



ELSEVIER

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng



Research
Artificial Intelligence—Article

面向物理属性的机器人认知学习

Fuchun Sun ^{a,*}, Wenbing Huang ^b, Yu Luo ^a, Tianying Ji ^a, Huaping Liu ^a, He Liu ^a, Jianwei Zhang ^c

^a Department of Computer Science and Technology, Tsinghua University, Beijing 100190, China

^b Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China

^c Department of Informatics, University of Hamburg, Hamburg 20148, Germany

ARTICLE INFO

Article history:

Received 30 January 2024

Revised 30 September 2024

Accepted 15 October 2024

Available online 9 November 2024

关键词

机器人学习

物理基础

认知学习

摘要

人类通过与环境的持续交互实现认知发展,不断提升感知与行为能力。然而,当前机器人缺乏类人的动作与演化能力,这成为制约机器人智能提升的瓶颈。现有研究多将机器人建模为从观测到动作的单向静态映射系统,忽略了感知与行为的动态过程。本文结合物理属性,提出了一种新的机器人认知学习方法。我们构建了一个理论框架,将机器人概念化为由感知体(P-body)、认知体(C-body)和行为体(B-body)构成的三体物理系统。各子体均具备物理动力学属性,并在闭环交互中运行。重要的是,三种关键交互构成了子体间的联系:C-body依赖于P-body提取的状态,并反过来提供长期奖励以优化P-body的感知策略;此外,C-body通过设定子目标引导B-body的动作,而随后P-body获取的状态又促进C-body的认知动力学学习;最后,B-body遵循C-body生成的子目标,并在P-body的感知状态约束下执行动作,进而进入下一轮交互。这些交互推动各子体的联合演化,实现最优设计。为验证所提方法,本文利用四足机器人D’Kitty,开展了一项导航任务,该机器人配有可移动全局相机。该导航任务要求在感知、规划与D’Kitty的运动间实现精细协同。相较于传统方法,本文的框架能够显著提升任务执行性能。综上,本文结合物理属性,在P-body、C-body与B-body间引入物理交互,建立了一种机器人认知学习的范式,该框架在导航任务中的成功应用也验证了其在提升机器人智能方面的有效性。

© 2024 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. 引言

人类以复杂且高度智能的方式与世界交互[1]。他们用眼睛感知、用手脚行动,而这些均在大脑的控制下协同进行。以穿过杂乱的房间为例,所需特定能力有:感知能力,使个体能够识别门、墙壁及障碍物;行为能力,通过该能力规划导航路线并避免碰撞;认知能力,通过大脑功能向感知和行为系统发出指令。作为反馈系统[2],人类通过感知与行为的相互作用过程不断完善认知能力。当观

测结果偏离预期时,试错会增强其规划能力。

在机器人中模仿这种闭环多功能系统具有一定的挑战性。人工智能和机器人领域研究人员通常将机器人建模为智能体[3–4]。通过模仿学习(IL)[5]或强化学习(RL)优化策略,将观测输入(如摄像头图像和力传感器的触觉反馈)映射为动作,以最大化反馈奖励。虽然这些方法使机器人能够执行实际任务,如导航[6]、运动[7]以及魔方操控[8],但仍存在局限性。首先,现有方法仅实现从观测到动作的静态单向映射,忽略了机器人所具有的感知、行为

* Corresponding author.

E-mail address: fcsun@tsinghua.edu.cn (F. Sun).

2. 相关研究

2.1. 视觉感知

近年来, 得益于深度学习的快速发展, 图像分类[11–13]、目标检测[14–18]和实例分割[19]等多种视觉应用领域展现出显著突破。在此, 我们回顾了目标检测任务的最新进展。该类算法以图像作为输入, 并输出关注目标的类别及其位置信息。传统的目标检测算法大体可分为两类: 两阶段目标检测器与单阶段目标检测器。两者都包含特征提取、候选区域生成、位置坐标回归与类别分类等步骤。两阶段目标检测方法的典型代表是快速区域卷积神经网络 (RCNN) [20], 该算法提出了一种锚点机制, 在每个位置预测目标边界和目标分数, 而非传统的选择性搜索方法。另一种值得一提的例子是特征金字塔网络 (FPN) [21], 它通过融合不同深度的特征层的特征信息, 提升了对不同尺度目标的检测效果。“只需看一次” (YOLO) 检测系统[22]作为一种单阶段目标检测方法, 相较于两阶段方法, 大幅缩短了检测时间, 但精度有所降低。

其他单阶段检测方法还包括单次检测器 (SSD) [23], 该方法通过提取特征图并利用不同尺度和长宽比的先验框进行检测; 而 RetinaNet [24]则引入焦点损失以解决样本分布极度不均衡的问题, 从而提升了单阶段检测器的精度。此外, 还出现了无锚点目标检测方法[25–27], 该类方法利用网络结构特征替代锚点。近期, Transformer 的发展[28–29]也进一步推动了目标检测的研究进展[30–31]。

上述方法均属于被动视觉感知范畴, 大多数现有算法也归类于此。被动视觉感知方法依赖于高质量采集的静态图像作为输入, 但这在实际场景中往往难以获取。具体而言, 在复杂动态环境中, 被动视觉感知方法受遮挡、过曝光、目标移动等噪声干扰, 难以实现目标的精准定位与分类。相比之下, 人类能够通过调整自身位置与视角来获取感兴趣的图像。这一现象启发了主动感知研究的兴起。在以往的研究中, 我们曾探索过配有固定视觉传感器的机器人[32–34], 推动了主动目标检测的发展。然而, 这类方法依赖于一种人工设计的奖励机制, 导致RL训练的智能体需通过最大化回报提升感知性能[35–36]。在我们的架构中, P-body 能够移动并调整至最优配置以提升检测效果, 这属于主动感知方法。此外, 我们的 P-body 同时具备数据驱动与知识驱动的特征。具体来说, P-body 的感知策略不仅在感知奖励的引导下借助RL从数据中习得, 还会受到 C-body 生成的子目标引导的影响。在生成子目标的条件下, P-body 的主动感知过程将利用来自

C-body 的知识, 而这些知识又会通过 B-body 与环境的交互不断更新和优化, 直至达成最终任务目标。

2.2. 强化学习

RL 在人工智能领域取得了令人瞩目的进展, 已在多个领域 (如 Atari、Go 等) 超越了人类表现。RL 将目标导向型智能问题形式化, 即在任务和环境分别实现累计奖励的最大化[37]。RL 算法通常可分为无模型 RL 和基于模型的 RL 两大类。在无模型 RL 中, 智能体训练方法又可进一步分为两种主流方法。

(1) **基于价值的方法**。在这类方法中, 智能体学习状态 (s) - 动作 (α) 价值函数逼近器 $Q_{\theta}(s, \alpha)$, 并据此选择动作, 其中, θ 为神经网络参数。深度 Q 网络 (DQN) [38]是首个成功从高维感知输入中直接学习价值函数的深度学习模型。

(2) **基于策略的方法**。此类方法侧重于直接建模和优化策略 $\pi_{\theta}(\alpha|s)$ 。例如, REINFORCE 算法[39]通过估计回报 (利用蒙特卡罗法收集的完整回合样本) 进行梯度上升, 以此更新策略参数。

更具体地说, 演员-评论家方法介于策略评估与策略优化之间。例如, 深度确定性策略梯度 (DDPG) 智能体算法[40]能够同时学习确定性策略和 Q 函数, 并使二者相互促进优化; 而软演员-评论家 (SAC) 智能体算法[41]则通过引入熵正则化以及其他技巧来学习随机策略。与传统的无模型 RL 方法不同, 我们提出的 Bcent 框架中的 B-body 可以在 C-body 提供的规划提示下进行引导, 从而在与环境交互时实现更优的探索性能。

基于模型的 RL 方法通常在数据效率上表现更佳, 但其渐近性能往往弱于无模型方法。根据已学习的动力学模型不同用途, 基于模型的算法通常可分为三类[42]。

(1) **Dyna-style 算法**。在 Dyna 算法中, 利用所学动力学模型生成虚拟数据, 训练在基于虚拟数据的优化策略与基于真实样本的模型修正之间迭代。基于模型的策略优化 (MBPO) [43]采用神经网络集成来建模动力学过程, 并使用 SAC [41]作为策略优化算法。

