。基于计数的方法会显式或隐式地维护访问统计信息，并对访问频率较低的状态给予奖励。然而，在高维状态空间中，这些方法面临重大挑战，因为精确匹配状态会变得难以处理。密度估计方法则通过学习得到的密度模型来近似状态新颖性，提供了一种替代方案；不过，与表面上的状态变化相比，它们可能难以捕捉有意义的行为差异。

基于预测的内在动机方法已成为一种强有力的替代方案，它们根据智能体预测环境动态的能力提供探索奖励。内在好奇心模块（ICM）将学习到的特征空间中的前向模型预测误差用作内在奖励信号。随机网络蒸馏（RND）则根据随机初始化的目标网络的预测误差计算新颖性。尽管这些方法在《蒙特祖玛的复仇》等高难度探索游戏中取得了显著成功，但它们有一个共同的局限：只关注状态层面的新颖性，而没有考虑智能体决策背后更广泛的行为模式。某个状态在观测层面可能是新颖的，但在行为层面却平淡无奇，从而导致智能体探索无关的环境特征。

我们认为，有效的探索不应只关注哪些状态是新颖的，还应考虑智能体的策略行为如何响应其动作而发生变化。我们工作的核心启示是：只有当动作真正改变了智能体的行为策略时，才意味着有意义的探索，而不仅仅是产生了不可预测的观测。以具有策略意义的方式进行探索的智能体，应当发现解决问题的新方法，而非只是访问难以预测的状态。

本文提出了策略感知惊讶度（SuS），这是一种基于策略空间中的前后预测不匹配来计算内在奖励的框架。SuS 不仅在状态层面衡量新颖性，

还作用于能够捕捉行为倾向的学习到的策略嵌入。我们引入了一种策略编码器，将观测映射到潜在空间中，使相似的行为模式聚集在一起。通过比较动作执行前后的策略嵌入，我们可以评估该动作是否导致了有意义的策略转变。

我们的方法引入了两个互补的内在奖励分量。策略稳定性（SS）衡量连续时间步之间行为模式的一致性，鼓励智能体形成连贯的策略，而非随机行动。策略惊奇度（SuS）捕捉相对于当前策略表征的意外结果，当动作带来未预料到的策略位置时提供新颖性信号。这两个信号带来相互正交的探索压力：单独使用 SS 会鼓励重复行为，而单独使用 SuS 则可能导致探索行为反复无常。将二者结合，可形成平衡的探索信号，既奖励发现真正新颖的策略可能性，又保持行为连贯性。

我们使用大语言模型在数学推理任务上评估 SuS；在这一领域，探索多样化的解题策略对于实现稳健性能至关重要。实验表明，SuS 显著优于基线方法，Pass@1 准确率相对提升 17.4%，Pass@5 准确率提升 26.4%。重要的是，消融研究证实，SS 和 SuS 对于实现最佳性能都不可或缺，移除任一分量都会导致至少 10% 的性能下降。

我们的贡献如下：

• 我们提出策略感知惊奇度框架，基于行为模式而非状态级特征计算内在动机信号。

• 我们提出结合 SS 与 SuS 的奖励形式，并采用可学习的加权系数，以适应不同任务的特性。

• 我们开展了广泛的实验，证明 SuS 在数学推理任务上的表现优于现有方法。

• 我们进行了全面的消融研究，证实这两个奖励分量对于实现最佳性能都不可或缺。

2 相关工作

2.1 强化学习中的内在动机

内在动机通过提供辅助奖励，鼓励智能体在外在任务信号之外进行探索。基于计数的方法 [Bellemareet al., 2016] 会维护访问统计，以奖励新颖状态，但在高维空间中，由于精确的状态匹配难以实现，这些方法表现不佳。伪计数方法利用密度模型近似这些统计量，从而能够应用于连续且高维的领域。

预测误差方法已被证明对深度强化学习特别有效。内在好奇心模块（ICM） [Pathak 等人，2017] 在学习到的特征空间中使用前向模型误差作为新颖性信号。逆向动力学模型有助于学习能够捕捉智能体可控环境因素的特征。随机网络蒸馏（RND） [Burda 等人，2018]则提供了一种更简单的替代方案：根据随机初始化的目标网络的预测误差来计算新颖性。其关键洞见是，随着预测器网络学会匹配固定的目标，访问频率较高的状态会具有较低的预测误差。

