arXiv:2304.00367v1 [cs.RO] 1 Apr 2023

通过对比行为摘要传达自主机器人能力

Peter Du、Surya Murthy 和 Katherine Driggs-Campbell

摘要—随着人工智能的发展，基于学习的自主智能体的能力日益增强，人类观察者也越来越难以高效地构建智能体行为的心理模型。为了成功部署自主智能体，人类不仅应能够理解各个智能体的局限性，还应了解它们之间的相对表现。为此，我们需要有效的方法来生成符合人类理解的智能体行为摘要。过去，单智能体行为总结通常通过生成解释来说明智能体为何在某一时间步选择特定动作。然而，对于复杂任务而言，针对每个动作的解释可能无法传达智能体的全局策略。因此，研究人员开始采用多时间步摘要，以更好地帮助人类评估智能体的整体能力。近来，多步摘要也被用于生成对比示例，以评估多个智能体。然而，过去的方法主要依赖非结构化搜索方法来生成摘要，并且要求智能体具有离散的动作空间。本文提出一种自适应搜索方法，用于高效生成对比行为摘要，并支持连续状态与动作空间。我们开展了一项用户研究，以评估这些摘要在帮助人类辨别给定任务中性能更优的自主智能体方面的有效性。结果表明，在有限的观察时间预算下，自适应搜索能够高效识别具有信息量的对比场景，使人类能够准确选择表现更好的智能体。

I. 引言

自主智能体正在迅速开发并部署到此前由人类主导的领域中。基于学习的技术近期取得的进展，使自主智能体能够越来越胜任需要与其他人互动的复杂任务 [1]–[4]。如今，从在人群中行驶的自动驾驶车辆，到穿行于密集人群中的机器人，这些智能体已应用于各种领域。为了让系统设计者部署合适的智能体，从而实现成功且安全的人类与自主系统交互，我们需要相应的方法，帮助人类理解自主智能体的行为，并识别特定智能体的能力相较于其他智能体存在的局限性。

过去，已有不少研究致力于解决如何向人类传达自主智能体行为的问题。这些研究大致可分为两类：局部方法和全局方法。局部方法着眼于单个状态，旨在解释智能体为何在该状态下做出特定决策。对智能体行为的局部解释可能无法展现智能体的整体策略，因为某些行为可能需要跨越多个状态的一系列相互依赖的决策。全局方法试图解决这一问题，而是通过向人类观察者传达完整的动作序列，来解释智能体的整体行为 [5], [6]。

看到自主智能体执行任务的示例后，人们自然会开始熟悉其行为，并完善自己对智能体将如何行动的心理模型 [7]。向人们展示智能体在各种场景中的行动示例，是让他们了解智能体全局行为的一种直接但低效的方式。然而，这种做法并不实际，因为人类观察者用于构建智能体心理模型的时间有限，而且并非所有场景都能同样有效地向观察者传递信息。例如，展示自动驾驶车辆在笔直道路上行驶、且没有与其他驾驶者互动的场景，并不能让人类观察者充分了解车辆在安全关键情境下会如何反应。

为解决这一问题，研究者提出了基于状态重要性生成信息更丰富的行为摘要的算法 [5], [8]。随后，他们对这些算法进行了改进，通过使用分歧度量生成涉及多个智能体的对比摘要，以应对智能体比较任务 [9], [10]。然而，这些方法缺少一种更有结构的搜索方法来探索场景空间（它们依赖仿真器随机化，或在智能体训练期间并行运行），而且无法推广到具有连续动作的智能体。

本文针对这些局限性，提出了一种用于对比行为摘要的结构化搜索方法。我们的贡献如下。

1) 我们提出一种全局行为摘要方法，帮助人类观察者比较多个自主智能体的能力。我们放宽了现有方法的要求，不再将分析局限于在离散动作空间中运行的智能体。2) 我们提出一种基于自适应搜索的算法，用于生成对比行为摘要，并能高效探索大型多智能体场景空间。3) 我们使用这些方法为多个基于学习的群体导航机器人生成对比摘要。我们开展了一项用户研究，以评估这些摘要能否帮助真实的人类高效、准确地辨别智能体性能差异。