(2) **基于时间反向传播的策略搜索**。这类方法利用模型的导数, 并基于模型的解析梯度来改进策略。迭代线性二次高斯 (iLQG) 方法[44]假设奖励函数为二次函数、动力学过程为线性形式, 然后通过动态规划在这些简化参数化条件下推导出控制器。

(3) **射击算法**。这类方法从预设的分布中采样候选动作, 在模型下对采样动作进行评估, 并选择最优动作, 以应对非线性动力学和非凸奖励函数问题。在推荐系统

(RS) [45]中, 智能体从均匀分布中生成候选动作序列, 而交叉熵方法 (CEM) [46]和带有轨迹采样的概率集成 (PETS) [47]则通过迭代调整采样分布。在Bcent框架中, C-body的知识库可被视为对这些学习动力学模型的一种扩展。此外, 该知识库还能为P-body和B-body提供子目标, 引导它们提升训练效率。

3. 方法

如图1所示, 我们的体系结构有三个关键的物理组成部分: P-body、C-body和B-body, 三者在闭环中紧密协作。具体而言, P-body从环境中获取全面信息, 并将其提炼为低维状态, 供C-body使用; 随后, C-body基于接收到的状态向B-body下发具体的子目标; 最终, B-body执行精确动作, 实现C-body所设定的子目标。随着环境不断变化, 这一循环过程持续迭代推进, 直至机器人完成最终目标。

在接下来的章节中, 我们将详细阐述各子体的作用与功能。

3.1. 感知体

P-body的首要目标是在处理从环境中获取的高维观测数据后, 为C-body提供低维表征。P-body的物理实现可以涉及多种传感器, 如用于视觉输入的摄像机、用于触觉反馈的力传感阵列以及用于音频数据采集的录音装置。P-body的一个基本特征在于其动态感知过程。其中, “动态”意味着P-body可根据需求调整自身配置, 提升观测质量。例如, 当P-body被构建为一台可移动摄像机时, 其观测结果与摄像机位置、朝向等配置密切相关。我们将控制此类配置变化的策略称为感知策略。此外, 如前所述, 感知策略还会受到来自C-body的提示 (即子目标) 的影响。

P-body内部的计算流程可用数学公式表达如下:

$$\mathbf{s}_p(t) = f(\mathbf{o}(t)|\mathbf{c}_p(t)) \text{ (state extraction)} \quad (1)$$

$$\mathbf{a}_p(t) = \pi_p(\mathbf{s}_p(t)|\mathbf{s}_c(t+\Delta t)) \text{ (perception policy)} \quad (2)$$

$$\mathbf{c}_p(t+1) = g_p(\mathbf{c}_p(t), \mathbf{a}_p(t)) \text{ (perception dynamics)} \quad (3)$$

式中, \mathbf{a}_p 表示P-body策略的动作。上述方程中, t 表示时间步 ($t=0, \dots, T$); T 为任务时域; $\mathbf{o}(t)$ 表示来自环境的高维多模态观测 (如图像、深度和音频); $\mathbf{s}_p(t)$ 是一个低维抽象状态向量, 用于表征感知状态; $\mathbf{c}_p(t)$ 是感知状态, 其定义为传感器参数与配置的组合, 如传感器位置和内、外部参数, 这些参数由策略 $\pi_p(\mathbf{s}_p(t)|\mathbf{s}_c(t+\Delta t))$ 确定, 其中, $\mathbf{s}_c(t+$

Δt)表示C-body在延迟 Δt 后生成的未来子目标; f 是嵌入函数, 基于配置 $\mathbf{c}_p(t)$ 将高维观测向量 $\mathbf{o}(t)$ 映射为低维抽象状态向量 $\mathbf{s}_p(t)$ 。随后, P-body会基于当前配置采取的动作, 转移到一个新的配置 $\mathbf{c}_p(t+1)$, 该转移由转移概率函数 $g_p(\mathbf{c}_p(t), \mathbf{a}_p(t))$ 决定。该公式化过程的基本原理源于马尔可夫决策过程 (MDP)。

3.1.1. 通过P-C交互实现的无模型感知策略学习

在MDP的框架下, 策略 π_p 训练通常包含两种主要方法: 基于模型的方法[48]和无模型的方法 (如近端策略优化 (PPO) [49]和信赖域策略优化 (TRPO) [50])。对于P-body, 我们选择无模型策略, 这是因为学习观测 $\mathbf{o}(t)$ 的转移概率可能比较复杂且成本高。一个重要问题在于奖励信号 r_p 应如何提供? 一种选择方式是将 r_p 设定为感知状态 $\mathbf{s}_p(t)$ 对真实状态 $\mathbf{s}^*(t)$ 估计精度的衡量标准, 但该标准在许多场景中可能并不适用。

3.1.2. P-C交互

机器人通常执行的是长期目标 (如到达指定位置), 所以必须确保在整个运动轨迹中, 尤其是在目标点处, 将感知误差降至最低。此外, 在Bcent框架下, C-body会向P-body传递子目标, 从而为其提供引导并提升感知性能。与传统的主动感知方法[50–51]相比, 我们的P-C交互有以下显著优势。

- **高效性:** P-body对感知动态的显式建模使得C-body的规划信息能够通过动态模型辅助P-body运动, 从而增强感知的可解释性和有效性。相比之下, 基于RL的主动感知方法通常依赖最大化感知奖励引导传感器运动, 这种方法探索成本较高。

- **灵活性:** 我们的框架[见公式 (1) ~ (3)]具有普适性, 能够兼容多种策略实现方式, 包括预设规则、控制方法以及RL。

3.2. 认知体

C-body在我们的模型中发挥着核心作用, 类似于人类大脑在协调和指挥各种身体机能方面的职责。更具体而言, C-body的主要任务是基于另外两个子体与环境交互所积累的经验, 提取和更新知识。作为反馈, C-body会向这两个子体提供引导与指令。

C-body内部的计算流程定义如下:

$$\mathbf{s}_c(0) = f(\mathbf{o}(0)|\mathbf{c}_p(0)) \text{ (initial state extraction)} \quad (4)$$

$$\mathbf{a}_c(t) = \pi_c(\mathbf{s}_c(t)|\mathbf{k}_c) \text{ (cognition policy)} \quad (5)$$

$$\mathbf{s}_c(t+\Delta t) = g_c(\mathbf{s}_c(t), \mathbf{a}_c(t)|\mathbf{k}_c) \text{ (cognition dynamics)} \quad (6)$$

在上述公式中， $s_c(t)$ 表示来自C-body的规划指令，其作用是引导P-body和B-body的探索； $a_c(t)$ 表示C-body在每个规划时刻的动作，该动作会生成下一步的规划指令。此外， $\Delta t \in \mathcal{N}^+$ （正整数）表示C-body的更新间隔，其时间取值可以是 $t = 0, \Delta t, \dots, K\Delta t$ （ K 个时间步）； π_c 和 g_c 分别表示C-body中的认知转移概率函数；而 k_c 表示C-body的知识库。

需要注意的是， $s_c(t)$ 可以以多种形式引导P-body和B-body的探索与执行，它代表由认知动力学生成的计划子目标或指令。与采用单位更新间隔（即 $\Delta t = 1$ ）的P-body（以及后续的B-body）不同，这里我们允许 $\Delta t > 1$ ，偏离了默认设定。此修改源于C-body更侧重于长期、粗粒度的目标，并且在更大的时间尺度上演化。

值得强调的是，认知策略和认知动态均依赖于C-body的知识储备（记为 k_c ）。本文虽然将 k_c 定义为认知

转移概率函数 g_c 的参数，但它同样可以表现为记忆缓冲区、知识数据库、知识图谱或其他信息性结构。 k_c 知识的引入使得C-body能够模拟人脑的功能，包括知识的运用与更新两大维度。

我们在图2和图3中进一步对规划指令进行说明。具体来说，在图3的右侧，D’Kitty机器人从起点出发，目标是到达终点，中间需绕过一个由橙色区域表示的障碍物。在知识库中，该任务可分解为一个三步技能序列：“移动至障碍物”“绕过障碍物”和“移动至目标”，每个技能均配备了相应的动作原语与参数空间。