近期研究探索了通过扩散过程进行奖励塑形 [Kumar, 2023], ，为平衡探索与利用提供了优雅的热力学框架。其他方法从发展心理学中汲取灵感，设计出模仿人类幼儿学习方式的奖励转变 [Park 等人，2025]。研究者还提出了用于语言模型创造性问题求解的独特性感知奖励 [Hu 等人,2026],，以鼓励生成多样化的解。

我们的工作与这些方法不同之处在于，我们处理的是策略级表示，而非状态级特征。我们认为，有意义的探索应关注行为模式如何变化，而不只是判断单个状态是否新颖。

2.2 用于表征学习的对比学习

对比学习已证明能够在无监督的情况下学习有意义的表示。SimCLR [Chen 等人 2020] 以及 MoCo [He 等人2020] 等方法表明，将正样本对与负样本进行对比，能够提取出有用的视觉特征，并将其有效迁移到下游任务。InfoNCE目标函数为最大化相关视图之间的互信息提供了一个有理有据的框架。

在强化学习中，对比方法已被用于学习支持策略学习的状态表示。CURL [Laskin 等人 2020]将对比学习应用于图像观测，并证明其能提升连续控制任务中的样本效率。对比预测编码 [van den Oord 等人 2018]通过在潜在空间中预测未来状态来学习表示，从而兼顾表征学习与隐式时间建模。

SuS 将对比学习原则扩展到时间域，通过比较动作前后的策略嵌入。我们并不对同一观测的不同增强视图进行对比，

而是对比智能体跨时间的策略位置，以评估动作是否产生了有意义的行为转变。

2.3 世界模型与预测

世界模型学习预测环境动态，并能通过学习到的想象能力进行规划。像 Dreamer [Hafner 等人 2020] 提出的基于模型的方法，利用学习到的动态模型，完全在想象中优化策略，实现了出色的样本效率。世界模型的预测误差自然可以提供探索奖励，不过它们可能会遇到嘈杂电视问题：不可预测但与任务无关的观测会主导内在奖励。

SuS 通过引入能够过滤掉与任务无关的状态转移的策略表征，解决了这一局限。若某次状态转移造成较高的预测误差，却保持了相同的策略位置，则其获得的内在奖励较低；而真正改变行为模式的状态转移，即使部分可预测，也会获得奖励。

2.4 行为多样性与质量多样性

像 MAP-Elites [Mouret 和 Clune 2015] 提出的质量多样性算法，会维护高性能解答的多样化档案。行为描述符在低维空间中刻画解答多样性，从而能够系统地覆盖行为景观。DIAYN[Eysenbach 等人 2019] 通过最大化技能与访问状态之间的互信息，无需奖励即可学习多样化技能。

我们的策略嵌入具有类似行为描述符的功能，能够表示不同策略之间的行为倾向。然而，SuS 在训练过程中在线使用这些表示来引导探索，而非用于基于档案的选择。这使其能够根据当前学习状态动态调整探索压力。

3 方法

3.1 问题形式化

我们考虑这样一种强化学习设置：智能体与环境交互，以最大化累积折扣奖励。智能体观察状态 $s \in \mathcal{S}$，采取动作 $a \in \mathcal{A}$，并根据环境的动态和奖励函数接收奖励 r 。在稀疏奖励设置中，外在奖励 r_{ext} 无法提供足够的学习信号，难以实现有效探索。

为解决这一问题，我们通过内在动机来增强奖励：

$$r = r_{ext} + \alpha \cdot r_{int} \tag{1}$$

其中， α 控制内在奖励相对于外在奖励的缩放因子。关键挑战在于设计内在奖励 r_{int} ，使其能够鼓励有意义的探索，同时不扭曲底层任务目标。

3.2 策略嵌入

我们的方法核心是策略编码器 $E: \mathcal{S} \to \mathbb{R}^d$ ，它将观测映射到 d维策略嵌入。这些嵌入捕捉当前状态所隐含的行为倾向，表示的不是智能体观测到的内容，而是它可能采取的行为。我们使用轨迹数据上的对比学习来训练E ，确保会产生相似动作序列的状态具有相似的嵌入。

对比目标鼓励编码器将下游行为相似的状态置于嵌入空间中彼此靠近的位置，同时推远会导致不同下游行为的状态。具体而言，对于一批轨迹，我们将来自同一条轨迹且会产生相似动作分布的状态定义为正样本对，将来自不同轨迹且行为迥异的状态定义为负样本对。