P. Du、S. Murthy 和 K. Driggs-Campbell 任职于伊利诺伊大学厄巴纳-香槟分校电气与计算机工程系。邮箱：{peterdu2,suryakm2,krdc}@illinois.edu

II. 背景

A. 解释智能体动作

随着基于学习的技术不断进步，自主智能体在复杂任务中的性能得以提升，解释这些智能体决策的方法也日益受到关注。许多研究分析了如何向人类传达自主智能体的重要行为特征，以帮助人类更好地理解智能体的能力与局限性 [11]–[13]。

已有多种用于解释智能体行为的“局部”方法 [14]–[16]。这些方法通常针对智能体在单个状态下执行的某个单个动作，并试图对其进行解释。在关键状态等特定场景中 [17], , 局部解释可能很有帮助，但未必能有效传达智能体的整体策略。

人类通过观看智能体行为的示例动作序列，能够了解智能体的特征，并建立关于其将如何行动的心理模型 [7]。因此，人们对行为摘要的“全局”方法越来越感兴趣；这类方法旨在向人类观察者传达智能体的全局策略，而非解释单个动作。全局方法并不解释智能体选择某个特定动作的原因，而是尝试生成连续的动作序列（轨迹），为试图构建心理模型的人类提供有用信息。在 [5], [8], 中，作者提出了一种为单个智能体生成全局行为摘要的算法。该方法依赖一种状态重要性度量，用于衡量智能体在特定状态下Q值的差异。在智能体执行过程中，系统会收集重要性较高的状态，以形成供人类观察者查看的候选轨迹。

尽管单智能体摘要方法已被证明能够提升人们准确辨别智能体能力的水平，但当人们希望在多个智能体之间进行比较 时，这些方法在传达哪个智能体更优方面效果较差。在这些研究中，每个智能体的摘要都是独立生成的，因此可能无法捕捉不同智能体行为出现分歧的有意义场景。文献 [9] 和 [10]提出了针对比较需求、用于生成两个智能体行为摘要的改进算法。作者没有利用状态重要性为单个策略生成轨迹，而是提出了一种耦合方法：根据正在比较的两个智能体之间的分歧来确定行为摘要。这种耦合使生成的摘要能够展示两个智能体选择不同动作的场景。

迄今讨论的所有全局行为摘要方法都有一个共同特征：它们依赖离散智能体动作空间来评估相关指标，从而确定所生成的轨迹（摘要）。例如，在 [9], [10] 中，s_D “分歧状态” A_1 （两个智能体 A_2 的策略 π_1 和 π_2 满足 $\pi_1(s) \neq \pi_2(s)$ 的状态）这一概念被用于确定对比行为摘要的起始位置。对于映射到连续动作空间的策略而言，这种分歧状态的定义会变得难以处理，因为策略之间的差异会导致几乎所有状态都被归类为分歧状态。可以通过添加误差阈值来缓解这一问题，使得需要满足 $|\pi_1(s) - \pi_2(s)| < \epsilon$，然而， ϵ 的选择可能会极大地影响所生成的摘要，并且必须针对每个应用手动调节。

生成的摘要也高度依赖于用户如何步进包含这些智能体的仿真，因为这些算法几乎没有为探索场景空间提供指导。本文探讨使用基于差异度的度量来生成对比行为摘要，以帮助人们比较自主智能体。然而，与以往方法不同，我们在场景空间中进行自适应搜索，以最大化该差异度度量。通过针对完整轨迹优化该度量，我们消除了对离散智能体动作空间的限制，并展示了如何为同时具有连续状态与动作空间的智能体高效生成信息丰富的对比行为摘要。

B. 自适应场景搜索

自适应搜索技术已应用于解决黑盒或半黑盒学习系统的验证问题。自适应压力测试（AST）就是其中一种技术，它将验证问题描述为马尔可夫决策过程（MDP），并使用强化学习（RL）生成策略，以产生相关的失败示例。这种方法已成功应用于飞机防撞、自动驾驶汽车等多个领域 [18], [19]。AST 由三个组件模块组成，分别是仿真器 $\mathscr{S}$、奖励函数 R 和强化学习求解器 S。仿真器包含环境和被测智能体。强化学习求解器从一组环境动作中选择动作 $\mathcal{A}$ ，以驱动仿真器步进。在每个步进中，求解器都会收到一个奖励，以更新其策略。用于自适应故障搜索的奖励函数示例如下：

$$R = \begin{cases} 0 & s \in E \\ -\alpha - \beta f(s) & s \notin E, t \geq T \\ -g(a) - \eta h(s) & s \notin E, t < T \end{cases}$$

