偏好变换器：使用 Transformer 为强化学习建模人类偏好

Changyeon Kim[1*] **Jongjin Park**[1*] **Jinwoo Shin**[1] **HonglakLee**[2,3] **PieterAbbeel**[4] **Kimin Lee**[5]韩国科学技术院 [2]密歇根大学 [3]LG AI Research [4]加州大学伯克利分校 [5]Google Research

arXiv:2303.00957v1 [cs.LG] 2 Mar 2023

摘要

基于偏好的强化学习（强化学习）提供了一种框架，可利用人类对两种行为的偏好来训练智能体。然而，基于偏好的强化学习难以扩展，因为学习与人类意图一致的奖励函数需要大量人类反馈。本文提出偏好变换器，这是一种使用 Transformer 对人类偏好进行建模的神经网络架构。不同于以往假设人类判断基于马尔可夫奖励、且各种奖励对决策贡献相同的方法，我们提出了一种基于非马尔可夫奖励加权和的新型偏好模型。随后，我们采用 Transformer 架构设计该偏好模型，其中堆叠了因果自注意力层和双向自注意力层。我们证明，偏好变换器能够利用真实人类偏好解决多种控制任务，而以往方法则无法奏效。我们还展示了偏好变换器能够自动捕捉人类决策中的时间依赖关系，从而生成定义明确的奖励，并关注轨迹中的关键事件。代码可在项目网站获取：https://sites.google.com/view/preference-transformer。

1 引言

在具有合适奖励函数的各种领域中，强化学习（RL）在解决序贯决策问题方面已取得成功（Mnih et al.，2015；Silver et al.，2017；Berner etal.，2019年；Vinyals 等人，2019年）。然而，奖励工程带来了诸多挑战。为了设计稠密且精确的奖励，它通常需要大量仪器设备（例如，热成像摄像机（Schenck & Fox，2017）、加速度计（Yahyaet al.，2017），或运动追踪器（Penget al.，2020））。此外，由于许多问题具有多重目标，因此很难用单一标量评估结果的质量。例如，为实现稳定运动，我们需要关注速度、消耗的能量和躯干垂直度等多个目标（Tassa et al.，2012；Faustet al.，2019年）。将多个目标聚合为单一标量需要大量的人类努力和广泛的任务知识。

为了避免奖励工程，可以通过多种方式从人类数据中学习奖励函数，例如实值反馈（Knox & Stone，2009；Daniel 等人，2014）、专家演示（Ng 等人，2000；Abbeel & Ng，2004）、偏好（Akrour 等人，2011；Wilson 等人，2012；Sadigh 等人，2017）以及语言指令（Fu 等人，2019年；Nair 等人，2022）。尤其是，近年来人们对基于偏好的强化学习（Akrour 等人，2012；Christiano等，2017；Lee 等人，2021b）的研究兴趣不断增加，因为做出相对判断（例如成对比较）既容易提供，又包含丰富的信息。通过从人类对轨迹的偏好中学习奖励函数，近期研究表明，智能体能够学习新颖行为（Christiano等，2017；Stiennon 等人，2020）或避免奖励利用（Lee 等人，2021b）。

然而，现有方法仍然需要大量的人类反馈，这使得将基于偏好的强化学习扩展到各种应用中变得困难。我们假设，这一困难源于大多数以往工作在偏好建模中所采用的共同底层假设。具体而言，以往工作通常假设：（a）奖励函数具有马尔可夫性（即，只取决于当前状态和动作）；（b）人类根据奖励之和并以相等权重来评估一条轨迹（智能体的行为）的质量。由于以下原因，这些假设可能存在缺陷：

*贡献相同。联系方式：{changyeon.kim, jongjin.park}@kaist.ac.kr

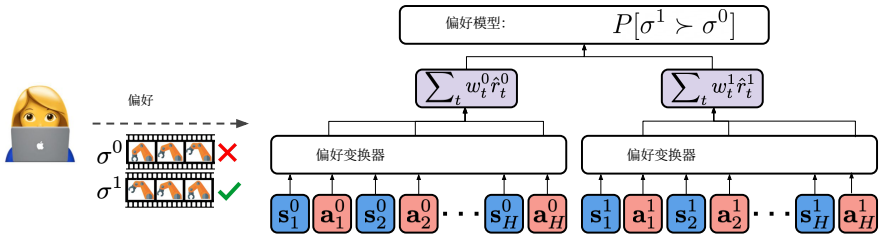


图 1：我们的框架示意图。给定两个轨迹片段之间的偏好（σ^0，σ^1），偏好变换器会生成每个片段上的非马尔可夫奖励 $\hat{r}_t$ 及其重要性权重w_t。随后，我们基于非马尔可夫奖励的加权和（即，$\sum_t w_t \hat{r}_t$）对偏好预测器进行建模，并使其与人类偏好保持一致。

首先，存在许多奖励取决于访问状态的任务（即，非马尔可夫性），因为很难将所有任务相关信息编码到状态中（Bacchus 等人，1996；1997）。这在基于偏好的学习中尤其如此，因为轨迹片段是按顺序提供给人类的（例如，视频片段（Christiano 等人，2017；Lee 等人，2021b）），从而使较早发生的事件能够影响人类对后续事件的评分。此外，由于人类对引人注目的时刻高度敏感（卡尼曼，2000），因此需要在轨迹内部进行信用分配。例如，在使用眼动追踪器研究人类对电子游戏的注意力时（Zhang 等人，2020），人类玩家需要在可能带来较大奖励或惩罚的重要状态上花费更长的反应时间，并进行多次眼动，以便作出决策。

本文旨在提出一种替代性的偏好模型，以克服先前工作中常见假设的局限性。为此，我们提出了一种基于非马尔可夫奖励加权和的新型偏好模型，该模型能够捕捉人类决策中的时间依赖关系，并推断轨迹中的关键事件。受变换器在Vaswani 等人，2017提出的用于建模序列数据的近期成功启发（Brown 等人，2020；Chen 等人，2021），我们提出了偏好变换器，这是一种用于构建所提出偏好模型的基于变换器的架构（见图1）。偏好变换器以轨迹片段为输入，从而能够提取与任务相关的历史信息。通过堆叠双向和因果自注意力层，偏好变换器生成非马尔可夫奖励和重要性权重作为输出。随后，我们利用它们定义偏好模型。

本文的主要贡献如下：

- 我们提出了一个更通用的框架，基于非马尔可夫奖励的加权和来建模人类偏好。
- 我们提出偏好变换器，这是一种基于变换器的架构，由我们为所提出的框架设计的新型偏好注意力层构成。
- 偏好变换器使强化学习智能体能够通过从Fu 等人，2020）基准测试中学习奖励函数，解决 D4RL 的复杂导航、运动任务，以及 Robomimic（Man- dlekar 等人，2021）基准测试中的机器人操作任务，并从真实人类偏好中学习。
- 我们分析了学得的奖励函数和重要性权重，表明偏好变换器能够生成定义明确的奖励，并捕捉轨迹中的关键事件。

2 相关工作

基于偏好的强化学习。近年来，各种方法利用人类偏好训练强化学习智能体，而无需进行奖励工程（Akrour 等人，2012；Christiano 等人，2017；Ibarz 等人，2018；Stiennon 等人，2020；Lee 等人，2021b；c；Nakano 等人，2021；Wu 等人，2021；III & Sadigh，2022；Knox 等人，2022；Ouyang 等人，2022；Park 等人，2022；Verma & Metcalf，2022）。Christiano 等人（2017）表明，基于偏好的强化学习能够利用深度神经网络有效解决复杂的控制任务。为提高反馈效率，研究者提出了多种方法，例如预训练（Ibarz 等人，2018；Lee 等人，2021b）、数据增强（Park 等人，2022）、探索（Liang 等人，2022）和元学习（III & Sadigh，2022）。基于偏好的强化学习也成功用于针对困难任务微调大规模语言模型（如 GPT-3；Brown 等人，2020）（Stiennon 等人，2020；Wu 等人，2021；Nakano 等人，2021；Ouyang 等人，

2022）。由于语言输入存在部分可观测性，这些方法将基于 Transformer 的模型（即，预训练语言模型）用作奖励函数；而我们则针对控制任务采用了融入全新偏好建模的 Transformer。

用于强化学习和模仿学习的 Transformer。Transformer（Vaswani 等，2017）已被用于强化学习中的各种目的（Vinyals 等人，2019；Zambaldi 等人，2019；Parisotto 等人，2020；Chen 等，2021；Janner 等，2021）。研究发现，在星际争霸（Vinyals 等人，2019；Zambaldi 等人，2019）和 DMLab-30（Parisotto 等人，2020）等复杂环境的基准测试中，使用 Transformer 对强化学习智能体进行建模可以提高样本效率和泛化能力。对于离线强化学习，Chen 等（2021）和 Janner 等（2021）也通过将强化学习问题表述为序列建模问题，利用了 Transformer。此外，Transformer 已成功应用于模仿学习（Dasari 和 Gupta，2020；Mandi 等人，2022；Reed 等人，2022）。Dasari 和 Gupta（2020）以及 Mandi 等人（2022）展示了 Transformer 在单样本模仿学习（Duan 等人，2017）中的泛化能力，而 Reed 等人（2022）则利用 Transformer 进行多领域和多任务模仿学习。在本文中，我们证明了 Transformer 同样有助于提高基于偏好的学习的效率。