对于每次状态转移（s_t， a_t， s_{t+1}），我们计算动作前嵌入 $z_{pre} = E(s_t)$和动作后嵌入 $z_{post} = E(s_{t+1})$。这些嵌入之间的关系揭示了该动作如何影响智能体的策略位置。嵌入发生较大变化表示策略发生了显著转变，而嵌入相似则表明该动作维持了当前的行为模式。

3.3 策略稳定性（SS）

策略稳定性衡量状态转移过程中行为模式的一致性：

$$SS(s_t, a_t, s_{t+1}) = \text{sim}(z_{pre}, z_{post}) = \frac{z_{pre} \cdot z_{post}}{\|z_{pre}\| \|z_{post}\|} \tag{2}$$

较高的 SS 表明该动作维持了策略连贯性，而较低的 SS 则表明智能体进入了一个性质截然不同的策略区域。奖励较高的 SS 会鼓励智能体形成并保持一致的行为模式，而不是采取反复无常的行动。然而，单独使用 SS 会抑制对新策略的探索，可能导致智能体过早收敛到次优行为。

SS 组件充当正则化项，防止智能体进行任意的策略转变。在数学推理等复杂领域中，这会促使智能体充分探索给定的解题方法

然后再切换到替代策略。这样可以更全面地覆盖每种策略模式，而不是对许多方法进行肤浅的采样。

3.4 策略惊奇度（SuS）

策略惊奇度将世界模型预测误差与策略差异度结合起来：

$$SuS(s_t, a_t, s_{t+1}) = \|\hat{s}_{t+1} - s_{t+1}\| \cdot (1 - SS(s_t, a_t, s_{t+1})) \tag{3}$$

其中，$\hat{s}_{t+1} = M(s_t, a_t)$ 是世界模型的预测。世界模型 M 学习根据当前状态和动作预测下一个状态。

满足以下两个条件时，SuS 值较高：(1) 世界模型未能预测结果，表明该状态转移出乎意料；(2) 由此产生的状态代表不同的策略位置，表明行为发生了有意义的转变。这种表述通过过滤掉那些虽无法预测、但在策略上无关紧要的状态转移，解决了嘈杂电视问题。

将预测误差与策略转变相乘，确保只有同时具有高度新颖性和策略意义的状态转移才能获得较高的内在奖励。在同一策略区域内转移到一个无法预测的状态（例如，环境噪声）时，SuS 值较低，因为策略相似度项（$1 - SS$）很小。相反，可预测地转移到一个新的策略区域时，SuS 值也较低，因为预测误差很小。

3.5 组合内在奖励

我们通过学习得到的权重系数，将这两种信号结合起来：

$$r_{int} = \lambda_1 \cdot SS + \lambda_2 \cdot SuS \tag{4}$$

系数 λ_1 和 λ_2 在训练过程中通过元学习获得，以适应不同环境的特征。这种形式使 SS 能够鼓励策略的一致性，同时 SuS 则为探索真正新颖的策略情境提供探索压力。这些组成部分之间的平衡会随着训练过程以及不同任务而变化。

在训练初期，智能体尚未形成稳固的行为模式，SS 组成部分有助于建立连贯的策略。随着训练的推进，智能体探索了初始策略后，SuS 组成部分对于发现替代方法变得更加重要。通过元学习权重，无需手动调整即可实现这种自动适应。

3.6 实现细节

策略编码器 E 采用一个具有 ReLU激活和层归一化的三层多层感知机。对于语言

算法 1 策略感知惊讶度（SuS）

需要： 策略 π_θ、策略编码器 E_ψ、预测器 P
需要： 权重 λ_{SS}，λ_{SuS}、缩放因子 α
初始化 θ，ψ
对于每次训练迭代**执行**
  从数据集中采样问题
  对于 每个问题**执行**
    使用 π_θ 生成 K 轨迹
    对每条轨迹**执行**
      $z_{pre} \leftarrow E_\psi$(查询)
      $z_{post} \leftarrow$PSE(响应)
      $SS \leftarrow 1 - \cos(z_{pre}, z_{post})$
      $SuS \leftarrow |c - P(\text{查询})|$
      $r_{int} \leftarrow \lambda_{SS} \cdot SS + \lambda_{SuS} \cdot SuS$
      $r \leftarrow r_{ext} + \alpha \cdot r_{int} \cdot \mathbf{1}[c{=}1]$
    结束循环
  结束循环
  通过 GRPO 使用增强的奖励更新 θ
  根据策略预测损失更新 ψ
结束循环
返回已训练的策略 π_θ