更具体而言，在第一个技能“移动至障碍物”中，动作原语包括：通过检测障碍物的距离和形状，接近转折点1，最终到达障碍物前方。需要注意的是，障碍物为矮墙或坑洞，机器人无法通过，因此必须绕过障碍物才能到达目标位置。在到达第一个转折点后，D’Kitty机器人进入第二个

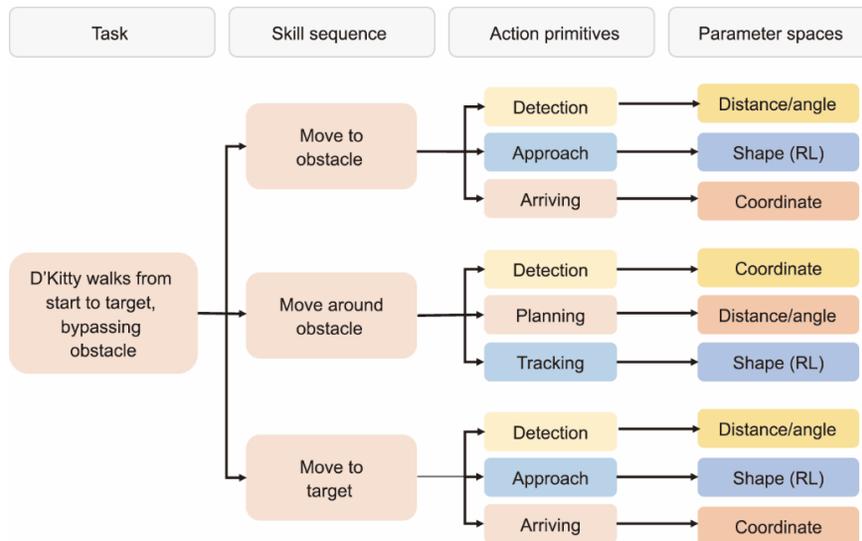


图2. 知识库的设计与构建。在初始化阶段，我们通过从人类/专家演示中提取技能，或采用简单的手工编码来建立知识库。为区分技能与动作原语的来源，我们使用橙色代表由P-body提供的动作（如检测），粉色代表C-body作出的决策（如确定转折点），蓝色则代表由B-body执行的行为（如接近与跟踪）。

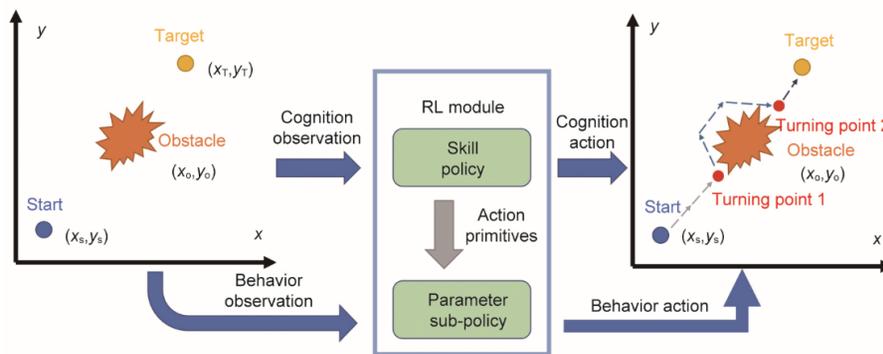


图3. 知识的应用与部署。我们展示了技能序列的完整过程，包括“移动至障碍物”“绕过障碍物”和“移动至目标”，每个技能序列都包含其固有的动作原语。通过子目标生成，C-body的技能可以适应不同场景。 (x, y) 表示任务中的位置坐标轴， (x_s, y_s) 表示起点坐标， (x_t, y_t) 表示目标点坐标， (x_o, y_o) 表示障碍物坐标， (x_{T1}, y_{T1}) 表示目标坐标。

技能“绕过障碍物”，该过程包括重新检测障碍物获取坐标，规划绕行路径并实时追踪规划轨迹；最后，D’Kitty执行第三个技能“移动至目标”，检测目标点并移动至目标位置。通过依次执行上述三项技能，D’Kitty能够自主完成识别障碍与目标点的双重检测，基于检测结果规划合理路线，最终依次执行动作以实现各个子目标，在这一过程中，C-body作为技能策略用于生成规划指令，B-body作为参数子策略用于最终到达目标点。

3.2.1. 知识运用

C-body中知识利用的目标是基于任务目标 g 生成相应的子目标 $\{s_c(k\Delta t)\}_{k=1}^{\lfloor \frac{T}{\Delta t} \rfloor}$ 。其中， $\lfloor x \rfloor$ 表示不大于 x 的最大整数，且 k 为正整数。本文通过基于模型的RL实现这一过程。我们聚焦目标导向型任务，奖励函数 r_c 通过衡量在时刻 $T(T=K\Delta t)$ 时的感知状态 $s_p(T)$ 与目标状态 s^* 之间的接近程度来计算，具体定义为 $r_c(s^*, s_p(T)) = -d(s^*, s_p(T))$ ，其中， $d(\cdot, \cdot)$ 表示感知状态空间中的距离度量函数（如本文采用欧几里得范数）。我们选择在C-body中进行基于模型的策略学习主要有两点原因：

- 认知动力学在简洁且低维的状态空间内运行，使其更适合在这一抽象领域中学习模型 $g_c(s_c(t), a_c(t)|k_c)$ ；

- 学习认知动力学是至关重要的，因为正如后文所述，C-body需要为P-body和B-body提供子目标，这些子目标依赖于认知模型的输出。

研究已经提出了多种基于模型的RL方法，我们选择了iLQR方法[52]，该方法已被广泛应用于机器人控制领域。子目标的生成涉及一个最优控制问题的构建：

$$\{s_c\}_{k=1}^{\lfloor \frac{T}{\Delta t} \rfloor} = \arg \max_{s_c} \sum_k r_c(s^*, s_p(T)) \quad (7)$$

$$\text{s.t. } s_c((k+1)\Delta t) = g_c(s_c(k\Delta t), s_p(k\Delta t)|k_c) \quad (8)$$

$$s_c(0) = s_p(0), \quad s_c(T) = s^* \quad (9)$$

该问题的解产生了一个子目标序列 $\{s_c(0), s_c(\Delta t), s_c(T)\}$ ，并进一步作为引导信息传递给P-body和B-body。

3.2.2. 知识更新

在Bcent框架中，知识的初始化可通过两种方式实现：从先验知识中学习，或通过随机初始化学习。当存在先验知识时，可以通过函数逼近或硬编码的方式将初始化知识嵌入知识库中。否则，对于没有先验知识的通用场景，可以通过随机超参数来对知识库进行初始化。接下来出现一个问题：我们如何更新知识 k_c ？更具体来说，是如何更新认知模型 $g_c(s_c(k\Delta t), s_p(k\Delta t)|k_c)$ 的参数？本质上，

$s_c(t+\Delta t)$ 表示由C-body预测得到的期望状态，它应当尽可能逼近实际反馈的 $s_p(t+\Delta t)$ 。换句话说，认知模型的学习目标就是通过最小化均方误差来减少残差 $\|s_c(t+\Delta t) - s_p(t+\Delta t)\|_2^2$ 。考虑到 $s_c(t+\Delta t) = g_c(s_c(t), a_c(t)|k_c)$ ，我们建议在模型学习过程中，将 $a_c(t)$ 实例化为实际状态 $s_p(t)$ 。实验结果表明，这种实例化方式能够显著加速学习过程。总体而言，模型学习问题可以表述如下：

$$k_c = \arg \min_{k_c} \sum_{k=1}^{\lfloor \frac{T}{\Delta t} \rfloor} \|s_p((k+1)\Delta t) - g_c(s_c(k\Delta t), s_p(k\Delta t)|k_c)\|_2^2 \quad (10)$$

式中， $\{s_c(k\Delta t)\}_{k=1}^{\lfloor \frac{T}{\Delta t} \rfloor}$ 来自于公式(7)中的知识利用。

3.2.3. P-C交互

一旦习得最优的认知转移概率模型 g_c ，它能够促进P-body与B-body的策略学习。子目标被整合到感知策略[见公式(2)]以及P-body的长期奖励 r_p 中，用于主动感知。