其中， s 是仿真器的当前状态（或观测），E 是搜索的目标状态集合， T 是有限的时间范围， g 是基于环境动作的奖励， f、 h 是用于引导搜索过程的基于启发式的奖励， α、β 是可调参数。

后续研究探索了多智能体验证和奖励增强，以生成更具信息量的失效模式。差分自适应压力测试对比两个智能体，并搜索其中一个智能体会失败的场景 [20]。在这种表述中，对于预期会失败的智能体，采用标准的单智能体 AST 奖励；而第二个智能体则使用取反后的奖励。因此，搜索旨在找到一种场景：一个智能体最终失败，而另一个则不会。奖励增强和关键状态也已应用于 AST 表述

以提升搜索所返回的失效模式的多样性和相关性 [21] [22]。先前研究表明，自适应搜索能够高效地搜索大规模多智能体状态空间，并生成信息丰富的场景，呈现自主智能体的失效空间。本文利用自适应搜索的这些特性，提出一种生成对比行为摘要的方法，使人类观察者能够准确判断一个智能体相较于其他智能体的能力。

III. 方法

A. 耦合仿真环境

对于两个需要生成对比行为摘要的智能体，我们创建一个耦合模拟器 $\mathscr{S}_c$来执行自适应搜索。在模拟器中，有两个独立的仿真实例 $\mathscr{S}_c^{(1)}$和 $\mathscr{S}_c^{(2)}$，分别对应智能体 $\mathscr{A}^{(1)}$ 和 $\mathscr{A}_{(2)}$，其策略分别为 $\pi^{(1)}$ 和 $\pi^{(2)}$。仿真实例包含智能体所需的所有环境信息，以及可能与智能体交互的任何环境参与者。我们将仿真的状态表示如下：令 s_{env} 表示仿真环境的状态， $s^{(i)}$ 表示第 i 个智能体的状态。耦合模拟器 $\mathscr{S}_c$ 可通过环境动作 a_{env}向前步进。为确保行为摘要展示智能体在相同 场景中运行的示例，a_{env} 在两个仿真实例 $\mathscr{S}_c^{(1)}$和 $\mathscr{S}_{(2)c}$之间共享。环境动作可以改变模拟器中的任何参数或动作，但不能改变智能体本身。给定一系列固定的环境动作，仿真实例的演化是确定性的，因此两个智能体能够体验相同的环境。

将仿真向前推进一个时间步分为两个步骤。在搜索过程中，强化学习求解器首先从固定的环境动作空间中生成一个动作 a_{env}。该动作会被传递给仿真器中的两个仿真实例，并用于更新环境状态 s_{env}。随后，每个智能体都会根据自身的策略 $\pi^{(i)}$更新其状态 $s^{(i)}$。

B. 基于散度的奖励

我们设计搜索过程，以寻找这样的场景：处于对比中的智能体即使在相同环境中采取动作，仍表现出显著不同的行为。因此，搜索得到的场景具有对比性。社会科学领域的研究认为，对比示例可能更容易向人类解释，因为人们只需理解两个案例之间的差异，而不必确定所示场景的所有成因。换言之，外行可能会觉得对比示例更直观、更具信息量 [23]。

我们考察影响给定行为摘要中两个智能体之间对比度的两个方面：状态和动作。在仿真的每个步进中，智能体所选动作之间的散度会直接影响下一时间步对其行为差异的感知。在一条轨迹的过程中，这种差异会使智能体的状态逐渐分化，从而生成对比度较高的摘要。这里，我们使用动作向量之差的范数来量化智能体动作的散度。对于任意两个进行比较的智能体（$\mathscr{A}^{(1)}$ 和 $\mathscr{A}^{(2)}$），设 $\pi^{(1)}$和 $\pi^{(2)}$ 分别为这两个智能体的策略。设这两个智能体在给定环境状态 s_{env} 下的动作由 $\pi^{(1)}(s_{env})$ $\pi^{(2)}(s_{env})$ $\in \mathbb{R}^m$生成。环境状态 s_{env} 下的动作散度 d_a 定义如下：

$$d_a(s_{env}) = \sum_{i=1}^{m} \left[\pi^{(1)}(s_{env})[i] - \pi^{(2)}(s_{env})[i]\right]^2$$