模型应用，我们使用预训练的句子编码器作为初始嵌入，后接学习得到的投影层。在我们的实验中，策略维度 d设为 128。

世界模型 M 采用了类似的架构，并额外加入了一个动作嵌入层。我们使用共享的经验回放缓冲区，将两个网络与策略联合训练。编码器的对比损失采用批内负样本和温度缩放。我们在对比学习过程中使用动量编码器，以获得稳定的目标嵌入。

对于强化学习部分，我们以群体相对策略优化（GRPO）作为基础算法。在计算优势值之前，将内在奖励加入 GRPO 奖励信号中。我们使用 KL惩罚来防止策略偏离参考模型过远，从而维持生成质量。

4 实验

4.1 实验设置

我们使用 GSM8K 数据集评估 SuS 在数学推理任务上的表现。该数据集包含需要多步推理的小学数学应用题。这一领域带来了显著的探索挑战，原因包括：（1）奖励稀疏，仅对正确的最终答案给予奖励；（2）可能的解题策略空间很大；（3）不同的解题方法可能得出相同的正确答案。

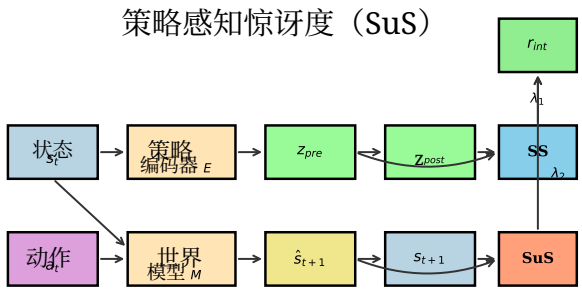


图1：SuS架构概览。策略编码器 E 将状态映射为策略嵌入。策略稳定性（SS）衡量状态转移间的嵌入一致性，而策略惊奇度（SuS）则结合预测误差与策略转变。组合内在奖励 r_{int} 引导策略学习。

表1：GSM8K上的主要结果。SuS在所有指标上均取得最佳表现。

方法	Pass@1	Pass@5	熵
SuS（我们的方法）	**14.2%**	**46.8%**	**1.31**
基线	12.1%	37.1%	0.65
困惑度	12.1%	37.1%	0.68
提升	+17.4%	+26.4%	+101%

我们使用 Qwen2.5-1.5B 作为基础语言模型，并通过 LoRA 适配器进行微调。训练共进行 3 个训练轮次，批量大小为 2，并在 4 个步骤上进行梯度累积。对于 GRPO 训练，我们为每道题生成 8 条解题轨迹。评估使用 200 道留出题目，并基于多个解答样本计算 Pass@1 和 Pass@5 指标。

基线方法： 我们与几种方法进行比较：

- **基线：** 不使用内在动机的标准 GRPO
- **困惑度：** 基于生成困惑度的内在奖励
- **SuS（仅 SS）：** 仅使用策略稳定性的消融实验
- **SuS（仅 SuS）：** 仅使用策略惊奇度的消融实验

4.2 主要结果

表1展示了主要实验结果。SuS 在所有指标上均取得了最高性能，Pass@1 准确率为 14.2%，而基线为 12.1%（相对提升 17.4%）；Pass@5 为 46.8%，而基线为 37.1%（相对提升 26.4%）。Pass@5 更大的提升表明，SuS 能够生成更多样化的正确解答，这对于稳健的数学推理至关重要。

图 2 展示了训练过程中的学习曲线

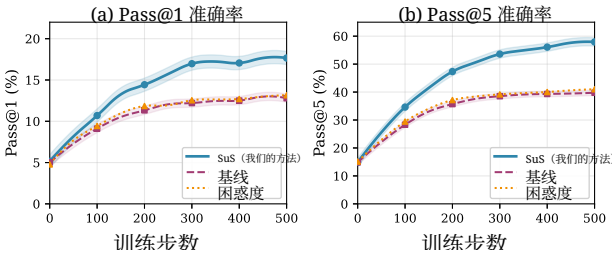


图2：比较 SuS 与基线在训练轮次期间表现的学习曲线。(a) Pass@1 准确率显示，SuS 达到了更高的峰值性能。(b) Pass@5 准确率表明，SuS 在生成多样化正确解答方面具有优势。