非马尔可夫奖励学习。非马尔可夫奖励（Bacchus 等人，1996；1997）已被用于处理现实中的奖励设定：奖励取决于访问状态，或者奖励会延迟发放，甚至在每个回合结束时才给出。近期有多项研究聚焦于回报分解：在这一方法中，会额外训练一个模型，使其根据给定的状态-动作序列预测轨迹回报，并将其用于奖励重新分配。Arjona-Medina 等人（2019年）和Early 等人（2022）使用长短期记忆网络（Hochreiter 和 Schmidhuber，1997）来捕捉奖励学习中的序列信息。Gangwani 等人（2020）和Ren 等人（2022）提出了无需任何先验任务知识的简化方法，其假设是回合回报在各时间步均匀分布。在本研究中，我们采用基于 Transformer 的奖励模型来学习非马尔可夫奖励，该模型能够捕捉人类决策中的时间依赖关系。

3 预备知识

我们考虑这样一种强化学习（RL）框架：智能体在离散时间内与环境交互（Sutton 和 Barto，2018）。形式上，在每个时间步 t，智能体从环境接收当前状态 $\mathbf{s}_t$，并根据其策略 π选择动作 $\mathbf{a}_t$。环境给出奖励 $r(\mathbf{s}_t,\ \mathbf{a}_t)$，随后智能体转移到下一个状态 $\mathbf{s}^{t+1}$。强化学习的目标是学习一个能够最大化期望回报的策略，期望回报 $\mathcal{R}_t = \sum_{k=0}^{\infty}\gamma^k r(\mathbf{s}^{t+k},\ \mathbf{a}^{t+k})$定义为奖励的折扣累积和，折扣因子为 γ。

在许多应用中，很难设计出能够体现人类意图的合适奖励函数。基于偏好的强化学习（Akrour 等人，2011；Pilarski 等人，2011；Wilson 等人，2012；Christiano 等人（2017））通过从人类偏好中学习奖励函数来解决这一问题。与 Wilson 等人（2012）和 Christiano 等人（2017）类似，我们考虑对长度为 H, $\sigma = \{(\mathrm{s}_1, \mathrm{a}_1)\ ..., (\mathrm{s}_H, \mathrm{a}_H)\}$ 的两个轨迹片段进行偏好比较。给定一对片段（σ^0, σ^1），一位（人类）教师指出更偏好哪个片段，即 $y \in \{0, 1, 0.5\}$。偏好标签 $y = 1$ 表示 $\sigma^1 \succ \sigma^0$，$y = 0$ 表示 $\sigma^0 \succ \sigma^1$，而 $y = 0.5$ 表示两者同样受偏好，其中 $\sigma^i \succ \sigma^j$表示片段 i 优于片段 j这一事件。每条反馈都作为三元组（σ^0, σ^1, y）存储在偏好数据集 $\mathcal{D}$ 中。

为了获得奖励函数 $\hat{}r$ 参数化形式为 ψ，大多数先前工作（Christiano 等人，2017；Ibarz 等人，2018；Lee 等人，2021b；c；III & Sadigh，2022；Park 等人，2022）定义了一个遵循 Bradley-Terry 模型（Bradley 与 Terry，1952）的偏好预测器：

$$P[\sigma^1 \succ \sigma^0; \psi] = \frac{\exp\left(\sum_t \hat{r}(\mathbf{s}_t^1, \mathbf{a}_t^1; \psi)\right)}{\exp\left(\sum_t \hat{r}(\mathbf{s}_t^1, \mathbf{a}_t^1; \psi)\right) + \exp\left(\sum_t \hat{r}(\mathbf{s}_t^0, \mathbf{a}_t^0; \psi)\right)}. \tag{1}$$

然后，给定一个偏好数据集 $\mathcal{D}$，通过最小化该偏好预测器与实际人工标注之间的交叉熵损失，来更新奖励函数 $\hat{}r$：

$$\mathcal{L}^{\mathtt{CE}}(\psi) = -\mathop{\mathbb{E}}_{(\sigma^0, \sigma^1, y)\sim\mathcal{D}}\Big[(1-y)\log P[\sigma^0 \succ \sigma^1; \psi] + y \log P[\sigma^1 \succ \sigma^0; \psi]\Big]. \tag{2}$$

可以使用任意强化学习算法更新策略 π，使其最大化相对于学得奖励的期望回报。

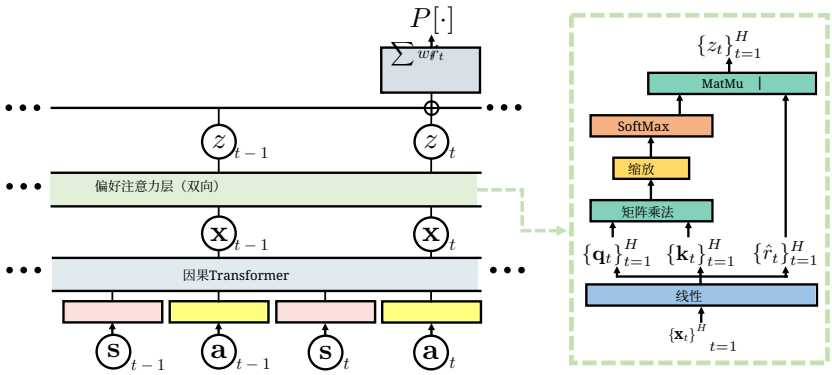


图2：偏好变换器概览。我们首先通过因果Transformer构造隐藏嵌入 $\{\mathbf{x}_t\}$，其中每个嵌入都表示从初始时间步到时间步 t的上下文信息。带有双向自注意力的偏好注意力层根据这些隐藏嵌入计算非马尔可夫奖励 $\{\hat{r}_t\}$ 及其凸组合 $\{z_t\}$，然后我们聚合 $\{z_t\}$，以对非马尔可夫奖励 $\sum_t w_t \hat{r}_t$的加权和进行建模。

4 偏好变换器

本节介绍偏好变换器（PT），这是一种用于建模人类偏好的 Transformer 架构（概览见图 2）。首先，我们在第 $P[\sigma^1 \succ \sigma^0]$ **4.1** 节提出一种新的偏好预测器，它基于非马尔可夫奖励的加权和，能够反映智能体行为的长期上下文，并捕捉轨迹片段中的关键事件。随后，我们在第 4.2 节介绍一种新颖的基于变换器的架构，用于建模所提出的偏好预测器。

4.1 偏好建模

如第 3 节所述，大多数既有研究都假设奖励具有马尔可夫性（即，仅取决于当前状态和动作），并假设人类根据奖励之和（各项权重相同）来评估轨迹片段的质量。基于这些假设，偏好预测器 $P[\sigma^1 \succ \sigma^0]$定义如方程 1所示。然而，这种表述在建模真实人类偏好方面存在若干局限。首先，在许多情况下，很难用马尔可夫奖励来明确描述任务（Bacchus 等人，1996；1997；Early 等人，2022）。此外，由于人类对显著时刻十分敏感（卡尼曼，2000），轨迹中的信用分配也可能是必要的。

To 为解决上述问题，我们引入一种新的偏好预测器，假设偏好某个片段的概率 y of p随非马尔可夫奖励加权和呈指数变化：

$$P[\sigma^1 \succ \sigma^0; \psi] = \frac{\exp\left(\sum_t w\left(\{(\mathbf{s}_i^1, \mathbf{a}_i^1)\}_{i=1}^H; \psi\right)_t \cdot \hat{r}\left(\{(\mathbf{s}_i^1, \mathbf{a}_i^1)\}_{i=1}^t; \psi\right)\right)}{\sum_{j\in\{0,1\}} \exp\left(\sum_t w\left(\{(\mathbf{s}_i^j, \mathbf{a}_i^j)\}_{i=1}^H; \psi\right)_t \cdot \hat{r}\left(\{(\mathbf{s}_i^j, \mathbf{a}_i^j)\}_{i=1}^t; \psi\right)\right)}. \tag{3}$$

为捕捉时间依赖关系，我们考虑一个非马尔可夫奖励函数 $\hat{r}$，它接收完整的前序子轨迹 $\{(\mathbf{s}_i,\mathbf{a}_i)\}_{i=1}^t$ 作为输入。重要性权重 w 也被引入为整个轨迹片段 $\{(\mathbf{s}_i,\mathbf{a}_i)\}_{i=1}^H$的函数，使偏好预测器能够在片段内进行信用分配。我们指出，我们的公式是传统设计的一种推广版本；如果奖励函数仅取决于当前状态-动作对，且重要性权重始终为 1，那么我们的偏好预测器就等同于公式 1中的标准模型。

4.2 架构

In o为了按照公式 3对偏好预测器进行建模，我们提出一种新的变换器架构 - ture，称为偏好变换器，由以下可学习组件构成：

因果Transformer。我们使用 Transformer 网络（Vaswani 等人，2017）作为骨干网络，这一选择受到其在建模序列数据方面优势的启发（Brown 等人，2020；Ramesh 等人，2021；