仅有较高的动作散度，并不足以引导场景搜索找到智能体之间具有较高对比度的摘要。例如，考虑一条轨迹，其中两个智能体在每个时间步都交替采取动作 a 和 $-a$，且彼此采取相反的动作。在正向和负向动作之间交替，会产生较高的动作散度，但最终得到的状态几乎没有对比度。在查看摘要时，人类是通过智能体的状态来感知高度对比的行为的。因此，我们选择的场景搜索奖励会将轨迹末尾的状态散度 d_s 纳入考量。设摘要的智能体最终状态为$s^{(1)}, s^{(2)} \in \mathbb{R}^n$。状态散度 d_s 定义为：

$$d_s(s^{(1)}, s^{(2)}) = \sum_{i=1}^{n} \left[s^{(1)}[i] - s^{(2)}[i]\right]^2$$

结合动作散度和状态散度，用于自适应搜索对比行为摘要的奖励函数如下：

$$R = \begin{cases} \alpha \cdot d_s(s^{(1)}, s^{(2)}) & s^{(1)} \text{ or } s^{(2)} \in E \\ \beta \cdot h(s_{env}, s^{(1)}, s^{(2)}) & s^{(1)}, s^{(2)} \notin E, t \geq T \\ d_a(s_{env}) & s^{(1)}, s^{(2)} \notin E, t < T \end{cases}$$

其中， E 是一组智能体目标状态， T 是有限搜索时域， h 是用于引导搜索的可选启发式方法，而 α和 β是可调常数参数。

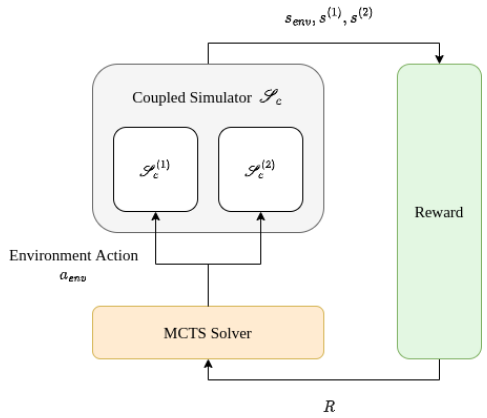

图 1：自适应场景搜索

C. 自适应场景搜索

给定 N 个智能体 $\mathscr{A}(1) \ldots, \mathscr{A}(N)$，我们通过 $\frac{1}{2}N$ 次（$N-1$）自适应搜索，为所有唯一的智能体对生成对比行为摘要。搜索过程的框图如图 1 所示。开始搜索前，我们先创建一个长度为 K 的优先队列 PQ，用于存放按对应累计奖励排序的轨迹序列。接着，我们初始化一个蒙特卡洛树搜索求解器，用于生成环境动作以推进仿真。每次搜索迭代开始时，仿真环境和待对比的智能体都会重置为固定初始状态 s_{init}。在每次迭代的每个时间步中，蒙特卡洛树搜索求解器都会推进仿真，收集奖励函数 R 提供的奖励信号，并更新其内部策略。如果仿真达到终止状态，或搜索的时间范围已到，则将从 s_{init} 开始的当前环境动作序列加入优先队列 PQ，并开始下一次搜索迭代。

算法 1: 自适应场景搜索

输入： $n_{itr}, s_{init}, \mathscr{S}_c, K, T$

输出： X

$PQ \leftarrow priorityQueue(K)$;
$S \leftarrow solverMCTS$;
$R \leftarrow rewardFunction$;
$i \leftarrow 0$;
当 $i \leq n_{itr}$ **时执行**
  $step \leftarrow 0$;
  $T \leftarrow emptyList$;
  $\mathscr{S}_c.reset(s_{init})$;
  当 非 $\mathscr{S}_c.isTerminal$ 且 $step < T$ **时**
    $a_{env} \leftarrow S.getAction(\mathscr{S}_c)$;
    $\mathscr{S}_c.step(a_{env}, \pi^{(1)}, \pi^{(2)})$;
    $\mathcal{R} \leftarrow R(\mathscr{S}_c)$;
    $S.updatePolicy(\mathcal{R})$;
    $T.append((a_{env}, \mathcal{R}))$;
    $step \leftarrow step + 1$;
  end
  $trajReward \leftarrow T.getReward + R(\mathscr{S}_c)$;
  $PQ.add((T, trajReward))$;
  如果 $PQ.size \geq K$，**那么**
    $PQ.popMin$
  end
  $i \leftarrow i + 1$;
end
$X \leftarrow PQ.getTrajectories(K)$;
返回 X