3.2.4. C-B交互

对于B-body，子目标 $\{s_c(k\Delta t)\}_{k=1}^{\lfloor \frac{T}{\Delta t} \rfloor}$ 将被用作期望状态，引导行为策略的优化，具体内容将在下一小节中介绍。

在前述的P-C和B-C交互过程中，C-body会利用历史信息与先验知识不断更新认知体 k_c 。这一特性凸显了本文方法相较于以往研究的优势。

- **类人认知：**C-body能够模拟环境特征并优化规划过程中的知识体系，使其更贴近于人类的认知发展轨迹。

- **样本效率：**引入 k_c 能够利用历史信息与先验知识，从而提升认知训练的样本效率。

- **协同性：**C-body向P-body与B-body提供规划信息，使二者能够以更高的精度与效率执行各自的动作。

3.3. 行为体

B-body根据从C-body获取的子目标 $\{s_c(k\Delta t)\}_{k=1}^{\lfloor \frac{T}{\Delta t} \rfloor}$ 来统筹其动作。为便于说明，我们将概述实现子目标 $s_c(\Delta t)$ 的过程，其余子目标的计算流程遵循类似模式。我们采用了如下的MDP，其中， $t = 0, 1, \dots, \Delta t$ 。

$$s_p(t) = f(o(t)|c_p(t)) \quad (\text{state extraction}) \quad (11)$$

$$a_b(t) = \pi_b(s_p(t)|s_c(t+\Delta t)) \quad (\text{behavior policy}) \quad (12)$$

$$o(t+1) = g_b(o(t), a_b(t)) \quad (\text{behavior dynamics}) \quad (13)$$

为了清晰表述，我们沿用了P-body的状态提取阶段。除此之外， $a_b(t)$ 是由B-body的行为策略生成的动作，将其应用于环境，从而产生下一个观测值 $o(t+1)$ 。 π_b 是行为体

的策略函数，而 g_b 是行为动力学模型。

B-body 由两个主要目标驱动：首先，它旨在准确追踪 C-body 提供的子目标。我们将这种奖励称为目标导向奖励 $r_{c1} = -d(s_p(k\Delta t), s_c(k\Delta t))$ ，用于在每个第 k 阶段衡量 $s_p(k\Delta t)$ 与 $s_c(k\Delta t)$ 之间的距离。其次，除了到达目标位置之外，B-body 还需要满足其他条件（例如，在导航任务中，机器人必须学会站立与避障，才能最终到达目标）。因此，我们额外引入一个从环境中获得的奖励 r_{c2} 。综合两者，我们计算了奖励 $r_c = \lambda_b r_{c1} + (1 - \lambda_b) r_{c2}$ ，其中， λ_b 为权重参数，用于平衡这两个奖励的重要性。为促进行为策略 π_b 的学习，我们倾向于采用无模型 RL 方法。因行为动力学通常较为复杂，这一方法选择受其影响，难以被精确建模。然而，在动力学较为简单或已知的情境下，采用基于模型的策略学习则可能带来更高的探索效率。

在接收到 C-body 提供的子目标后，B-body 将原始的长期复杂任务拆解为短期、简单的顺序子任务。在每个子任务中，B-body 通过执行动作实现给定子目标。当子任务完成后，B-body 会将实际探索结果反馈给 C-body，以更新知识库。

相较于传统的 RL 方法，B-body 具有以下优势。

- **可分解性**：通过利用规划信息，B-body 能够将原始任务分解为一系列更简单的子任务。每个子任务的成功完成都有利于总体目标的实现。
- **交互性**：与环境交互后，B-body 计算得到的交互残差能够有效促进 C-body 知识体系的更新。
- **效率提升**：将任务分解为顺序子任务，使 B-body 能够提高交互效率，从而高效、流畅地执行任务。

4. 实验

所有实验均使用一款名为 D'Kitty 的四足机器人。该机器人的主要目标是导航至预设的目标位置。以下为我们实验框架的实现说明。

- **顶置摄像头充当 P-body**，负责追踪 D'Kitty 的位置。为提升追踪能力，摄像头能够在三维（3D）方向上自由移动，即上下、左右和前后。
- **D'Kitty 具有 12 个运动自由度**，充当 B-body 角色，其任务包括学习如何站立和向前移动，这一过程依赖于内部传感器检测到的腿部角度信息，以及由 P-body 提供的二维（2D）位置坐标。
- **规划器充当 C-body**，负责生成通向目标位置的虚拟子目标轨迹，该轨迹用来引导 P-body 和 B-body 的运动。

后续章节将展示一系列实验，以证明在导航任务背景

下三者交互的必要性。更具体地说，我们先对感知任务进行评估，随后再对导航任务展开研究。

在进行这些实验时，Bcent 框架的符号实例化定义如下： $c_p(t)$ 表示相机的内、外参数， $o(t)$ 表示从相机获取的俯视图像， $s_p(t)$ 表示从图像中提取出的 D'Kitty 的位置与速度以及由内部传感器检测到的 D'Kitty 的角度和角速度， $a_p(t)$ 为相机配置的控制策略， k_c 表示 D'Kitty 与任务的先验知识结构参数， $s_c(t)$ 为 D'Kitty 在下一个 Δt 步的期望位置与姿态， $a_c(t)$ 为 D'Kitty 位置和姿态的认知策略，而 $a_b(t)$ 则表示 D'Kitty 的控制输入（如关节力矩）。

尽管实验场景看似简单，仅包含一个障碍物，但其实非常复杂。传统基于控制的方法难以满足任务需求，主要面临两个挑战。第一个挑战在于机器人需要感知环境，如识别障碍物的位置和大小。如果缺少感知模块（即 P-body），仅依赖低层次传感将难以实现避障功能。第二个挑战是因为 D'Kitty 机器人配置特殊，配备了四足和 12 个运动自由度。在本任务中，机器人必须学会如何站立和移动，而传统的无学习方法难以实现四条腿的协同控制。

总之，上述感知能力和行为能力更适合通过基于学习的方法来实现，而非依赖传统基于控制的方法。

4.1. 感知评估

在实验中，摄像头能够在三维空间中移动，以此全面评估主动感知在机器人任务中的重要性，并在理想环境下充分探索主动感知的潜力。在时刻 t ，P-body（由摄像头表示）基于摄像头的内、外部参数 $c_p(t)$ ，接收到环境中的俯视图像，记为 $I(t) = (o(t)|c_p(t))$ 。在此输入的基础上，P-body 利用目标检测神经网络来检测运动中 D'Kitty 的 2D 状态 $s_p(t)$ ，如公式（1）中的 $s_p(t) = f(I(t))$ 所述。与以往的目标检测方法相比，我们的 P-body 具有两个显著优势。

- **主动感知**：P-body 具备移动能力，能够确保 D'Kitty 始终保持在其视野范围内。这一能力通过其感知策略 π_p 得以实现。
- **P-C 交互**：C-body 设想的轨迹为 P-body 提供了有效引导，使其能够更好地跟踪 D'Kitty 的行为，从而提升感知策略 $a_p(t) = \pi_p(s_p(t)|s_c(t+\Delta t))$ ，其中， $s_c(t+\Delta t)$ 代表公式（4）中在 $t+\Delta t$ 时的子目标。

在本节中，我们将展示这些特定设计的优势。实验流程的概览见图 4 和图 5。

我们为训练编制了两类不同的数据集：离线数据集和在线数据集。离线数据集通过结合 DeepMind Control Suite 和 MuJoCo 以及 Robel 环境构建而成，其实现方式是在实验装置中加入一台固定摄像头。P-body 的执行细节见

图6。在离线数据集上完成训练后，建立被动感知模块，并在随后的主动感知模块训练中保持不变。在线数据集则是通过移动摄像头与环境的实时交互生成。需要注意的是，在线图像无需存储，因为主动感知模型会动态地与环境进行交互。所有图像的输入尺寸均统一为480×480。

4.1.1. 评估指标

本研究的评估采用五个关键指标：精确率、召回率、 F_1 分数、 F_2 分数以及平均精度均值（mAP）。在计算机视觉研究领域，使用mAP来评估目标检测模型的性能已成为一种常见做法。