Janner 等，2021；Chen 等，2021）。具体而言，我们使用 GPT 架构（Radford 等，2018），即采用因果掩码自注意力的 Transformer 架构。给定长度为 H 的轨迹片段 $\sigma = \{(\mathbf{s}_1, \mathbf{a}_1) \cdots, (\mathbf{s}_H, \mathbf{a}_H)\}$，我们生成 $2H$ 个输入嵌入（对应状态和动作），并通过线性层和随后的层归一化（Ba 等，2016）进行学习。我们学习一个共享的位置嵌入（即同一时间步的状态和动作共享相同的位置嵌入），并将其加到每个输入嵌入上。随后将输入嵌入馈入因果 Transformer 网络，该网络生成输出嵌入 $\{\mathbf{x}_t\}_{t=1}^{H}$，使得第 t 个输出仅依赖于截至 t 的输入嵌入。

偏好注意力层。为使用公式3中定义的非马尔可夫奖励加权和来建模偏好预测器，我们引入了偏好注意力层。如图 2 所示，偏好注意力层接收来自因果Transformer$\{\mathbf{x}_t\}_{t=1}^{H}$ 的隐藏嵌入，并生成奖励 $\hat{r}$ 和重要性权重 w_t。形式上，第 t 个输入 $\mathbf{x}_t$ 通过线性变换映射为键 $\mathbf{k}_t \in \mathbb{R}^d$、查询 $\mathbf{q}_t \in \mathbb{R}^d$和值 $\hat{r}_t \in \mathbb{R}$，其中 d 为嵌入维度。我们指出，自注意力中的第 t 个值，即 $\hat{r}_t$，被视为对非马尔可夫奖励 $\hat{r}(\{(\mathbf{s}_i, \mathbf{a}_i)\}_{i=1}^{t})$进行建模，因为隐藏嵌入 $\mathbf{x}_t$ 仅取决于轨迹片段中之前的输入。具体而言，在每个时间步 t，我们使用 t 个状态-动作对 $\{(\mathbf{s}_i, \mathbf{a}_i)\}_{i=1}^{t}$ 来近似训练期间的奖励 $\hat{r}_t$。

按照自注意力（Vaswani 等人，2017），第 i 个输出 z_i 定义为值的凸组合，其注意力权重由第 i个查询和键计算得出：

$$z_i = \sum_{t=1}^{H} \texttt{softmax}(\{\langle \mathbf{q}_i, \mathbf{k}_{t'} \rangle\}_{t'=1}^{H})_t \cdot \hat{r}_t.$$

然后，奖励的加权和可以通过对输出取平均值来计算 $\{z_i\}_{i=1}^{H}$，如下所示：

$$\frac{1}{H}\sum_{i=1}^{H} z_i = \frac{1}{H}\sum_{i=1}^{H}\sum_{t=1}^{H} \texttt{softmax}(\{\langle \mathbf{q}_i, \mathbf{k}_{t'} \rangle\}_{t'=1}^{H})_t \cdot \hat{r}_t = \sum_{t=1}^{H} w_t \hat{r}_t \tag{4}$$

其中 $w_t = \frac{1}{H}\sum_{i=1}^{H} \texttt{softmax}(\{\langle \mathbf{q}_i, \mathbf{k}_{t'} \rangle\}_{t'=1}^{H})_t$。这里，$w_t$ 对应于对重要性权重 $w(\{(\mathbf{s}_i, \mathbf{a}_i)\}_{i=1}^{H})_t$ 进行建模，这一权重出现在方程 3 中，因为这种偏好注意力没有因果掩码，因此依赖于完整序列（即双向自注意力）。总之，我们通过对变换器网络的输出取平均值来建模非马尔可夫奖励的加权和，并假设偏好该片段的概率与之成正比。偏好变换器的完整架构如图2所示。

4.3 训练与推理

训练。给定偏好数据集 2，我们通过最小化公式中的交叉熵损失来训练偏好变换器 $\mathcal{D}$。通过使偏好变换器所建模的偏好预测器与人工标注相一致，我们发现偏好变换器能够诱导出合适的奖励函数，并捕捉轨迹片段中的重要事件（支持性结果见图 3 ）。

推理。在强化学习训练中，所有状态-动作对都使用学得的奖励函数进行标注。由于我们训练的是非马尔可夫奖励函数，因此向偏好变换器提供 H 过去的状态转移（$\mathbf{s}_{t-H+1}$、$\mathbf{a}_{t-H+1}$、$\cdots$、$\mathbf{s}_t$、$\mathbf{a}_t$），并将偏好注意力层的第 t 个值 $\hat{r}_t$用作时间步 t的奖励。

5 实验

我们设计实验以探究以下问题：

- 偏好变换器能否利用真实人类偏好解决复杂的控制任务?
- 偏好变换器能否诱导出良好对齐的奖励，并关注关键事件?
- 偏好变换器在合成偏好下（即脚本化教师设置）表现如何?

表 1：使用不同奖励函数时，IQL 在 AntMaze、Gym-Mujoco 运动任务上的平均归一化得分，以及在 Robosuite 操作任务上的成功率。使用来自真实人类教师的同一偏好数据集，我们训练了偏好变换器（PT）、基于多层感知机的马尔可夫奖励（MR；Christiano 等，2017；Lee 等，2021b）和基于 LSTM 的非马尔可夫奖励（NMR；Early 等，2022）。结果展示了 8 次运行的平均值和标准差。

数据集	使用任务奖励的 IQL	使用偏好学习的 IQL		
		MR	NMR	偏好变换器（我们的方法）
antmaze-medium-play-v2	73.88 ± 4.49	31.13 ± 16.96	62.88 ± 5.99	70.13 ± 3.76
antmaze-medium-diverse-v2	68.13 ± 10.15	19.38 ± 9.24	20.13 ± 17.12	65.25 ± 3.59
antmaze-large-play-v2	48.75 ± 4.35	24.25 ± 14.03	14.13 ± 3.60	42。38 ± 9.98
antmaze-large-diverse-v2	44.38 ± 4.47	5.88 ± 6.94	0.00 ± 0.00	19.63 ± 3.70
antmaze-v2 总计	58.79	20.16	24.29	49.35
Hopper 中等回放 v2	83.06 ± 15.80	11.56 ± 30.27	57.88 ± 40.63	84.54 ± 4.07
Hopper 中等专家 v2	73.55 ± 41.47	57.75 ± 23.70	38.63 ± 35.58	68.96 ± 33.86
walker2d-medium-replay-v2	73.11 ± 8.07	72.07 ± 1.96	77.00 ± 3.03	71.27 ± 10.30
walker2d-medium-expert-v2	107.75 ± 2.02	108.32 ± 3.87	110.39 ± 0.93	110.13 ± 0.21
locomotion-v2 总计	84.37	62.43	70.98	83.72
lift-ph	96.75 ± 1.83	84.75 ± 6.23	91.50 ± 5.42	91.75 ± 5.90
lift-mh	86.75 ± 2.82	91.00 ± 4.00	90.75 ± 5.75	86。75 ± 5.95
can-ph	74.50 ± 6.82	68.00 ± 9.13	62.00 ± 10.90	69。67 ± 5.89
can-mh	56.25 ± 8.78	47.50 ± 3.51	30.50 ± 8.73	50.50 ± 6.48
Robosuite 总分	78.56	72.81	68.69	74.66

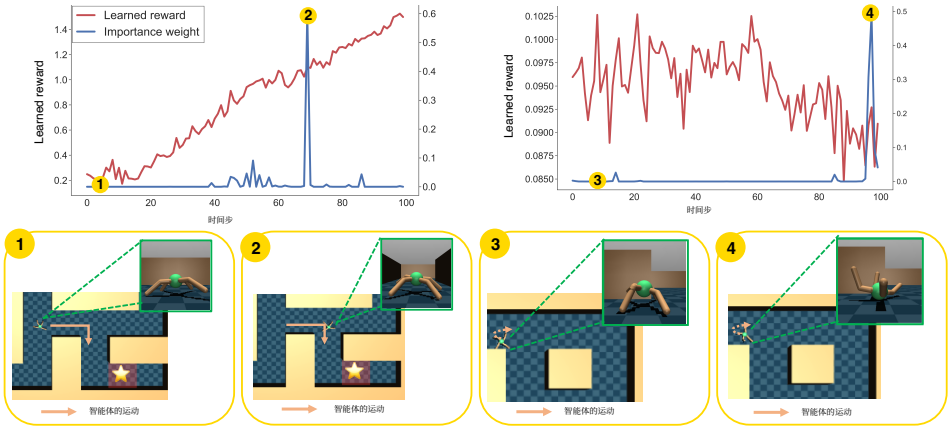
5.1 设置

与 Shin &Brown (2021) 类似，我们使用 D4RL（Fu 等人，2020）基准和 Robomimic（Man- dlekar 等人，2021）基准，在离线设定下评估偏好变换器（PT）在若干复杂控制任务上的表现。[1] 具体而言，我们考虑三个不同的领域：AntMaze、Gym-Mujoco 运动（Todorov 等人，2012；Brockman 等人，2016）基准（来自 D4RL），以及 Robomimic 基准中的 Robosuite 机器人操作（Zhu 等人，2020），并采用不同的数据收集方案。在奖励学习过程中，我们从离线数据集中均匀随机地选取查询（成对的轨迹片段），并向真实的人类训练者（即作者）收集偏好。[2] 随后，我们利用收集到的人类偏好数据集学习奖励函数，并使用隐式 Q 学习（IQL；Kostrikov 等人 2022）训练强化学习智能体。IQL 是一种近期提出的离线强化学习算法，在 D4RL 基准上表现出色。我们遵循 IQL 的原始实现（例如，架构和超参数），采用马尔可夫策略和价值函数。此外，我们还在附录 G中提供了使用非 MDP 模型进行离线强化学习实验的结果。