IV. 评估

A. 去中心化结构化循环神经网络人群导航

我们测试了在机器人群体导航场景中使用自适应搜索生成对比行为摘要的方法。图 2 展示了该场景的一个示例。在此设置中，我们有 10 个环境智能体（编号为 0 至 9），代表一群人。一个自主机器人需要穿过人群并避免与人类接触，从而导航至目标状态。

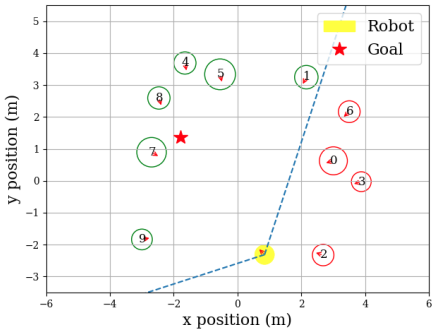

图 2：机器人群体导航场景

我们采用一个去中心化结构化循环神经网络（DS-RNN）强化学习智能体，并对其进行训练，使其能够穿越密集的人群，到达目标状态 [2]。DS-RNN 是机器人群体导航领域的先进方法，融合了空间推理和时间推理，使智能体能够在仿真和现实世界中安全、高效地应对具有挑战性的情境。

在我们的评估中，我们使用无模型深度强化学习（DRL）训练 DS-RNN 人群导航智能体。我们通过选取 DS-RNN 策略在训练过程三个不同阶段的快照，生成三个用于比较的智能体，从而得到一组具有真实性能排名的智能体。这些智能体分别记为 $\mathscr{A}(lo)$ $\mathscr{A}(med)$ $\mathscr{A}(hi)$，按性能从差到优排列。

表 II：DS-RNN 智能体

智能体	训练时间步	平均回合奖励
$\mathscr{A}(lo)$	5.0×10^6	10.0
$\mathscr{A}(med)$	7.4×10^6	16.5
$\mathscr{A}(hi)$	1.2×10^7	21.5

B. 行为摘要生成

我们使用自适应搜索，为所有成对的 DS-RNN 智能体生成对比行为摘要：($\mathscr{A}(lo)\mathscr{A}(med)$)、($\mathscr{A}(med)\mathscr{A}(hi)$)、($\mathscr{A}(lo)\mathscr{A}(hi)$)。对于每一对，我们考虑两个初始状态，它们代表智能体可能需要应对的不同场景。第一个场景要求机器人从空间的左下角前往右上角，第二个场景则要求机器人从左上角前往右下角。我们为每一对智能体和每个初始状态分别运行一个独立的自适应搜索实例。搜索结束后，我们从优先队列中选出奖励最高的轨迹，作为该实例的行为摘要。因此，我们生成了一组摘要，其中包含 6 条成对对比轨迹。

除了通过自适应搜索获得的集合外，我们还通过均匀

表 I：参与者对智能体进行正确排序的准确率

	第一部分（成对排序）			第二部分（全局排序）		
生成方案	正确	错误	**准确度**	正确	错误	**准确度**
N优先轨迹	65	55	**54.17%**	12	28	**30.00%**
自适应搜索	114	6	**95.00%**	26	14	**65.00%**

采样环境动作空间，生成一组基线（N-first）摘要。这种方法类似于人类观察者在固定观看预算内观看若干条轨迹，以构建智能体心理模型的情形。由于轨迹是均匀采样的，观察者将看到智能体更常表现出的行为。为使输出与自适应搜索相匹配，我们将观看预算设为 6 条摘要。

C. 用户研究流程

我们开展了一项用户研究，以评估行为摘要的有效性。我们通过 Amazon Mechanical Turk 招募了共 40 名参与者（男女各占 50%），并要求他们以调查问卷的形式回答问题。在开始之前，我们向参与者展示了一个机器人群体导航任务示例，以确保他们理解该场景。如第 IV-B 节所述，我们采用两组初始状态和目标状态，代表两个不同的场景。调查问卷开始时，参与者需要随机选择一个 1 到 4 之间的数值，据此将其分配到特定场景。为鼓励参与者完成调查问卷，在成功提交后，他们将获得 4 美元的报酬。