4.1.2. 结果与分析

我们在表1 [33,52]中呈现了量化结果，并将最佳结果以加粗字体标出。在该表中，标记为“fixed” [33]的方法表示相机固定不动的情景。“fixed”方法表现最差，因为D’Kitty经常移出相机视野。术语“random”表示相机通过初始化三个向量来确定沿 x 、 y 和 z 轴的移动范围，实现摄像头移动。“active”方法[52]指主动感知，即感知智能体通过RL方法在人工设计的奖励下进行训练，以调整传感器的配置和位置。“policy”表示本文方法的一种变体，该变体没有子目标的引导，而“policy+sub-goal”则代表本文提出的完整实施方案。

实验结果表明，根据最基本且严格的评估指标（即mAP），我们的方法优于其他三种基线方法。这一结果证明了我们方法的优越性。值得注意的是，“policy”方法的

召回率高于我们的方法。我们推测，由于缺乏子目标的引导，“policy”方法在初期往往收集过多样本。与我们的方法相比，这可能导致其召回率较高，但精确率有所下降。

4.2. 运动与导航评估

本小节的实验对D’Kitty在运动（即站立与行走等任务）和导航（即到达指定目标位置的策略）方面进行了全面验证。

4.2.1. 观测空间与子目标空间

在这些评估中，D’Kitty在某一既定时间步 t 的完整观测空间 $\mathbf{o}(t)$ 内涵盖了丰富的变量。其中包括D’Kitty的当前位置与速度，以及其朝向与目标位置之间的对齐程度。此外，该空间还包含12个关节角度及其对应的角速度。进一步来说，状态向量 $\mathbf{s}_p(t)$ 和子目标 $\mathbf{s}_c(t)$ 与D’Kitty在环境中的2D位置紧密相关。

4.2.2. 预测性规划

在接收到初始状态与目标位置后，C-body会生成一条预测性的状态轨迹，具体表示为 $\{\mathbf{s}_c(0), \mathbf{s}_c(\Delta t), \mathbf{s}_c(T)\}$ 。随后，根据每一个预测状态 $\mathbf{s}_c(t)$ ，B-body的行为策略会执行一系列动作，其目标是在 Δt 时间步内，实现随后的子目标 $\mathbf{s}_c(t+\Delta t)$ 。

4.2.3. 知识更新

在经历一个由 Δt 时间步组成的交互窗口后，计算的积累残差误差会传递给C-body，其形式为 $\mathbf{s}_c(t+\Delta t)-\mathbf{s}_p(t+$

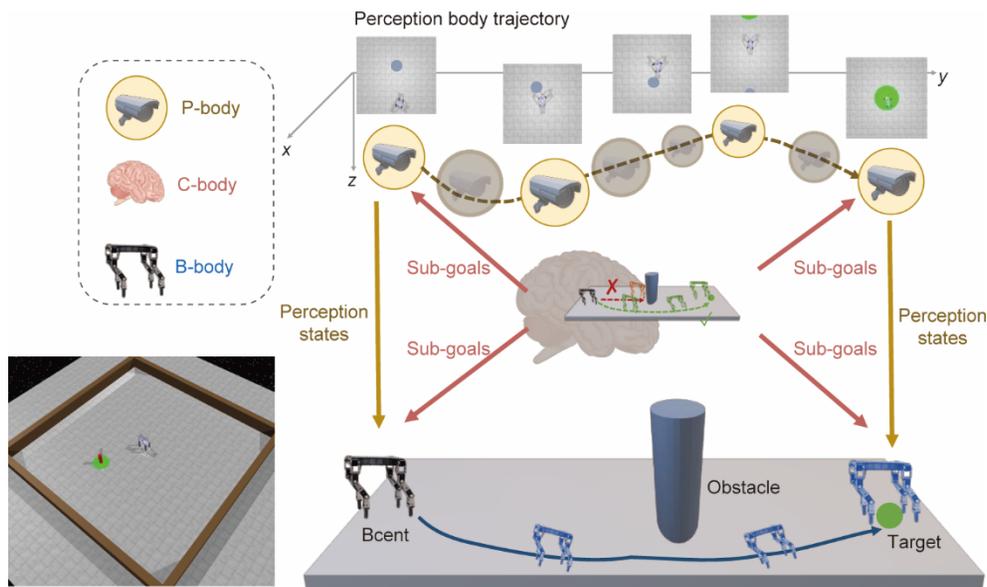


图4. D’Kitty运动与导航任务的实现示意图。图中，P-body为主动感知相机，C-body执行路径导航，B-body则为实际移动的D’Kitty机器人。随着D’Kitty前进，相机会主动跟随其运动。同时，C-body根据相机的感知信息和D’Kitty的状态进行决策。通过为B-body和P-body生成子目标，C-body能更有效地引导D’Kitty到达目标点，并避开行进路径上的障碍物。

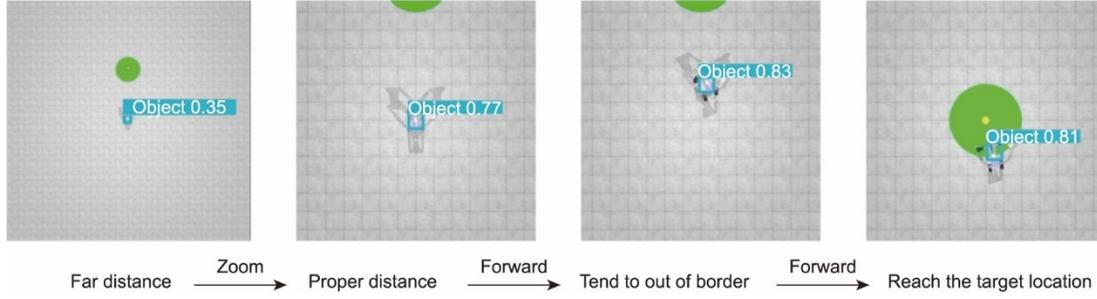


图5. P-body的工作示意图。P-body（即相机）最初位于远处，以便捕捉环境的全景。随后，部署在P-body上的感知模块会搜索并定位D’Kitty的大致位置。然而，由于距离较远，感知模块对结果的置信度较低（如图所示为35%的概率）。接着，P-body会拉近视角，从而获得D’Kitty更精确的位置。同时，D’Kitty正接近其目标位置。若D’Kitty超出视野范围，P-body会紧密跟随D’Kitty，对其位置进行检测与更新。

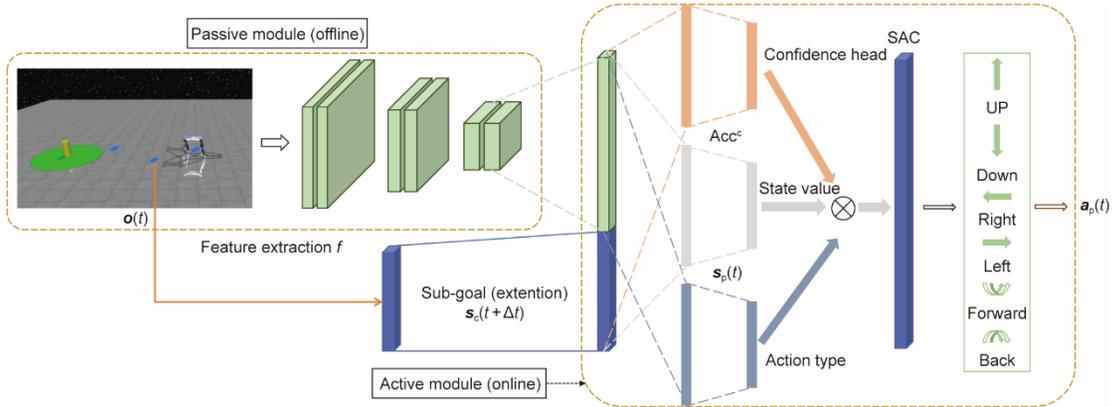


图6. P-body的网络结构。该结构主要分为两部分：被动模块与主动模块。被动模块充当特征提取器，用于生成置信度头和状态值；主动模块则利用被动模块的输出及子目标来预测P-body的动作。Acc^c：特征提取输出的置信度头。

表1 本实验的量化结果

Method	Precision	Recall	F_1 scores	F_2 scores	MAP-0.5
Fixed [33]	0.164	0.594	0.257	0.389	0.160
Random	0.320	0.597	0.417	0.509	0.311
Active [52]	0.362	0.605	0.428	0.542	0.342
Policy	0.381	0.70	0.494	0.599	0.398
Policy + sub-goal	0.529	0.631	0.576	0.608	0.597