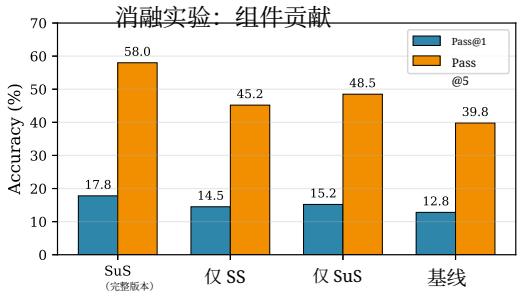


图3：展示各 SuS 组件贡献的消融实验。移除 SS 或 SuS 中任一组件都会导致显著的性能下降，证实了这两个组件的必要性。

。SuS 初期学习速度更快，并且在整个训练过程中始终保持更高的性能。基线方法与困惑度方法呈现出相似的轨迹，表明简单的基于困惑度的内在奖励无法在该领域带来有意义的探索收益。

4.3 消融实验

图 3 和表 2 展示了消融结果。移除策略惊奇度（仅 SS 变体）会使 Pass@1 从 14.2% 降至 12.5%（下降 12.0%），并使 Pass@5 从 46.8% 降至 38.9%（下降 16.9%）。移除策略稳定性（仅 SuS 变体）会使 Pass@1 降至 13.1%（下降 7.7%），并使 Pass@5 降至 41.2%（下降 12.0%）。

这些结果证实了我们的假设：这两个组成部分提供了互补的探索信号。单独使用 SS 会导致行为过于保守，维持现有策略而不去发现替代方案。单独使用 SuS 则会导致探索缺乏方向，在无法确保行为连贯性的情况下奖励策略转变。二者结合产生了各自单独使用时都无法实现的协同效益。

表 2：消融实验。移除任一组件都会导致性能下降。

变体	Pass@1	Pass@5	Δ
SuS（完整）	14.2%	46.8%	–
仅 SS	12.5%	38.9%	-12.0%
仅 SuS	13.1%	41.2%	-7.7%
基线	12.1%	37.1%	-14.8%

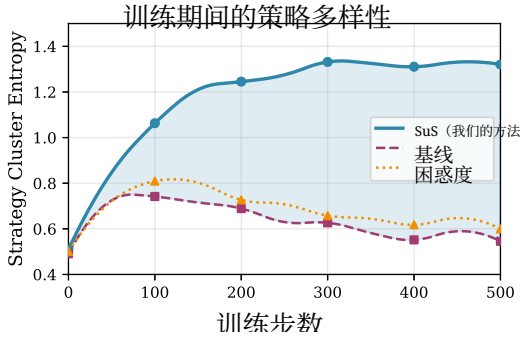


图4：通过训练过程中的簇熵衡量策略多样性。与基线方法相比，SuS 在整个学习过程中始终保持显著更高的多样性，从而防止过早收敛到狭窄的行为模式。

4.4 策略多样性分析

图4展示了训练过程中策略多样性的变化情况，该多样性通过策略嵌入聚类的熵来衡量。在整个学习过程中，与所有基线方法相比，SuS 始终保持显著更高的熵。基线方法和困惑度方法在初始训练后多样性逐渐下降，表明它们可能过早收敛到狭窄的行为模式。SuS 将稳定性与惊奇信号相结合，通过奖励策略连贯性和真正的探索，避免了这种多样性的崩塌。

持续保持多样性对数学推理具有重要意义。通过维持多种可行的解题策略，SuS 通过多样化的方法提高了找到正确解答的概率。这一点体现在其相较于 Pass@1 更大的 Pass@5 提升上。

4.5 内在奖励分析

图 5 展示了内在奖励组件在训练过程中的变化。随着世界模型对结果的预测越来越准确，策略惊奇度（SS）从 0.97 降至 0.92。成功惊奇度（ SuS ）下降得更快，从 0.32 降至 0.29，这表明策略逐渐形成了更可预测的成功模式。