用户研究调查问卷由两部分组成。第一部分中，我们评估参与者在给定对比行为摘要的情况下，通过成对示例辨别优胜智能体的能力。参与者需要观看以短视频片段形式呈现的摘要，每段视频都包含两台朝同一目标移动的机器人。根据所观察到的机器人智能体的行为，参与者需要从中选出优胜智能体，并简要说明作出这一选择的理由。随后，参与者需要使用从“非常容易”到“非常困难”的 9 点语义差异量表，量化他们感知到的辨别较优智能体的难度。针对全部 6 个对比摘要，重复这一流程。

在调查问卷的第二部分中，我们通过成对对比行为摘要来考察多智能体排名。参与者会观看三个视频片段，涵盖三个 DS-RNN 智能体（lo、 med、hi）的所有成对摘要。观看视频后，参与者需要根据观察到的行为，将智能体从最好到最差进行排名。此外，我们还请他们简要说明为何选择该排名。参与者还需使用从“非常小”到“非常大”的 9 点语义差异量表，评定他们感知到的智能体之间的差异。最后，他们需要使用 5 点李克特量表，评定自己对“区分智能体对很容易”这一陈述的认同程度。在所有示例中，DS-RNN 智能体均以通用数字命名，以避免使参与者对某个特定智能体产生偏见

。在用户研究的第一部分和第二部分中展示的对比行为摘要采用了独特的初始位置和目标位置，以确保调查问卷的两个部分相互独立。

V. 结果

本节将介绍用户研究的结果，比较自适应搜索与被动观看在智能体行为比较中的效果。

表 I 展示了用户在看到由 N-first 和自适应搜索生成方案提供的对比行为摘要后，正确识别优胜策略并按正确顺序排序的结果。我们观察到，调查问卷第一部分中的成对排序任务对参与者来说更容易。看到一份展示两台机器人的行为对比摘要后，使用自适应搜索生成方案时，参与者能够以较高的准确度正确识别优胜智能体；使用 N-first 时，准确度则中等。调查问卷第二部分要求参与者根据成对行为摘要对三台不同的机器人进行排序，任务难度更高。此处，只有将三台 DS-RNN 智能体全部按正确顺序排序，用户的回答才会被标记为正确。与第一部分相比，两种生成方案在此任务中的准确度都有所下降。使用 N-first 方案提供的成对摘要时，参与者的准确度较低，能够正确排列机器人智能体的受访者不到三分之一。然而，相比之下，使用自适应搜索生成的摘要时，用户仍能达到中等准确度，约有三分之二的受访者能够正确

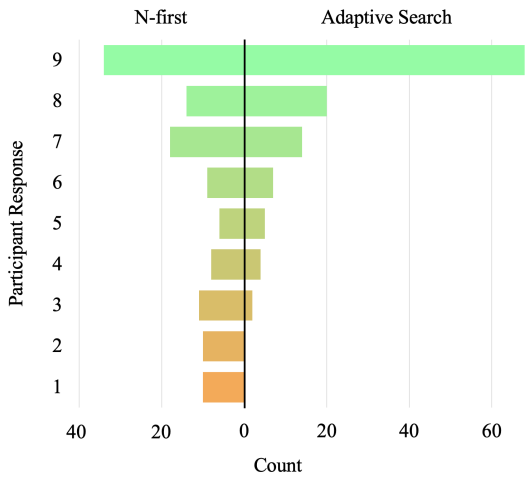

图 3：参与者对调查问卷第一部分任务难度的回答（1 =非常困难， 9 =非常容易）。

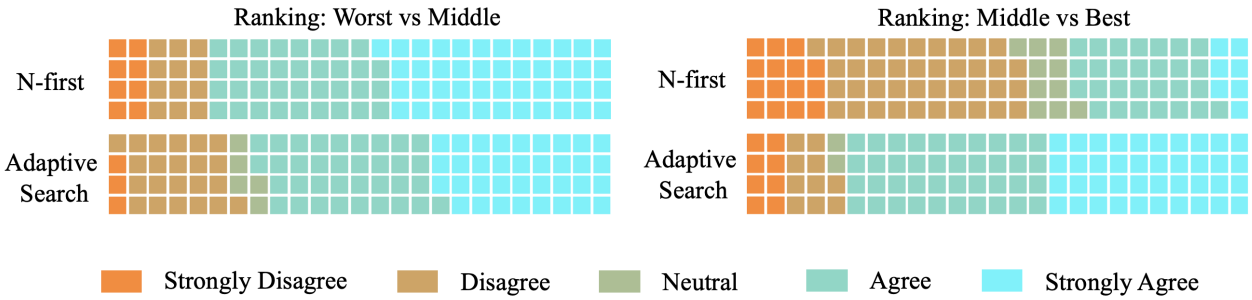

图4：参与者对以下陈述的认同度：很容易区分排名最差的机器人与排名居中的机器人（middlerankedandbestrankedrobot）。橙色表示分歧较大，蓝色表示认同度较高。