Δt)。该信息作为触发因素，通过公式 (10) 用于更新其知识参数 k_c 。同时，B-body 也会对其策略进行优化以减小该误差，从而在后续训练迭代中提升性能。

本质上，这一协同交互循环将感知、规划与动作无缝融合，不仅促进了C-body 获取更优知识，同时也使B-body 能够微调其策略，以实现更优的执行效果。

4.2.4. 实现细节

在C-body 内部，认知转移概率模型 g_c 通过多层感知器 (MLP) 实现。在每个训练回合中，C-body 会提供子目标，引导B-body 与环境的交互。从这些交互中获得的经验被存储在缓存 \mathcal{L} 中，用于C-body 的自监督训练。为克服P-body 与B-body 同时训练所引起的非平稳性问题，采用了“all-in all-out”的训练策略。该策略通过循环处理

缓存，将经验数据划分为训练集和测试集，用于模型的迭代训练。此外，图7展示了C-body 与B-body 的实现细节及其交互过程。

对于B-body，该任务被划分为一系列带有目标条件的子任务，每个子任务间存在间隔 Δt 。每个子任务都会结合来自C-body 的子目标，以便与环境进行交互。其行为策略通过PPO 算法进行训练，并采用三层MLP 作为策略逼近器。所有超参数都列于表2中。

4.3. 评估结果

为评估训练与测试阶段，D’Kitty 机器人的初始位置固定在原点(0,0)，并保持标准站姿；目标位置则在 x - y 坐标空间内随机选取。每完成50 个训练回合后，会进行10 次测试，并计算平均累积奖励与成功率用于展示。

4.3.1. 奖励函数与成功指标

奖励 (r) 函数的定义如下，可参见文献[54]：

$$r(\mathbf{s}, \mathbf{a}) = r_{\text{upright}} - 4d_{t,\text{goal}} + 2h_{t,\text{goal}} + r_{\text{bonus_small}} + r_{\text{bonu_big}} \quad (14)$$

$$r_{\text{bonus_small}} = 5(d_{t,\text{goal}} < 0.5 \text{ or } h_{t,\text{goal}} > \cos 25^\circ) \quad (15)$$

$$r_{\text{bonus_big}} = 10(d_{t,\text{goal}} < 0.5 \text{ or } h_{t,\text{goal}} > \cos 25^\circ) \quad (16)$$

式中， r_{upright} 是鼓励机器人保持直立的奖励，当 $\alpha_{\text{upright}} =$

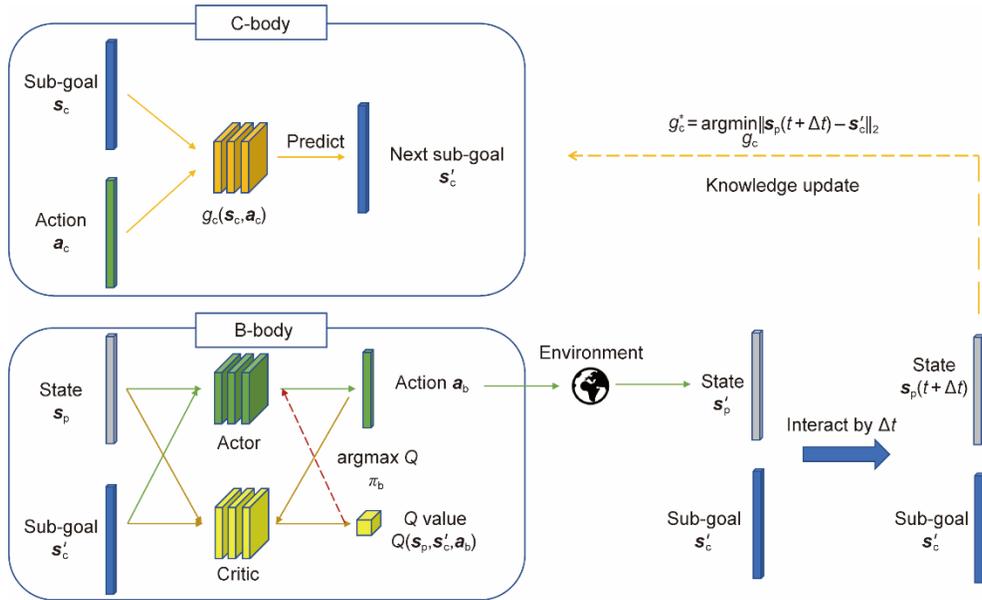


图7. C-body与B-body的网络结构。在该框架中，C-body根据当前状态 s_c 和动作 a_c 生成下一个子目标 s'_c ；B-body则通过执行动作与环境进行交互，并在时间间隔 Δt 后追踪下一个子目标。到达状态 $s_p(t+\Delta t)$ 被视为预测子目标 s'_c 的监督信号，用于训练认知转移概率模型 g_c 。 Q ：B-body中策略的状态-动作价值函数。

表2 B-body、C-body与P-body的超参数

Different body	Hyper-parameter	Value
Behavior body	Q hidden sizes	(128, 128)
	Policy hidden sizes	(128, 128)
	Q learning rate	3×10^{-4}
	Policy learning rate	3×10^{-4}
	Discount factor γ	0.99
	Clip	0.2
	Generalized advantage estimation factor τ_{gae}	0.95
	Samples per updating	1024
	Batch size	64
	Cognition body	Model hidden sizes
Horizon of MPC		3
Weight factor		0.1
Holdout ratio		None
Cache size		3000
Batch size		64
Model learning rate		3×10^{-4}
Perception body	Camera resolution	(480, 480)
	Model hidden sizes	(512, 512)

$\cos 15^\circ$ 时， $r_{\text{upright}} = 1$ ，否则 $r_{\text{upright}} = -500$ 。其中， α_{upright} 表示机体法线与地面之间夹角的余弦值； $d_{t,\text{goal}}$ 表示当前状态到目标的距离； $h_{t,\text{goal}}$ 表示机器人朝向与目标连线之间夹角的余弦值； $r_{\text{bonus-small}}$ 和 $r_{\text{bonus-big}}$ 则是在达到期望目标时的奖励。由该奖励函数可见，该任务要求模拟的D’Kitty既能到达目标，又能保持平衡。

成功的判定指标是：在训练回合的最后一步中，目标

距离是否在某一阈值范围内，并且D’Kitty是否保持足够直立：

$$\mathbb{E}_{\tau \sim \pi} \left[\mathbf{1}(d_{T,\text{goal}}^{(\tau)} < 0.5) \times \mathbf{1}(h_{T,\text{goal}}^{(\tau)} > \cos 25^\circ) \right] \quad (17)$$

式中， \mathbb{E} 表示期望值； τ 是单一任务的状态-动作轨迹； $d_{T,\text{goal}}^{(\tau)}$ 表示任务完成时D’Kitty当前位置与目标之间的距离； $h_{T,\text{goal}}^{(\tau)}$ 表示任务完成时，最终角度的余弦值。

4.3.2. 结果与分析

成功率与奖励情况的评估结果分别在图8和图9中展示。此外，我们还在附录A的视频S1中提供了动态运动过程可视化结果。本文对比了两种方法的变体：一种采用真实状态，记为CB-body（由C-body和B-body组成）；另一种采用P-body预测的状态，记为PCB-body（由P-body、C-body和B-body组成）。同时还包括利用PPO [49]、SAC [41]或时序差分模型（TDM） [53]等多种基线方法，单独控制B-body。实验结果表明，CB在控制性能和成功率方面表现更佳，该方法展现了对局部最优问题的鲁棒性。而PPO由于探索能力受限，仅收敛到较低的奖励值，导致机器人只学会站立并朝向目标，而未能掌握行走动作。

与CB-body相比，PCB-body的性能略有下降。这主要源于P-body所预测的状态与真实状态存在一定差异，导致PCB-body的性能出现轻微退化。在缺乏真实感知信息（如D’Kitty机器人的位置与速度以及目标和障碍物的坐标）的条件下，P-body能够通过图像提取相关信息，随后C-body和B-body接收来自P-body的感知信息并驱动机