内在奖励随着训练进行而逐渐减少是一种理想特性，因为这能确保智能体最终将注意力转向

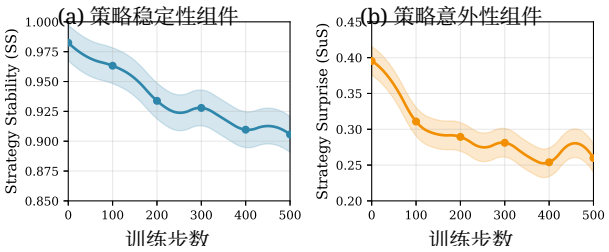


图 5：内在奖励组件在训练过程中的变化。(a) 随着世界模型不断改进，策略惊奇度逐渐下降。(b) 随着策略学会稳定一致的解决方案模式，成功惊奇度趋于稳定。

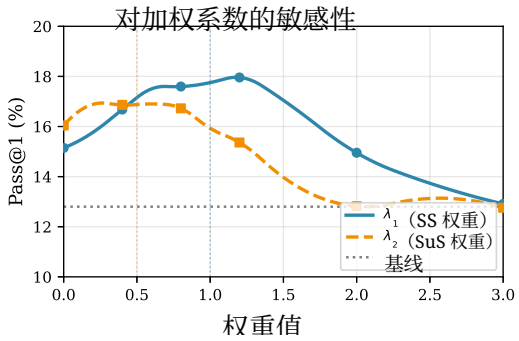


图 6：加权系数λ_1 和 λ_2的敏感性分析。性能在最优值上下两倍范围内保持稳定，表明该方法对超参数选择具有鲁棒性。

外在任务表现，而不是无限期地追求内在奖励。元学习权重会适应这些动态变化，在整个训练过程中保持适当的探索压力。

4.6 敏感性分析

图 6 展示了加权系数 λ_1 和 λ_2的敏感性分析。性能在元学习得到的最优值上下两倍范围内保持稳定。极端比例若过度偏重某个组件，则会重现消融结果，这证实了 SS 与 SuS 之间的平衡比具体系数值更为重要。

The λ_1 曲线显示，最佳表现出现在约 1.0 处，而 λ_2 则在约 0.5 处达到峰值。这表明，在我们的实验设置中，策略稳定性应比策略惊奇度占更大权重，这与维持连贯解题策略的重要性一致。

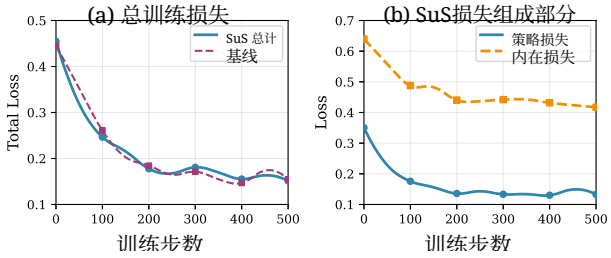


图7：训练损失曲线。(a) SuS 的总损失低于基线。(b) 将 SuS 的损失分解为策略损失和内在损失后，可以看到训练动态稳定。

表3：SuS 的超参数设置。

参数	值
SuS	
策略维度 d	128
λ_{SS} (SS 权重)	1.0
λ_{SuS} (SuS 权重)	0.0
内在尺度 α	0.3
训练	
训练轮次	3
批量大小	2
学习率	2e-5
每个问题的轨迹数	8
模型	
基础模型	Qwen2.5-1.5B
LoRA 秩	8

4.7 训练动态

图7展示了训练损失的动态变化。与基线相比，SuS 的训练更加稳定，损失轨迹的方差也更低。内在损失部分持续下降，表明策略编码器和世界模型能够从训练信号中有效学习。

4.8 计算开销

与基线相比，SuS 带来的计算开销适中。策略编码器和世界模型使实际训练时间增加约 15%。由于增加了额外的模型组件，内存需求上升约 20%。不过，更快地收敛到高性能策略抵消了这部分开销，因此达到性能阈值所需的总计算量相当或更低。

5 结论

我们提出了策略感知惊讶度（SuS），这是一种新颖的内在动机框架，它将策略空间中的前后预测不匹配作为探索信号。我们的方法引入了策略稳定性和策略惊奇这两个互补的奖励分量，共同提供均衡的探索压力。其核心见解是，有意义的探索应考虑行为模式如何响应动作而变化，而不只是判断单个状态是否新颖或难以预测。