对所有智能体进行排名。

图 3 绘出了调查问卷第一部分（成对排序）中参与者对难度的感知结果。参与者观看每份行为摘要并选出他们认为更优的智能体后，会被要求评价根据视频摘要辨别出优胜机器人的难度。我们看到，使用自适应搜索生成的对比摘要，在感知难度方面具有明显优势。多数参与者（占收集到的回答的 56%）表示，根据他们观看的摘要辨别出更优的机器人“非常容易”，且没有人回答“非常困难”。相比之下，使用 N-first 时，只有 28% 的回答将难度描述为“非常容易”，有 8% 的回答表示难以辨别两台机器人之间的优劣。

在调查问卷的第二部分中，我们请参与者判断是否容易区分每一对智能体。在每次全局排名中，用户都需要将智能体从最好到中等再到最差进行排序。因此，每位用户都需要针对上述两对智能体回答这个问题：排名最差与中等的智能体，以及排名中等与最好的智能体。图4展示了全局排名任务中的认同度回答。我们发现，使用 N-first 生成的摘要来区分（参与者认为）排名最差和中等的智能体时，多数参与者（80%）认同摘要提供了足够的对比度，因此很容易辨别出性能更好的策略。然而，非结构化的摘要生成在区分排名中等和最好的策略时效果不佳，只有 35% 的参与者认同这两个机器人很容易区分。这表明，超过某个性能阈值后，仅观察智能体的行为可能不再那么有效，难以判断其真正能力，因为在场景空间的很大一部分范围内，所有智能体的性能都开始趋于相近。使用自适应搜索生成的摘要时，与 N-first 相比，用户区分排名最差和中等策略的能力并未明显提升——约 70% 的受访者选择了“认同”或“强烈认同”。不过，在受访者对区分排名中等和最好的机器人有多困难这一问题上的回答中，自适应搜索与 N-first 之间存在明显差异，具体而言，在自适应搜索与 N-first 之间，区分排名中等和最好的机器人这一任务的难度存在明显差异。使用自适应搜索时，即使智能体性能开始趋同，行为摘要仍能提供对比度较高的轨迹，参与者认同度约为 80%。相比之下，使用 N-first 摘要时，多数参与者不同意这两个机器人很容易区分（如图4右上方图表中大片橙色区域所示）。这一结果凸显了采用高效探索方法生成信息丰富、具有对比性的行为摘要的必要性。随着不同策略的性能逐渐提升，对场景空间进行非结构化搜索，在发现能够体现智能体行为分歧且对比度较高的场景方面，会变得不那么有效。

VI. 结论

随着各种学习算法不断发展，性能优异的自主智能体层出不穷，针对特定任务选择合适的智能体进行部署也变得愈发困难。观察自主智能体行为的被动或非结构化方法，在确定智能体性能差异方面耗时低效；这是因为随着智能体行为趋于最优，高度分歧场景的范围也随之缩小。本文提出了一种用于对比行为摘要生成的自适应搜索方法，旨在以高效的方式找出成对智能体之间存在高度分歧的场景。用户研究的实验结果表明，通过该方法生成的摘要能够以易于辨识的方式，有效地向人类观察者传达表现更优的机器人智能体。

我们要求耦合仿真环境考虑智能体对。这缩小了每次搜索实例化所涉及的场景空间，但代价是比较 $N > 2$ 个智能体时需要多次运行搜索过程。未来，我们希望扩展仿真环境，使其能够在一次搜索实例化中支持所有 N 智能体。通过耦合所有智能体，生成的摘要或许能够更准确地传达全局排名。