评估时，我们衡量遵循原始基准（D4RL）底层任务奖励的专家归一化分数 [3] 以及成功率（Robomimic）。作为基线方法，我们考虑基于马尔可夫奖励（MR）或非马尔可夫奖励（NMR）的标准偏好建模。对于

基于 MR 的模型，我们将偏好预测器定义为 $P[\sigma^1 \succ \sigma^0;\ \psi_{\text{MR}}] = \exp(\sum_t \hat{r}(\mathbf{s}_t^1, \mathbf{a}_t^1; \psi_{\text{MR}}))/\sum_j \exp(\sum_t \hat{r}(\mathbf{s}_t^j, \mathbf{a}_t^j; \psi_{\text{MR}}))$使用与 Christiano 等人 (2017) 类似的、由多层感知机建模的奖励函数；Lee 等人 (2021b)。对于基于 NMR 的模型，我们采用基于 LSTM 的架构（Early 等人，2022）作为奖励函数

[1]为了专注于评估奖励学习的性能，我们采用离线设定。请注意，先前研究也考虑了假设可获得专家演示的类似设定（Ibarz 等人，2018）。[2] 我们在 AntMaze 中收集了 100（medium）/1000（large）个查询，在 Gym-Mujoco 运动中收集了 500（medium-replay）/100（medium-expert）个查询，并在 Robosuite 机器人操作任务中收集了 100（PH）/500（MH）个查询。除 Gym-Mujoco 运动的 medium-expert 任务和 antmaze-large 任务外，所有情况下收集反馈所需的人工时间最多为 10 分钟；这两项任务则需要 1 至 2 小时的人工时间。[3]由于偏好由熟悉机器人学和强化学习领域的专家提供，我们参照 Christiano 等人 (2017)，根据任务奖励衡量性能。



(a) 成功轨迹　　(b) 失败轨迹

图3：在 antmaze-large-play-v2 的 (a) 成功轨迹片段和 (b) 失败轨迹片段上，学习到的奖励函数（红色曲线）和重要性权重（蓝色曲线）的时间序列。两种情况下，重要性权重的峰值都对应着关键事件：向右转以到达目标（点 2），或发生翻转（点 4）。学习到的奖励也与人类意图高度一致：智能体接近目标时奖励增加，而智能体翻转时奖励降低。

并将偏好预测器定义为 $P[\sigma^1 \succ \sigma^0; \psi_{\text{NMR}}] = \exp(\sum_t \hat{r}(\{(s_t^1, a_t^1)\}_{t=1}^t; \psi_{\text{NMR}})) \sum_j \exp(\sum_t \hat{r}(\{(s_t^j, a_t^j)\}_{t=1}^t; \psi_{\text{NMR}}))$.[4] 请注意，两个基线方法都采用基于等权重奖励总和的相同偏好建模，而我们的方法则基于由 Transformer 建模的非马尔可夫奖励加权和。我们还报告了使用基准测试中的任务奖励训练的 IQL 的性能，作为参考。对于所有实验，我们报告 8 次运行的均值和标准差。更多实验细节（例如，任务描述、反馈收集和奖励学习）见附录 B 和 C。

5.2 使用真实人类教师的基准任务

表 1展示了 IQL 使用不同奖励函数时的表现。在几乎所有任务中，偏好变换器始终优于所有基线方法。尤其是，只有我们的方法几乎达到了使用任务奖励时 IQL 的表现，而基线方法在困难任务中无法奏效。这表明，偏好变换器能够从真实人类偏好中推导出合适的奖励函数，并教会智能体有意义的行为。具体而言，在复杂任务（即，AntMaze）中，偏好变换器与基线方法之间存在很大差距，因为在这项任务中，捕捉智能体行为的长期上下文（例如，智能体的移动方向）和关键事件（例如，目标位置）非常重要。这些结果表明，我们基于变换器的偏好模型在奖励学习方面非常有效。

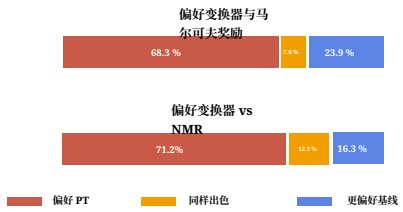


图 4：在 4 项 AntMaze 任务上的人类评估平均结果。数字表示评估者在 40 次试验中的回答统计数据。与马尔可夫奖励和 NMR 相比，偏好变换器获得了更高的评分。

人工评估。为检验学到的奖励是否确实与人类偏好一致，我们也进行了人工评估。我们从使用两种不同奖励训练的智能体中生成一个查询（即两条轨迹），由一位人类评估者（作者）判断哪条轨迹更好。[5] 图 4 展示了偏好变换器与基线方法在 40 组

[4]原始研究（Early 等人，2022）仅考虑了回报预测，而我们在基于偏好的学习中采用了所提出的长短期记忆网络架构。[5]为确保比较公平，轨迹经过匿名处理。此外，如果两个智能体都未能完成任务，评估者会跳过该查询，以便进行有意义的比较。

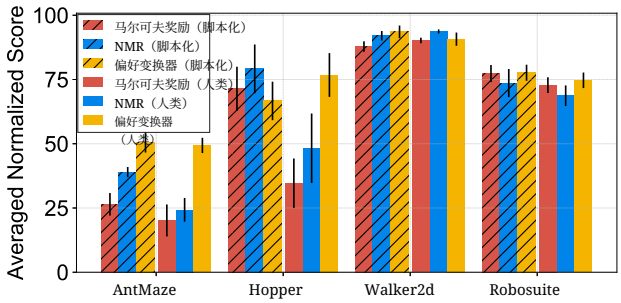


图5：在 AntMaze、Gym-Mujoco 运动（Hopper 和 Walker2d）以及 Robosuite 机器人操作任务中，使用人类偏好和合成偏好训练的 IQL 在各种奖励函数下的平均归一化得分。结果显示了 8 次运行的均值和标准差。我们的方法（PT）在脚本化教师和人类教师下都取得了良好的表现，而基线方法（MR 和 NMR）在人类教师下的表现则显著下降。

查询上的比较结果。[6] 我们观察到，相较于使用马尔可夫奖励或 NMR 的智能体，人类评估者更偏好使用偏好变换器的智能体，这表明偏好变换器与人类偏好更加一致。

5.3 奖励与权重分析

我们评估偏好变换器能否推导出定义明确的奖励，并从人类偏好中捕捉关键事件。图3 展示了在`antmaze-large-play-v2`中的成功与失败轨迹片段上学习到的奖励函数（红线）和重要性权重（蓝线）。首先，我们发现奖励函数与人类意图高度一致。在成功轨迹中（图3a），随着智能体接近目标，奖励值不断增加；而失败轨迹的奖励较低，因为智能体难以逃离角落，随后发生翻转，如图3b所示。这表明偏好变换器能够在奖励函数中捕捉智能体行为的情境。我们还发现，重要性权重呈现出与奖励函数不同的趋势。有趣的是，在成功和失败轨迹中，重要性权重的峰值都对应着关键事件，例如右转以抵达目标或发生翻转。补充材料中提供了更多视频示例。

5.4 使用脚本教师的基准任务

与先前的研究（Christiano 等人，2017；Lee 等人，2021b；c）类似，我们也使用脚本教师生成的合成偏好来评估偏好变换器。我们考虑一种确定性教师，它根据基准中的任务奖励生成偏好，具体如下： $r: y = i$，其中 $i = \arg\max_i \sum_{t=1}^{H} r(\mathbf{s}_t^i, \mathbf{a}_t^i)$。我们指出，这种脚本教师是方程1中引入的偏好模型的一个特例。[7]

图 5 展示了 IQL 在人类教师和脚本教师提供的不同奖励函数下的表现（完整结果见附录 D）。尽管合成偏好基于对决策贡献相同的马尔可夫奖励，偏好变换器在脚本教师提供的数据上仍表现出色。我们认为，这是因为偏好变换器能够利用历史信息，推导出形状更好的奖励。此外，非马尔可夫表述（公式 3）可以视为马尔可夫表述（公式 1）的泛化版本，因为它可以通过忽略输入中的历史信息来学习马尔可夫奖励。我们还注意到，在某些简单任务上（例如，运动任务），与真实人类教师相比，基线方法在脚本教师提供的数据上表现更好。这表明脚本教师并未准确模拟真实的人类行为，而在脚本教师提供的数据上进行评估可能会产生误导性信息。为详细调查这一问题，我们测量了人类教师与脚本教师之间的一致率。图 6b 显示，不一致率相当高，因为脚本教师无法正确把握