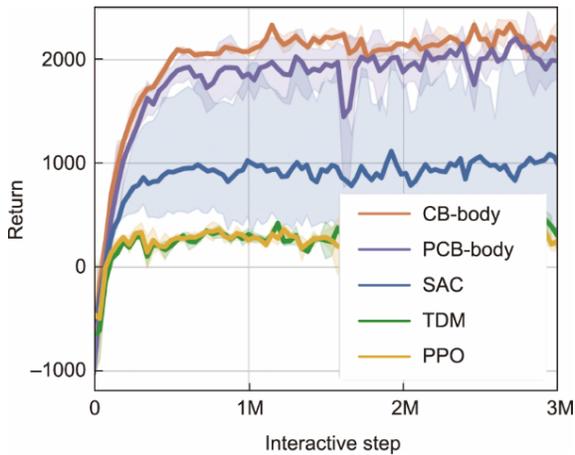


图8. 运动与导航任务的奖励。CB-body: C-body与B-body; PCB-body: P-body、C-body与B-body; TDM: 时序差分模型; M: 百万。

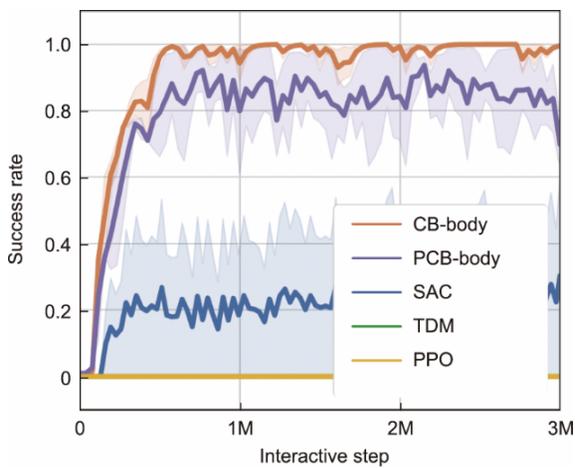


图9. D’Kitty运动与导航任务的成功率与奖励性能对比。我们将所提出的闭环框架（配备P-body、C-body与B-body）与其消融变体（仅使用B-body的PPO、SAC或TDM，以及使用真实状态的CB-body）进行比较。

机器人到达目标位置。总体而言，尽管PCB-body（由P-body、C-body与B-body组成）面对的环境较CB-body更具挑战性，但性能下降幅度极小，展现出在Bcent框架下三体固有协作的有效性。此外，PCB-body的表现显著优于基线方法，验证了三体交互在提升控制性能与任务成功率方面的作用。

我们还在表3中报道了训练耗时和计算基础设施，以展示计算效率。从结果可以看出，即使在三体的同步训练下，Bcent也能够常规计算环境中高效完成训练。

表3 计算基础设施及计算耗时

Item	CPU	GPU	Bcent (h)
P-body	AMD EPYC™ 7763	NVIDIA GeForce RTX 3090	6.84
C-body + B-body	AMD EPYC™ 7763	NVIDIA GeForce RTX 3090	18.74
P-body + C-body + B-body	AMD EPYC™ 7763	NVIDIA GeForce RTX 3090	27.65

CPU: central processing unit; GPU: graphics processing unit.

5. 讨论与结论

本文提出了一种创新的决策框架，该框架利用了P-body、C-body与B-body的协同发展实现决策优化。我们的框架通过引入C-body超越了传统智能体范式，更精准地模拟人类认知过程和动态决策过程。三体间的物理协同作用有助于整体性能的提升。通过对D’Kitty的导航与运动进行广泛实验，我们验证了所提出的各组件相较于现有方法及其他变体的必要性和有效性。

本研究为认知机器人领域的未来多项探索奠定了基础。

- **多认知体协作：**受到章鱼等生物的启发，它们的多个大脑区域分别承担不同的功能，因此探索多认知体如何协同实现更优性能，是一个极具吸引力的研究方向。

- **动态环境中的人机协作：**随着机器人在各种任务中越来越多地与人类及其他机器人协作，迫切需要研究机器人如何在动态环境中与人类并肩作业时有效应对动态环境。

- **云端分布式感知与学习：**云机器人概念正逐渐兴起。将分布式局部感知与云信息全局共享相结合，有望革新机器人的能力。

- **非马尔可夫决策：**传统决策过程通常具有马尔可夫性，智能体缺乏记忆过去轨迹的能力。采用非马尔可夫决策过程，可以让机器人充分利用完整的历史和经验，从而实现更全面、更有信息支撑的决策。

总之，本文框架在认知机器人领域引入了开创性的视角，强调不同认知体之间的协作。该研究为未来的研究与创新开辟了诸多令人振奋的方向。我们提出的Bcent框架具有较强的通用性，可为三体选择特定的状态和动作空间，从而应用于各种类型的机器人或任务。在人机交互场景中，潜在挑战主要体现在两个方面：在感知方面，为了更好地与人类交流，机器人需要理解人类意图，并在人机互动中了解人类的规划；在行为方面，机器人应能够从人类示范中学习，并模仿人类技能，以提升自身的行为能力。在Bcent框架中，除了感知与行为外，我们还利用认知机制来更新知识，从而实现持续学习和终身学习的能力，这有助于Bcent框架应对复杂的人机交互场景。

CRedit authorship contribution statement

Fuchun Sun: Writing-review & editing, Supervision, Project administration, Conceptualization. **Wenbing Huang:** Writing-original draft, Methodology, Formal analysis. **Yu Luo:** Writing-review & editing, Visualization, Validation, Formal analysis. **Tianying Ji:** Writing-review & editing, Software, Data curation. **Huaping Liu:** Supervision, Resources, Formal analysis. **He Liu:** Supervision, Resources, Formal analysis. **Jianwei Zhang:** Methodology, Investigation, Conceptualization

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was jointly funded by the National Science and Technology Major Project of the Ministry of Science and Technology of China (2018AAA0102900) and the “New Generation Artificial Intelligence” Key Field Research and Development Plan of Guangdong Province (2021B0101410002).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.eng.2024.10.013>.