参考文献

[1] C. Chen、Y. Liu、S. Kreiss 和 A. Alahi，《人群-机器人交互：基于注意力的深度强化学习实现人群感知机器人导航》，载于 IEEE机器人与自动化国际会议（ICRA），2019 年。[2] S. Liu、P. Chang、W. Liang、N. Chakraborty 和 K. Driggs-Campbell，《基于深度强化学习的机器人群体导航去中心化结构化循环神经网络》，载于 IEEE机器人与自动化国际会议（ICRA），2021 年。[3] Y. F. Chen、M. Everett、M. Liu 和 J. P. How，《基于深度强化学习的社会感知运动规划》，载于 IEEE/RSJ 智能机器人与系统国际会议（IROS），2017 年。[4] Y. F. Chen、M. Liu、M. Everett 和 J. P. How，《基于深度强化学习的去中心化非通信多智能体碰撞规避》，载于 IEEE机器人与自动化国际会议（ICRA），2017 年。[5] D. Amir 和 O. Amir，《亮点摘要：向人们概述智能体行为》，载于自主智能体与多智能体系统国际会议（AAMAS），2018 年。[6] B. Hayes 和 J. A. Shah，《通过自主策略解释提升机器人控制器透明性》，载于 ACM/IEEE 人机交互国际会议（HRI），2017 年。[7] A. Dragan 和 S. Srinivasa，《熟悉机器人运动》，载于 ACM/IEEE 人机交互国际会议（HRI），2014 年。[8] O. Amir、F. Doshi-Velez 和 D. Sarne，《概述智能体策略》，《自主智能体与多智能体系统》，第 33 卷，第 5 期，第 628–644 页，2019 年。[9] Y. Amitai 和 O. Amir，《“我不这么认为”：概述策略分歧以进行智能体比较》，载于 AAAI 人工智能会议，2022 年。[10] J. Gajcin、R. Nair、T. Pedapati、R. Marinescu、E. M. Daly 和 I. Dusparic，《用于比较强化学习智能体偏好的对比解释》，载于 AAAI 交互式机器学习研讨会，2022 年。[11] P. Sequeira 和 M. Gervasio，《可解释强化学习的有趣性要素：理解智能体的能力与局限》，《人工智能》，第 288 卷，第 103367 页，2020 年。[12] S. H. Huang、D. Held、P. Abbeel 和 A. D. Dragan，《使机器人能够传达其目标》，《自主机器人》，第 43 卷，第 2 期，第 309–326 页，2019 年 2 月。[13] S. Greydanus、A. Koul、J. Dodge 和 A. Fern，《可视化并理解 Atari 智能体》，载于国际机器学习会议（ICML），2018 年。[14] O. Z. Khan、P. Poupart 和 J. P. Black，《分解式马尔可夫决策过程的最小充分解释》，载于国际自动规划与调度会议，2009 年。[15] O. Khan、P. Poupart 和 J. Black，《自动生成的马尔可夫决策过程解释》，载于《人工智能应用中的决策理论模型：概念与解决方案》，第 144–163 页，2011 年 1 月。[16] F. Elizalde、E. Sucar、J. Noguez 和 A. Reyes，《基于马尔可夫决策过程生成解释》，载于墨西哥人工智能国际会议，2009 年。[17] S. H. Huang、K. Bhatia、P. Abbeel 和 A. D. Dragan，《通过关键状态建立适当的信任》，载于 IEEE 智能机器人与系统国际会议（IROS），2018 年。[18] M. Koren、S. Alsaif、R. Lee 和 M. J. Kochenderfer，《自动驾驶车辆的自适应压力测试》，载于 IEEE 智能车辆研讨会（IV），2018 年。[19] R. Lee、M. J. Kochenderfer、O. J. Mengshoel、G. P. Brat 和 M. P. Owen，《空中碰撞规避系统的自适应压力测试》，载于 IEEE 数字航空电子系统会议（DASC），2015 年。[20] R. Lee、O. Mengshoel、A. Saksena、R. Gardner、D. Genin、J. Brush 和 M. J. Kochenderfer，《碰撞规避系统的差分自适应压力测试》，载于 AIAA建模与仿真会议，2018 年。[21] A. Corso、P. Du、K. Driggs-Campbell 和 M. J. Kochenderfer，《通过奖励增强进行自适应压力测试，以验证自动驾驶车辆》，载于 IEEE 智能交通系统会议（ITSC），2019 年。

[22] P. Du 和 K. Driggs-Campbell，“使用领域专家所提供的关键状态进行自适应故障搜索”，载于*IEEE*机器人与自动化国际会议（*ICRA*），2021年。[23] T. Miller，“人工智能中的解释：来自社会科学的见解”，人工智能，第267卷，第1–38页，2019年。