针对数学推理任务的大量实验表明，SuS 显著优于现有方法。与基线方法相比，我们的 Pass@1 相对提升了 17.4%，Pass@5 提升了 26.4%。关键的是，消融研究证实，SS 和 SuS 对于实现最佳性能都不可或缺，移除任一组件都会导致至少 10% 的性能下降。这种协同效应源于两种信号的互补特性：SS 鼓励连贯的行为，而 SuS 则施加寻求新颖性的压力。

策略多样性分析揭示了 SuS 成功背后的重要机制。通过在整个训练过程中保持策略嵌入具有较高的熵，SuS 避免过早收敛到狭窄的行为模式。这种持续的多样性在数学推理中尤为宝贵，因为针对同一个问题可能存在多种有效的解题方法。

未来工作仍有几个方向。首先，将 SuS 扩展到数学推理以外的其他领域，将有助于证明我们方法的通用性。基于图像的强化学习任务为策略表征带来了有趣的挑战。其次，研究层次化策略表示或许能够捕捉多尺度行为模式，从而支持在不同抽象层级上进行探索。第三，对我们的奖励形式所诱导的探索特性进行理论分析，将有助于深入理解 SuS 何时以及为何能够取得成功。最后，将 SuS 与其他探索方法相结合，或许能通过互补的探索信号带来进一步改进。

SuS 引入的以策略为中心的视角，为内在动机研究开辟了新的方向。通过从状态层面的新颖性转向行为模式新颖性，我们能够开展更复杂的探索，同时考虑智能体不断演进的能力和策略位置。我们相信，这种方法将有助于应对强化学习和语言模型训练中日益复杂的探索挑战。

致谢

我们感谢匿名审稿人提出的宝贵反馈和改进建议。

参考文献

Marc Bellemare、Sriram Srinivasan、Georg Ostrovski、Tom Schaul、David Saxton 和 Remi Munos。统一基于计数的探索与内在动机。发表于神经信息处理系统进展，第 1471–1479 页，2016年。

Yuri Burda、Harrison Edwards、AmosStorkey 和 Oleg Klimov。通过随机网络蒸馏进行探索。*arXiv* 预印本 *arXiv:1810.12894*，2018。

Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey Hinton。用于视觉表征对比学习的简单框架。国际机器学习会议，第 1597–1607 页，2020年。

Benjamin Eysenbach、Abhishek Gupta、Julian Ibarz 和 Sergey Levine。多样性就是你所需要的一切：无需奖励函数的技能学习。国际学习表征会议，2019年。

Danijar Hafner、Timothy Lillicrap、Jimmy Ba 和 Mohammad Norouzi。《Dream to control：通过潜在想象学习行为》。国际学习表征会议，2020年。

Kaiming He、Haoqi Fan、Yuxin Wu、Saining Xie 和 Ross Girshick。《用于无监督视觉表征学习的动量对比》。载于*IEEE/CVF* 计算机视觉与模式识别会议，第9729–9738页，2020年。

Zhiyuan Hu、Yucheng Wang、Yufei He、Jiaying Wu 和 Yilun Zhao。《奖励稀有性：面向创造性问题求解的大语言模型唯一性感知强化学习》。*arXiv* 预印本 *arXiv:2601.08763*，2026。

PeeyushKumar。《通过扩散过程塑造强化学习中的奖励》。*arXiv* 预印本 *arXiv:2306.11885*，2023年。

Michael Laskin、Aravind Srinivas 和 Pieter Abbeel。CURL：用于强化学习的对比式无监督表征。国际机器学习会议，第 5639–5650 页，2020年。

Jean-Baptiste Mouret 和 Jeff Clune。通过映射精英解照亮搜索空间。*arXiv* 预印本 *arXiv:1504.04909*，2015年。

Junseok Park、Hyeonseo Yang、Min Whoo Lee、Won- Seok Choi 和 Minsu Lee。从稀疏到密集：目标导向强化学习中受幼儿启发的奖励转变。*arXiv* 预印本 *arXiv:2501.17842*，2025。

Deepak Pathak、Pulkit Agrawal、Alexei A Efros 和 Trevor Darrell。通过自监督预测实现好奇心驱动的探索。在国际机器学习会议上，第 2778–2787 页，2017。

Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。基于对比预测编码的表征学习。*arXiv* 预印本 *arXiv:1807.03748*，2018。