[6]每个集合包含来自使用不同随机种子训练的 8 个模型的 5 条滚动轨迹。[7]如果在方程 β 的指数项中引入一个理性常数（或温度）1，并将其设为$\beta \to \infty$，偏好模型便会变为确定性的。



(a) 轨迹示例　　(b) 一致率（%）

图6：人类教师与脚本教师之间的差异。(a) 展示给人类教师和脚本教师的 AntMaze 任务轨迹示例。人类教师能捕捉行为的上下文（即 方向），从而给出正确标签，而脚本教师则不能。(b) 人类教师与脚本教师之间的一致率。我们发现，所有任务的不一致率都相当高，这意味着基于脚本教师进行评估可能会产生误导性信息。

智能体行为的情境（见图 6a）。需要注意的是，在简单的网格世界领域中也观察到了人类教师与脚本教师之间的分歧（Knox 等人，2022）。这些结果表明，现有基于合成偏好的基于偏好的强化学习基准（Lee 等人，2021c）可能还不够，因此需要专门为基于偏好的强化学习设计新的基准。

5.5 学习复杂的新行为

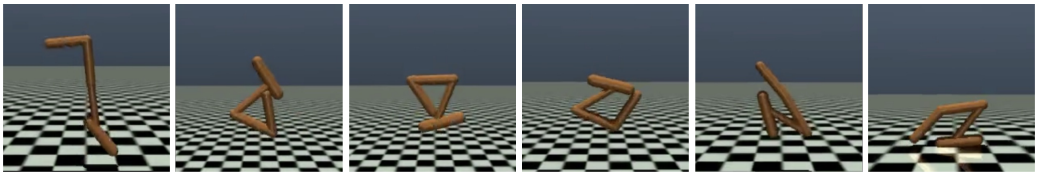


图7：Hopper完成双后空翻的六个画面。智能体通过300次人类反馈查询进行训练，以完成一连串后空翻。

我们展示了偏好变换器在帮助智能体学习复杂且新颖的行为方面的有效性；对于这类行为，合适的奖励函数难以设计。具体而言，我们展示了在 Gym-Mujoco 运动任务中，Hopper 每次跳跃都能完成多个后空翻。与以往基于偏好的强化学习方法所考虑的单次后空翻相比，这项任务更具挑战性（Christiano 等人，2017；Lee 等人，2021b），因为奖励函数必须捕捉包括旋转次数在内的非马尔可夫上下文。我们发现，如图7所示，Hopper 智能体学会了完成双后空翻，而使用马尔可夫奖励函数的智能体则未能学会这一行为。Hopper 后空翻的实现细节见附录B，所有行为（包括三连后空翻）的视频可在项目网站上观看。

6 讨论

本文提出了一个基于非马尔可夫奖励加权和的人类偏好建模新框架，并使用基于变换器的架构设计了该框架。我们提出了一种新颖的偏好注意力层，用于学习非马尔可夫奖励函数的加权和，使得模型能够通过自注意力机制推断每个奖励的重要性权重。实验表明，在离线强化学习基准测试中的复杂导航、运动和机器人操作任务上，偏好变换器显著优于当前的偏好建模方法。此外，我们观察到，所学习的偏好注意力层确实能够捕捉对人类决策至关重要的事件。我们认为，偏好变换器对于将基于偏好的强化学习（以及其他人在回路学习方法）扩展到各种应用中至关重要。

未来方向

偏好变换器还有若干未来研究方向。其中一个方向是将重要性权重用于强化学习或基于偏好的奖励学习。例如，可以利用重要性权重采样信息量更高的查询，从而提高基于偏好的奖励学习的反馈效率（Sadigh 等人，2017；Lee 等人，2021c）。另一个有趣的方向是利用重要性权重，通过加权更新来稳定Q学习（Kumar 等人，2020；Lee 等人，2021a）。与其他偏好模型相结合是未来研究的另一个重要方向。例如，Knox 等人 （2022）提出了一种基于简单网格世界环境中每个片段遗憾的新偏好模型。尽管他们提出的方法基于若干假设（例如，生成后继特征（Dayan，1993；Barreto 等人，2017）），但将其与基于遗憾的模型相结合将会是一个有趣的方向。

伦理声明

与其他领域（例如语言）不同，控制任务需要领域专家提供更多高质量的人类反馈。尽管可以通过训练标注员来提升数据集的质量，但这种训练需要付出大量努力。此外，来自众包平台（如亚马逊土耳其机器人）的反馈可能存在噪声。为应对这些问题，我们从非常熟悉机器人学、强化学习和目标任务的领域专家（作者）处收集了人类反馈。我们预计，我们的数据集带有高质量且干净的标签，并且我们的收集策略更贴近实际应用。然而与此同时，我们认为，使用公共众包平台进行评估将会很有意义，因此将其留待未来工作。

可复现性声明

我们在附录 B 中介绍了偏好变换器的实现细节，源代码可在摘要中的项目网站上获取。我们还将公开发布所收集的、带有真实人类偏好的离线数据集，以用于基准测试。

致谢与资金披露

我们感谢 Sihyun Yu、Subin Kim、Younggyo Seo 以及匿名评审，他们为改进本文提供了宝贵的反馈和建议。本研究获得了韩国政府科学技术信息通信部（MSIT）资助的信息通信技术规划评估院（IITP）项目支持（编号：2022-0-00953，具备问题解决能力的自主人工智能智能体；编号：2019-0-00075，人工智能研究生院项目（韩国科学技术院））。本材料所述工作还获得了 Google Cloud 研究额度计划提供的资助（项目编号：A696-P323-JC3B-RNWG）。

参考文献

Pieter Abbeel 和 Andrew Y Ng。通过逆强化学习进行模仿学习。发表于国际机器学习会议，2004年。

Riad Akrour、Marc Schoenauer 和 Michele Sebag。基于偏好的策略学习。发表于欧洲机器学习与数据库知识发现联合会议，2011。

Riad Akrour、Marc Schoenauer 和 Michèle Sebag。April：基于主动偏好学习的强化学习。载于欧洲机器学习与数据库知识发现联合会议，2012。

Jose A Arjona-Medina、Michael Gillhofer、Michael Widrich、Thomas Unterthiner、Johannes Brand- stetter 和 Sepp Hochreiter。Rudder：面向延迟奖励的回报分解。载于神经信息处理系统进展，2019年。

Jimmy Lei Ba、Jamie Ryan Kiros 和 Geoffrey E Hinton。层归一化。*arXiv* 预印本 *arXiv:1607.06450*，2016。

Fahiem Bacchus、Craig Boutilier 和 Adam Grove。行为奖励。载于全国人工智能会议，1996。Fahiem Bacchus、Craig Boutilier 和 Adam Grove。非马尔可夫决策过程的结构化求解方法。载于*AAAI*人工智能会议，1997。André Barreto、Will Dabney、Rémi Munos、Jonathan J Hunt、Tom Schaul、Hado P van Hasselt 和 David Silver。用于强化学习迁移的后继特征。载于神经信息处理系统进展，2017。Christopher Berner、Greg Brockman、Brooke Chan、Vicki Cheung、Przemysław D.ebiak、Christy Dennison、David Farhi、Quirin Fischer、Shariq Hashme、Chris Hesse 等。借助大规模深度强化学习玩 Dota 2。*arXiv* 预印本 *arXiv:1912.06680*，2019年。Ralph Allan Bradley 和 Milton E Terry。不完全区组设计的秩分析：I. 配对比较法。*Biometrika*，39(3/4):324–345，1952。Greg Brockman、Vicki Cheung、Ludwig Pettersson、Jonas Schneider、John Schulman、唐杰和 Wojciech Zaremba。OpenAI Gym。*arXiv* 预印本 *arXiv:1606.01540*，2016。Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少样本学习者。载于神经信息处理系统进展，2020。Lili Chen、Kevin Lu、Aravind Rajeswaran、Kimin Lee、Aditya Grover、Misha Laskin、Pieter Abbeel、Aravind Srinivas 和 Igor Mordatch。决策Transformer：通过序列建模实现强化学习。载于神经信息处理系统进展，2021。Paul F Christiano、Jan Leike、Tom Brown、Miljan Martic、Shane Legg 和 Dario Amodei。基于人类偏好的深度强化学习。载于神经信息处理系统进展，2017。C. Daniel、M. Viering、J. Metz、O. Kroemer 和 J. Peters。主动奖励学习。载于机器人学：科学与系统，2014。Sudeep Dasari 和 Abhinav Gupta。用于单样本视觉模仿的Transformer。载于机器人学习会议，2020。Peter Dayan。改进时序差分学习的泛化能力：后继表示。*Neural Computation*，5(4):613–624，1993。Yan Duan、Marcin Andrychowicz、Bradly Stadie、OpenAI 的 Jonathan Ho、Jonas Schneider、Ilya Sutskever、Pieter Abbeel 和 Wojciech Zaremba。单样本模仿学习。载于神经信息处理系统进展，2017。Joseph Early、Tom Bewley、Christine Evers 和 Sarvapali Ramchurn。通过可解释多实例学习，根据轨迹标签对非马尔可夫奖励进行建模。载于神经信息处理系统进展，2022。Aleksandra Faust、Anthony Francis 和 Dar Mehta。演化奖励以自动化强化学习。*arXiv* 预印本 *arXiv:1905.07628*，2019。Roy Frostig、Matthew James Johnson 和 Chris Leary。通过高级追踪编译机器学习程序。机器学习系统，4(9)，2018。Justin Fu、Anoop Korattikara、Sergey Levine 和 Sergio Guadarrama。从语言到目标：用于基于视觉的指令跟随的逆强化学习。*arXiv* 预印本 *arXiv:1902.07742*，2019。Justin Fu、Aviral Kumar、Ofir Nachum、George Tucker 和 Sergey Levine。D4RL：深度数据驱动强化学习数据集。*arXiv* 预印本 *arXiv:2004.07219*，2020。Tanmay Gangwani、Yuan Zhou 和 Jian Peng。通过轨迹空间平滑学习引导奖励。载于神经信息处理系统进展，2020。