References

- [1] Miriyev A, Kovač M. Skills for physical artificial intelligence. *Nat Mach Intell* 2020;2(11):658–60.
- [2] Åström KJ, Murray RM. *Feedback systems: an introduction for scientists and engineers*. Princeton: Princeton University Press; 2010.
- [3] Sünderhauf N, Brock O, Scheirer W, Hadsell R, Fox D, Leitner J, et al. The limits and potentials of deep learning for robotics. *Int J Robot Res* 2018;37(4–5): 405–20.
- [4] Wang W, Siau K. Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: a review and research agenda. *J Database Manage* 2019;30(1):61–79.
- [5] Osa T, Pajarinen J, Neumann G, Bagnell JA, Abbeel P, Peters J. An algorithmic perspective on imitation learning. *Found Trends Robotics* 2018;7(1–2):1–179.
- [6] Kretschmar H, Spies M, Sprunk C, Burgard W. Socially compliant mobile robot navigation via inverse reinforcement learning. *Int J Robot Res* 2016; 35(11):1289–307.
- [7] Kohl N, Stone P. Policy gradient reinforcement learning for fast quadrupedal locomotion. In: *Proceedings of the IEEE International Conference on Robotics and Automation*; 2004 Apr 26–May 1; New Orleans, LA, USA. New York City: IEEE; 2004. p. 2619–24.
- [8] Akkaya I, Andrychowicz M, Chociej M, Litwin M, McGrew B, Petron A, et al. Solving rubik’s cube with a robot hand. 2019. arXiv:1910.07113.
- [9] Zhang K, Yang Z, Basar T. Multi-agent reinforcement learning: a selective overview of theories and algorithms. In: Vamvoudakis KG, Wan Y, Lewis FL, Cansever D, editors. *Handbook of reinforcement learning and control*. Berlin: Springer; 2021. p. 321–84.
- [10] Ahn C, Kim E, Oh S. Deep elastic networks with model selection for multi-task learning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. New York City: IEEE; 2019. p. 6529–38.
- [11] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2016 Jun 27–30; VegasLas, NV, USA. New York City: IEEE; 2016. p. 770–8.
- [12] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. 2017. arXiv:1704.04861.
- [13] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*; 2012 Dec 3–6; TahoeLake, NA, USA. Trier: the dblp computer science bibliography; 2012. p. 1097–105.
- [14] Girshick R. Fast R-CNN. In: *Proceedings of the IEEE International Conference on Computer Vision*; 2015 Dec 7–13; Santiago, Chile. New York City: IEEE; 2015. p. 1440–48.
- [15] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2014 Jun 23–28; Columbus, OH, USA. New York City: IEEE; 2014. p. 580–7.
- [16] He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans Pattern Anal Mach Intell* 2015; 37(9):1904–16.
- [17] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2017 Jul 21–26; Honolulu, HI, USA. New York City: IEEE; 2017. p. 7263–71.
- [18] Redmon J, Farhadi A. YOLOv3: an incremental improvement. 2018. arXiv: 1804.02767.
- [19] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: *Proceedings of the IEEE International Conference on Computer Vision*; 2017 Oct 22–29; Venice, Italy. New York City: IEEE; 2017. p. 2961–9.
- [20] Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. In: *Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems*; 2015 Dec 7–12; Montreal, QC, Canada. Cambridge: The MIT Press; 2015. p. 91–9.
- [21] Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition-2017*; 2017 Jul 21–26; Honolulu, HI, USA. New York City: IEEE; 2017. p. 2117–25.
- [22] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real time object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition-2016*; 2016 Jun 27–30; VegasLas, NV, USA. New York City: IEEE; 2016. p. 779–88.
- [23] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot multibox detector. In: *Proceedings of the European Conference on Computer Vision*; 2016 Oct 11–14; Amsterdam, the Netherlands. Berlin: Springer; 2016. p. 21–37.
- [24] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*; 2017 Oct 22–29; Venice, Italy. New York City: IEEE; 2017. p. 2980–8.
- [25] Law H, Deng J. CornerNet: detecting objects paired keypoints. In: *Proceedings of the European Conference on Computer Vision (ECCV 2018)*; 2018 Sep 8–14; Munich, Germany. Berlin: Springer; 2018. p. 734–50.
- [26] Zhou X, Zhuo J, Krahenbuhl P. Bottom-up object detection by grouping extreme and center points. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition-2019*; 2019 Jun 15–20; LongBeach, CA, USA. New York City: IEEE; 2019. p. 850–9.
- [27] Zhu C, He Y, Savvides M. Feature selective anchor-free module for single-shot object detection. 2019. arXiv:1903.00621.
- [28] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T,

- et al. An image is worth 16 16 words: transformers for image recognition at scale. 2020. arXiv:2010.11929.
- [29] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 10–17; Montreal, QC, Canada. New York City: IEEE; 2021. p. 10012–22.
- [30] Fang Y, Liao B, Wang X, Fang J, Qi J, Wu R, et al. You only look at sequence: rethinking transformer in vision through object detection. Proceedings of the 35th Annual Conference on Neural Information Process 2021 Dec 6–14; online. San Diego: Neural Information Processing Systems; 2021.
- [31] Song H, Sun D, Chun S, Jampani V, Han D, Heo B, et al. ViDT: an efficient effective fully transformer-based object detector. 2021. arXiv:2110.03921.
- [32] Jing M, Ma X, Huang W, Sun F, Yang C, Fang B, et al. Reinforcement learning from imperfect demonstrations under soft expert guidance. Proc Conf AAAI Artif Intell 2020;34(04):5109–16.
- [33] Kong T, Sun F, Liu H, Jiang Y, Li L, Shi J. Foveabox: beyond anchor-based object detection. IEEE Trans Image Process 2020;29:7389–98.
- [34] Liu H, Wang F, Guo D, Liu X, Zhang X, Sun F. Active object discovery and localization using sound-induced attention. IEEE Trans Industr Inform 2021; 17(3):2021–9.
- [35] Bajcsy R, Aloimonos Y, Tsotsos JK. Revisiting active perception. Auton Robots 2018;42(2):177–96.
- [36] Liu H, Den Y, Guo D, Fang B, Sun F, Yang W. An interactive perception method for warehouse automation in smart cities. IEEE Trans Industr Inform 2021;17(2):830–8.
- [37] Silver D, Singh SP, Precup D, Sutton RS. Reward is enough. Artif Intell 2021; 299:103535.
- [38] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature 2015; 518(7540):529–33.
- [39] Sutton RS, McAllester D, Singh S, Mansour Y. Policy gradient methods reinforcement learning with function approximation. In: Proceed of the Annual Conference on Neural Information Processing Systems (1999); 1999 Nov 29–Dec 4; Denver, CO, USA. Cambridge: The MIT P 1999.
- [40] Lillierap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous co with deep reinforcement learning. In: Proceedings of the 4th Internati Conference on Learning Representations, ICLR 2016; 2016 May 2–4; San J Puerto Rico. Trier: the dblp computer science bibliography; 2016.
- [41] Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: off-policy maxi entropy deep reinforcement learning with a stochastic actor. In: Proceed of the 35th International Conference on Machine Learning; 2018 Jul 10 Stockholm, Sweden. New York City: Proceedings of Machine Lear Research; 2018. p. 1856–65.
- [42] Wang T, Bao X, Clavera I, Hoang J, Wen Y, Langlois E, et al. Benchmar model-based reinforcement learning. 2019. arXiv:1907.02057v1.
- [43] Janner M, Fu J, Zhang M, Levine S. When to trust your model: model-b policy optimization. In: Proceedings of the Annual Conference on Ne Information Processing Systems; 2019 Dec 8–14; Vancouver, BC, Canada. Diego: Neural Information Processing Systems Foundation, Inc.; 201 12498–09.
- [44] Tassa Y, Erez T, Todorov E. Synthesis and stabilization of complex beha through online trajectory optimization. In: Proceedings of the 2012 IEEE International Conference on Intelligent Robots and Systems; 2012 Oct 7 Vilamoura, Portugal. New York City: IEEE; 2012. p. 4906–13.
- [45] Zhou Z, Yan N. A survey of numerical methods for convection-diffusion optimal control problems. J Numer Math 2014;22(1):61–85.
- [46] De Boer PT, Kroese DP, Mannor S, Rubinstein RY. A tutorial on the cross entropy method. Ann Oper Res 2005;134(1):19–67.
- [47] Chua K, Calandra R, McAllister R, Levine S. Deep reinforcement learning handful of trials using probabilistic dynamics models. In: Proceedings of Annual Conference on Neural Information Processing Systems; 2018 Dec Montreal, QC, Canada. Red Hook: Curran Associates Inc.; 2018. p. 4759–70.
- [48] Yildiz C, Heinonen M, Lähdesmäki H. Continuous-time model-b reinforcement learning. In: Proceedings of the International Conferenc Machine Learning; 2021 Jun 18 – 24; online. New York City: Proceeding Machine Learning Research; 2021. p. 12009–18.
- [49] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal p optimization algorithms. 2017. arXiv:1707.06347.
- [50] Schulman J, Levine S, Abbeel P, Jordan M, Moritz P. Trust region p optimization. In: Proceedings of the International Conference on Mac Learning; 2015 Jul 6–11; Lille, France. New York City: Proceedings of Mac Learning Research; 2015. p. 1889–97.
- [51] Chang D, Johnson-Roberson M, Sun J. An active perception framework for autonomous underwater vehicle navigation under sensor constraints. IEEE Trans Control Syst Technol 2022;30(6):2301–16.
- [52] Amos B, Jimenez I, Sacks J, Boots B, Zico Kolter J. Differentiable MPC for end to-end planning and control. In: Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (NIPS 2018); 2018 Dec 2–8; Montreal, QC, Canada. San Diego: Neural Information Processing Systems; 2018.
- [53] Pong V, Gu S, Dalal M, Levine S. Temporal difference models: model-free deep RL for model-based control. In: Proceedings of the International Conference on Learning Representations; 2018 Apr 30–May 3; Vancouver, BC, Canada. Trier: the dblp computer science bibliography; 2018.
- [54] Ahn M, Zhu H, Hartikainen K, Ponte H, Gupta A, Levine S, Kumar V. ROBEL: robotics benchmarks for learning with low-cost robots. In: Proceedings of the Conference on Robot Learning; 2020 Nov 16–18; online; 2020.