Sepp Hochreiter 和 Jürgen Schmidhuber。长短期记忆。*Neural Computation*，9(8)：1735–1780，1997。Borja Ibarz、Jan Leike、Tobias Pohlen、Geoffrey Irving、Shane Legg 和 Dario Amodei。在 Atari 中从人类偏好和示范中学习奖励。载于神经信息处理系统进展，2018。

Donald Joseph Hejna III 和 Dorsa Sadigh。面向人类参与的强化学习的少样本偏好学习。载于机器人学习会议，2022。

Michael Janner、Qiyang Li 和 Sergey Levine。将强化学习视为一个大型序列建模问题。载于神经信息处理系统进展，2021。Daniel Kahneman。基于矩的评估：过去与未来。选择、价值与框架，第693–708页，2000。W Bradley Knox 和 Peter Stone。通过人类强化对智能体进行交互式塑造：TAMER 框架。载于国际知识获取大会，2009。

W BradleyKnox、StephaneHatgis-Kessell、SerenaBooth、Scott Niekum、PeterStone 和 Alessan- dro Allievi。用于学习奖励函数的人类偏好模型。*arXiv* 预印本 *arXiv:2206.02231*，2022。

Ilya Kostrikov、Ashvin Nair 和 Sergey Levine。使用隐式 Q 学习的离线强化学习。发表于国际学习表征会议，2022。

Aviral Kumar、Abhishek Gupta 和 Sergey Levine。DisCor：通过分布校正实现强化学习中的纠正反馈。发表于神经信息处理系统进展，2020。

Kimin Lee、Michael Laskin、Aravind Srinivas 和 Pieter Abbeel。Sunrise：用于深度强化学习中集成学习的简易统一框架。发表于国际机器学习大会，2021a。

Kimin Lee、Laura Smith 和 Pieter Abbeel。Pebble：通过经验重标记和无监督预训练实现反馈高效的交互式强化学习。发表于国际机器学习大会，2021b。

Kimin Lee、Laura Smith、Anca Dragan 和 Pieter Abbeel。B-pref：基准测试基于偏好的强化学习。发表于神经信息处理系统大会数据集与基准赛道（第一轮），2021c。

Xinran Liang、Katherine Shu、Kimin Lee 和 Pieter Abbeel。基于偏好的强化学习中用于探索的奖励不确定性。发表于国际学习表征会议，2022。Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。发表于国际学习表征会议，2019。Zhao Mandi、Fangchen Liu、Kimin Lee 和 Pieter Abbeel。迈向更具泛化能力的单次视觉模仿学习。发表于国际机器人与自动化会议，2022。

Ajay Mandlekar、Danfei Xu、Josiah Wong、Soroush Nasiriany、Chen Wang、Rohun Kulkarni、Li Fei- Fei、Silvio Savarese、Yuke Zhu 和 Roberto Martín-Martín。离线人类示范对于机器人操作学习究竟有何重要性。*arXiv* 预印本 *arXiv:2108.03298*，2021。

Volodymyr Mnih、Koray Kavukcuoglu、David Silver、Andrei A Rusu、Joel Veness、Marc G Belle-mare、Alex Graves、Martin Riedmiller、Andreas KFidjeland、GeorgOstrovski 等。通过深度强化学习实现人类水平控制。《自然》，518(7540):529，2015。

Suraj Nair、Eric Mitchell、Kevin Chen、Brian Ichter、Silvio Savarese 和 Chelsea Finn。从离线数据和众包标注中学习语言条件机器人行为。发表于机器人学习大会，2022。

Reiichiro Nakano、Jacob Hilton、Suchir Balaji、Jeff Wu、Long Ouyang、Christina Kim、Christo-pher Hesse、Shantanu Jain、Vineet Kosaraju、William Saunders 等。WebGPT：通过人类反馈进行浏览器辅助问答。*arXiv* 预印本 *arXiv:2112.09332*，2021。Andrew Y Ng、Stuart J Russell 等。逆强化学习算法。发表于国际机器学习大会，2000。Long Ouyang、Jeff Wu、Xu Jiang、Diogo Almeida、Carroll L Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray 等。通过人类反馈训练语言模型遵循指令。*arXiv* 预印本 *arXiv:2203.02155*，2022。Emilio Parisotto、Francis Song、Jack Rae、Razvan Pascanu、Caglar Gulcehre、Siddhant Jayakumar、Max Jaderberg、Raphael Lopez Kaufman、Aidan Clark、Seb Noury 等。稳定用于强化学习的 Transformer。发表于国际机器学习大会，2020。Jongjin Park、Younggyo Seo、Jinwoo Shin、Honglak Lee、Pieter Abbeel 和 Kimin Lee。SURF：通过数据增强实现半监督奖励学习，以提升基于偏好的强化学习的反馈效率。发表于国际学习表征会议，2022。Xue Bin Peng、Erwin Coumans、Tingnan Zhang、Tsang-Wei Lee、Jie Tan 和 Sergey Levine。通过模仿动物学习敏捷的机器人运动技能。发表于机器人学：科学与系统，2020。Patrick M Pilarski、Michael R Dawson、Thomas Degris、Farbod Fahimi、Jason P Carey 和 Richard S Sutton。通过演员-评论家强化学习在线训练人类操作的肌电假肢控制器。发表于国际康复机器人会议，2011。Alec Radford、Karthik Narasimhan、Tim Salimans 和 Ilya Sutskever。通过生成式预训练提升语言理解能力，2018。Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Radford、Mark Chen 和 Ilya Sutskever。零样本文本到图像生成。发表于国际机器学习大会，2021。Scott Reed、Konrad Zolna、Emilio Parisotto、Sergio Gomez Colmenarejo、Alexander Novikov、Gabriel Barth-Maron、Mai Gimenez、Yury Sulsky、Jackie Kay、Jost Tobias Springenberg 等。通才智能体。*arXiv* 预印本 *arXiv:2205.06175*，2022。Zhizhou Ren、Ruihan Guo、Yuan Zhou 和 Jian Peng。通过随机化回报分解学习长期奖励重新分配。发表于国际学习表征会议，2022。Dorsa Sadigh、Anca D Dragan、Shankar Sastry 和 Sanjit A Seshia。主动式基于偏好的奖励函数学习。发表于机器人学：科学与系统，2017。C. Schenck 和 D. Fox。液体倾倒的视觉闭环控制。发表于国际机器人与自动化会议，2017。Daniel Shin 和 Daniel S Brown。离线基于偏好的学徒学习。*arXiv* 预印本 *arXiv:2107.09251*，2021。David Silver、Julian Schrittwieser、Karen Simonyan、Ioannis Antonoglou、Aja Huang、Arthur Guez、Thomas Hubert、Lucas Baker、Matthew Lai、Adrian Bolton 等。在没有人类知识的情况下精通围棋。《自然》，550(7676):354，2017。Nisan Stiennon、Long Ouyang、Jeff Wu、Daniel M Ziegler、Ryan Lowe、Chelsea Voss、Alec Radford、Dario Amodei 和 Paul Christiano。通过人类反馈学习摘要生成。*arXiv* 预印本 *arXiv:2009.01325*，2020。Richard S Sutton 和 Andrew G Barto。《强化学习：导论》。麻省理工学院出版社，2018。Yuval Tassa、Tom Erez 和 Emanuel Todorov。通过在线轨迹优化合成并稳定复杂行为。发表于国际智能机器人与系统会议，2012。

Emanuel Todorov、Tom Erez 和 Yuval Tassa。Mujoco：用于基于模型的控制的物理引擎。载于国际智能机器人与系统会议，2012。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。Attention is all you need。载于神经信息处理系统进展，2017。

Mudit Verma 和 Katherine Metcalf。用于从人类偏好中进行奖励学习的符号引导的事后先验。*arXiv* 预印本 *arXiv:2210.09151*，2022。

Oriol Vinyals、Igor Babuschkin、Wojciech M Czarnecki、Michael Mathieu、Andrew Dudzik、Juny-oung Chung、David H Choi、Richard Powell、Timo Ewalds、Petko Georgiev 等。通过多智能体强化学习在星际争霸II中达到特级大师水平。《自然》，575(7782):350–354，2019年。

Aaron Wilson、Alan Fern 和 Prasad Tadepalli。基于轨迹偏好查询进行策略学习的贝叶斯方法。载于神经信息处理系统进展，2012年。

Jeff Wu、Long Ouyang、Daniel M Ziegler、Nissan Stiennon、Ryan Lowe、Jan Leike 和 Paul Chris- tiano。利用人类反馈递归式地总结书籍。*arXiv* 预印本 *arXiv:2109.10862*，2021年。

A. Yahya、A. Li、M. Kalakrishnan、Y. Chebotar 和 S. Levine。通过分布式异步引导式策略搜索进行集体机器人强化学习。载于智能机器人与系统国际会议，2017年。

Vinicius Zambaldi、David Raposo、Adam Santoro、Victor Bapst、Yujia Li、Igor Babuschkin、Karl Tuyls、David Reichert、Timothy Lillicrap、Edward Lockhart、Murray Shanahan、Victoria Langston、Razvan Pascanu、Matthew Botvinick、Oriol Vinyals 和 Peter Battaglia。具有关系归纳偏置的深度强化学习。载于国际学习表征会议，2019年。

Ruohan Zhang、Calen Walshe、Zhuode Liu、Lin Guan、Karl Muller、Jake Whritner、Luxin Zhang、Mary Hayhoe 和 Dana Ballard。Atari-head：Atari 人类眼动追踪与演示数据集。发表于*AAAI*人工智能会议，2020。

Yuke Zhu、Josiah Wong、Ajay Mandlekar 和 Roberto Martín-Martín。Robosuite：用于机器人学习的模块化模拟框架和基准。*arXiv* 预印本 *arXiv:2009.12293*，2020。

A 任务与数据集

本节将介绍 D4RL 基准测试中的控制任务详情（Fu 等人，2020）。

AntMaze。AntMaze 是一项导航任务，要求 Mujoco 蚂蚁机器人到达目标位置。数据集由经过预训练、旨在让机器人在不同迷宫布局中到达目标的策略生成。我们考虑两种迷宫布局：中等规模 和大型。数据集通过两种策略生成：多样化和游玩。多样化 数据集由预训练策略生成，起点和目标位置均经过随机化。游玩 数据集则由预训练策略生成，使用特定的人工挑选目标位置。任务奖励采用稀疏设定：只有当与目标位置的距离小于固定阈值时才给予奖励，否则奖励均为零。

Gym-Mujoco 运动。Gym-Mujoco 运动任务的目标是控制模拟机器人（Walker2d、Hopper），使其向前移动，同时尽量降低能量成本（动作范数），以实现安全行为。我们考虑两种数据生成策略：中等水平-专家和中等水平-回放。中等水平-专家 数据集由等量的专家演示和次优（部分训练）演示混合生成；中等水平-回放 数据集则对应于由部分训练策略收集的回放缓冲区。任务奖励由躯干的前进速度、控制惩罚和存活奖励构成。

Robosuite 机器人操作。Robosuite 机器人操作（Zhu 等人，2020）包含多种使用不同类型 7自由度模拟 Hand 机器人的任务。实验中，我们使用由 Franka Emika 的 Panda 机器人构成的模拟环境。我们选择了抬起一个立方体物体（*lift*）以及将一罐可乐从桌面放入目标箱中的任务（*can*）。数据集由一名熟练的遥操作员（*ph*）或 6 名熟练程度各异的遥操作员（*mh*）收集。任务奖励定义为稀疏奖励，具体细节请参见原论文。

B 实验详情

训练详情。偏好变换器的实现可在以下地址获取：

https://github.com/csmile-1006/PreferenceTransformer

我们的模型基于一个公开可用的 GPT JAX 重新实现版本（Frostig 等人，2018）[8]实现。所有实验均使用表2 中的超参数。所有实验中，我们都使用长度为 100 的片段。为提高使用学习得到的奖励函数训练 IQL 时的稳定性，我们对学习得到的奖励进行如下归一化：

$$\texttt{predicted reward} = \texttt{max timestep} \times \frac{\texttt{learned reward} - (\texttt{max returns})}{(\texttt{max returns}) - (\texttt{min returns})},$$

其中，最大时间步是回合的最大长度，而最大时间步和最小时间步分别表示数据集中最佳轨迹和最差轨迹的回报。训练时间因环境而异，但在 AntMaze 环境中使用 1000 条人类反馈进行奖励学习耗时不到 10 分钟，而使用学得的偏好变换器训练 IQL 则约需一小时。训练和评估模型时，我们使用一块 NVIDIA GeForce RTX 2080 Ti GPU 和 8 个 CPU 核心（Intel Xeon CPU E5-2630 v4 @ 2.20GHz）。我们使用 8 个随机种子训练奖励函数和 IQL。

偏好变换器的实现细节。在所有实验中，我们使用具有一层和四个自注意力头的因果变换器，后接一个具有单个自注意力头的双向自注意力层。偏好变换器使用 AdamW 优化器（Loshchilov 和 Hutter，2019年）进行训练，学习率为 1×10^{-4}，其中包含总梯度步数 5% 的线性预热步数、余弦学习率衰减、权重衰减 1×10^{-4}，批量大小为 256。强化学习训练使用公开发布的 IQL 实现[9]，并遵循原始的超参数设置。所有实验均使用与原始 IQL 相同的超参数。

评估详情。 我们遵循 IQL 在评估中的原始设置。我们在 AntMaze 中进行 100 次滚动轨迹评估，在 Gym-Mujoco 运动中进行 10 次滚动轨迹评估，在 Robosuite 机器人操作中进行 50 次滚动轨迹评估。对于评估指标，我们采用 D4RL 基准中提出的专家归一化分数

[8]https://github.com/matthias-wright/flaxmodels[9]https://github.com/ikostrikov/implicit_q_学习

：AntMaze 和 Gym-Mujoco 运动任务采用归一化得分 $= 100 \times \frac{\texttt{score} - \texttt{random score}}{\texttt{expert score} - \texttt{random score}}$，Robosuite 机器人操作任务采用成功率。

超参数。偏好变换器的超参数见表2。我们指出，增加偏好变换器的注意力层数和梯度更新次数可以提升模型性能，但为了加快训练和评估，我们使用当前这个层数较少的版本。

表 2：偏好变换器的超参数。

超参数	值
层数	1
注意力头数	4
嵌入维度	256
（因果变换器，偏好注意力层）	
批量大小	256
丢弃率（嵌入、注意力、残差连接）	0.1
学习率	0.0001
优化器	AdamW（Loshchilov 和 Hutter，2019年）
优化器动量	$\beta_1 = 0.9, \beta_2 = 0.99$
权重衰减	0.0001
预热步数	500
总梯度步数	10K

基线方法。 基线方法的实现细节如下：

- **马尔可夫奖励模型（MR）**。我们使用双层多层感知机，每层的隐藏维度为256。层间使用 ReLU激活函数作为激活函数，输出层则不使用激活函数。每个模型都通过优化公式 2 中定义的交叉熵损失进行训练，学习率为0.0003。

- **非马尔可夫奖励模型（NMR）**。对于非马尔可夫奖励模型，我们使用 JAX 重新实现 CSCInstance Space LSTM（Early 等人，2022），并遵循原论文中规定的架构。我们将隐藏维度加倍，以与 NMR 和 PT 保持一致。

Hopper后空翻的实现细节。 为了实现 Hopper后空翻，我们遵循Lee 等人（2021b）的做法，在在线设置中训练智能体。我们预训练策略 π，在最初的10,000个时间步中使用内在奖励进行探索并收集多样化的经验。随后，我们使用收集到的行为训练偏好变换器，并利用学到的奖励模型对回放缓冲区中的过往经验进行重标记。奖励学习和重标记阶段每经过10,000个时间步重复一次，每个阶段提供100次人类反馈查询。我们发现，3个奖励学习阶段（即300次查询）足以完成双后空翻。偏好变换器使用3层和8个注意力头，并在每个阶段训练100轮次。其他细节与表 2相同。

C 人类偏好

偏好收集。 我们从熟悉机器人任务的实际人类受试者（作者）那里收集反馈。具体而言，接受了各项任务相关指导的人类教师会观看呈现每个轨迹片段的视频，并选择哪一个片段更有助于智能体实现目标。每个轨迹片段长 3 秒（100 个时间步）。如果人类教师无法判断哪个片段更受偏好，则可以选择中性选项，对两个片段给予相同的偏好。

给人类教师的指导。

- AntMaze：首要任务是让蚂蚁机器人尽快到达目标位置，同时避免徘徊或跌落。如果蚂蚁机器人一直处于跌落、悬停或朝相反方向移动的状态，

相反 ite directi 继续前往 t h e goa l location, l 降低你的 pr iori ty to t h e segment even if the di

朝向目标的姿态更接近目标方向的那一段。如果两个机器人在这一指标上几乎持平，则根据机器人移动的距离来选择片段。

• Hopper：Hopper 机器人旨在尽可能向右移动，同时尽量降低能量消耗。如果 Hopper 机器人落地不稳，即使它在某一片段中移动的距离比另一个机器人更长，也应降低该片段的优先级。如果两个机器人在这一指标上几乎持平，则根据机器人移动的距离来选择片段。

• Walker2d：Walker2d 机器人的目标是尽可能向右移动，同时尽量降低能量消耗。如果 Walker2d 即将摔倒或步态异常（例如只用一条腿行走、打滑等），即使它在某一片段中移动的距离比另一个机器人更长，也应降低该片段的优先级。如果两个机器人在这一指标上几乎持平，则根据机器人移动的距离来选择片段。

• Robosuite：Panda 机械臂首先抓取物体，然后执行特定目的的动作（抬起一个立方体，或将一罐可乐移至目标箱）。优先选择两者中抓取到物体的一方。如果两台机器人在这一指标上的表现几乎持平，则根据目标物体移动的距离进行选择。

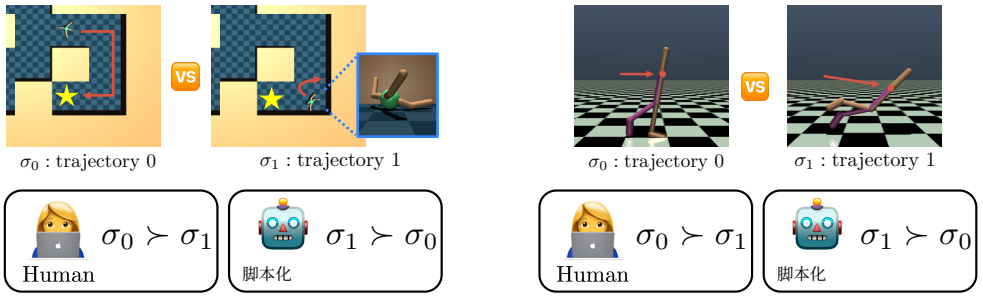
D 脚本教师的完整实验结果

表 3：在不同奖励函数下，IQL 在 AntMaze、Gym-Mujoco 运动任务上的平均归一化得分，以及在 Robosuite 操作任务上的成功率。我们使用来自脚本教师和真实人类教师的同一偏好数据集，训练偏好变换器（PT）和标准偏好建模方法；后者采用由多层感知机建模的马尔可夫奖励（MR），或由长短期记忆网络建模的非马尔可夫奖励（NMR）。结果显示了 8 次运行的平均值和标准差。

数据集	IQL（使用任务奖励）	IQL（使用偏好学习）					
		脚本教师			人类教师		
		MR	NMR	偏好变换器（我们的方法）	MR	NMR	偏好变换器（我们的方法）
antmaze-medium-play-v2	73.88 ± 4.49	64.75 ± 5.23	66。13 ± 3.36	66。13 ± 3。36	31.13 ± 16.96	62.88 ± 5.99	70.13 ± 3.76
antmaze-medium-diverse-v2	68.13 ± 10.15	4.25 ± 6.27	67.00 ± 5.90	68.13 ± 4.88	19.38 ± 9.24	20.13 ± 17.12	65.25 ± 3.59
antmaze-large-play-v2	48.75 ± 4.35	21.00 ± 14.23	10.75 ± 1.98	23.13 ± 13.10	24.25 ± 14.03	14.13 ± 3.60	42.38 ± 9.98
antmaze-large-diverse-v2	44.38 ± 4.47	15.75 ± 6.32	12.00 ± 3.59	44.50 ± 5.90	5.88 ± 6.94	0.00 ± 0.00	19.63 ± 3.70
蚂蚁迷宫-v2 平均值	58.79	26.31	38.97	50.00	20.16	24.29	49.35
Hopper 中等回放 v2	83.06 ± 15.80	62.77 ± 9.36	72.33 ± 0.01	94.19 ± 6.08	11.56 ± 30.27	57.88 ± 40.63	84.54 ± 4.07
Hopper 中等专家 v2	73.55 ± 41.47	80.00 ± 33.06	85.97 ± 37.91	39.14 ± 29.33	57.75 ± 23.70	38.63 ± 35.58	68.96 ± 33.86
walker2d-medium-replay-v2	73.11 ± 8.07	65.69 ± 8.17	73.63 ± 7.35	77。08 ± 9.84	72。07 ± 1.96	77.00 ± 3.03	71.27 ± 10.30
walker2d-medium-expert-v2	107.75 ± 2.02	109.95 ± 0.54	110.41 ± 0.82	109.99 ± 0.63	108。32 ± 3.87	110.39 ± 0.93	110.13 ± 0.21
locomotion-v2 平均值	84.37	79.60	85.56	80.10	62.43	70.98	83.72
lift-ph	96.75 ± 1.83	95。75 ± 2.71	80.00 ± 19.18	92.50 ± 4.24	84.75 ± 6.23	91.50 ± 5.42	91.75 ± 5.90
lift-mh	86.75 ± 2.82	92.25 ± 4.83	91.50 ± 5.42	93.00 ± 3.55	91.00 ± 4.00	90.75 ± 5.75	86.75 ± 5.95
can-ph	74.50 ± 6.82	62.25 ± 12.02	67.75 ± 6.80	69。75 ± 9.04	68.00 ± 9.13	62.00 ± 10.90	69.75 ± 5.89
can-mh	56.25 ± 8.78	59。00 ± 1.10	54.25 ± 5.57	55.25 ± 6.65	47.50 ± 3.51	30.50 ± 8.73	50.50 ± 6.48
Robosuite 平均值	78.56	77.31	73.37	77.63	72.81	68.69	74.68

E 查询的更多示例

在图8中，我们提供了更多查询示例，以突出人类教师与脚本教师之间的差异。在图8a中的示例中，脚本教师更偏好轨迹 1，因为蚂蚁离目标更近。然而，人类教师更偏好轨迹 0，因为蚂蚁在轨迹 1 中翻了个身。在图 8b中的示例中，脚本教师表现出非常短视的判断，因为手工设计的奖励无法捕捉行为的上下文（例如，跌倒）。这些示例表明，脚本教师缺乏权衡多重目标的能力，而人类则能通过平衡多重目标提供更合理的反馈。



(a) AntMaze

(b) Walker2d

图 8：来自 (a) antmaze-large-play-v2 和 (b) walker2d-medium-replay-v2 的轨迹示意图。由于人工设计的任务奖励并未包含所有与任务相关的信息，脚本教师会忽略翻转和跌倒等重要事件。

F 视频示例说明

补充材料中提供了视频示例，并对学到的重要性权重进行了可视化。为了便于观察，我们将权重表示为视频边缘上的颜色图，即，使用更明亮的颜色突出显示权重较高的帧。我们观察到，学到的重要性权重与人类意图高度一致，如图 3所示。

G 使用非马尔可夫模型的实验结果

我们研究了非马尔可夫奖励函数与非马尔可夫策略和价值函数结合时，是否能取得更好的性能。对于每个时间步 t，我们通过拼接历史信息 s_t 来扩充状态[10]，，并将其作为输入来训练策略和价值函数。表 4展示了偏好变换器与 NMR 的对比结果。我们观察到，PT 的表现优于 NMR，这再次表明了我们方法的优越性。然而，与表1中采用马尔可夫策略和价值函数的结果相比，所有方法的整体性能都有所下降。我们认为，这是因为原始的离线强化学习算法（IQL）基于马尔可夫设定，并在这一假设下调整了所有超参数。

表 4：在 AntMaze 和 Gym-Mujoco 运动任务中，采用不同奖励函数的非马尔可夫决策过程 IQL 的平均归一化得分。使用来自真实人类教师的同一偏好数据集，我们训练了偏好变换器（PT）和基于 LSTM 的非马尔可夫奖励（NMR；Early 等，2022）。结果展示了 8 次运行的平均值和标准差。

数据集	采用偏好学习的 IQL	
	NMR	PT（本文方法）
antmaze-medium-play-v2	54.50 ± 4.59	64.38 ± 3.16
antmaze-medium-diverse-v2	10.50 ± 13.60	53.00 ± 21.56
antmaze-large-play-v2	14.83 ± 5.78	25.63 ± 4.78
antmaze-large-diverse-v2	0.00 ± 0.00	8.13 ± 3.27
antmaze-v2 总计	19.96	37.79
Hopper 中等回放 v2	15.80 ± 26.11	55.77 ± 31.09
Hopper 中等专家 v2	55.80 ± 34.34	79.84 ± 29.32
walker2d-medium-replay-v2	41.91 ± 32.58	67.06 ± 13.39
walker2d-medium-expert-v2	94.20 ± 29.41	104.97 ± 10.63
locomotion-v2 总计	51.92	76.91

[10]具体而言，我们将上一个时间步的预测奖励 $\hat{r}_{t-1}$ 作为额外输入。