强化学习中的变压器概览

李文哲[1] *lwz21@mails.tsinghua.edu.cn*
Hao Luo[2,3*] *lh2000@pku.edu.cn*
林子川[4] *zichuanlin@tencent.com*
张崇杰[5] *chongjie@wustl.edu*
Zongqing Lu[2,3] *zongqing.lu@pku.edu.cn*
Deheng Ye[4] *dericye@tencent.com*

[1] *清华大学* [2] *北京大学* *3BAAI*
4Tencent Inc.
5圣路易斯华盛顿大学

摘要

变形器一直被认为是 NLP 和 CV 领域的主流神经架构，主要是在有监督的环境下使用。最近，强化学习（RL）领域也出现了使用变形器的类似热潮，但它面临着 RL 特性所带来的独特设计选择和挑战。然而，变形金刚在 RL 中的演化尚未得到很好的揭示。在本文中，我们试图系统回顾在 RL 中使用变形器的动机和进展，对现有作品进行分类，讨论各个子领域，并总结未来前景。

1 导言

强化学习（RL）为连续决策提供了一种数学形式。利用 RL，我们可以自动获得智能行为。RL 为基于学习的控制提供了一个通用框架，而深度神经网络作为一种具有高容量的函数逼近方式，已在多个领域取得了重大进展（Silver 等人，2016 年；Vinyals 等人，2019 年；Ye 等人，2020a;b）。

虽然深度强化学习（DRL）的通用性近年来取得了巨大发展，但样本效率问题阻碍了它在现实世界中的广泛应用。为了解决这个问题，一种有效的机制是在 DRL 框架中引入归纳偏差。DRL 中一个重要的归纳偏差是*函数逼近器架构的选择，*例如 DRL 代理的神经网络参数化。然而，与监督学习（SL）中的架构设计相比，如何为 DRL 设计架构的探索仍然较少。现有的大多数 RL 架构设计工作都是受（半）监督学习领域成功经验的启发。例如，在 DRL 中处理基于图像的高维输入的常见做法是引入卷积神经网络（CNN）（LeCun 等人，1998 年；Mnih 等人，2015 年）；处理部分可观测性的另一种常见做法是引入循环神经网络（RNN）（Hochreiter & Schmidhuber，1997 年；Hausknecht & Stone，2015 年）。

近年来，Transformer 架构（Vaswani 等人，2017 年）在广泛的 SL 任务中彻底改变了学习范式（Devlin 等人，2018 年；Dosovitskiy 等人，2020 年；Dong 等人，2018 年），并显示出优于 CNN 和 RNN 的性能。在

其显著优势中，Transformer 架构能够对长依赖关系建模，并具有出色的可扩展性（Khan 等人，2022 年）。受 SL 成功的启发，人们对在 RL 中应用变换器产生了浓厚的兴趣，希望将变换器的优势带到 RL 领域。

Transformers 在 RL 中的应用可以追溯到 Zambaldi 等人（2018 年），他们将自我注意机制用于结构化状态表示的关系推理。之后，许多研究人员试图将自我注意应用于表征学习，以提取实体间的关系，从而实现更好的策略学习（Vinyals 等人，2019 年；Baker

[*] 平等贡献； ψ 平等建议。

等人，2019 年）。除了利用变换器进行状态表征学习外，之前的研究还利用变换器捕捉多步时间依赖关系，以解决部分可观测性问题（Parisotto 等人，2020；Parisotto & Salakhutdinov，2021）。最近，离线 RL（Levine 等人，2020 年）因其利用大规模离线数据集的能力而备受关注。在离线 RL 的推动下，最近的研究表明，Transformer 架构可直接作为顺序决策模型（Chen 等人，2021 年；Janner 等人，2021 年），并可推广到多个任务和领域（Lee 等人，2022 年；Carroll 等人，2022 年）。

本调查旨在介绍*强化学习中*的*变形器*领域，简称为 TransformRL。尽管目前大多数强化学习研究（Devlin 等人，2018 年；Dosovitskiy 等人，2020 年；Bommasani 等人，2021 年；Lu 等人，2021 年）都将 Transformer 视为基础模型，但 RL 界对它的探索仍然较少。事实上，与 SL 领域相比，在 RL 中使用变换器作为函数近似器面临着独特的挑战。首先，RL 代理的训练目标通常是当前策略的函数，这就导致了变形器学习过程中的非静态性。其次，现有的 RL 算法通常对训练过程中的设计选择高度敏感，包括网络架构和容量（Henderson 等人，2018 年）。第三，基于 Transformer 的体系结构通常具有较高的计算和内存成本，使得 RL 学习过程中的训练和推理成本都很高。例如，在用于视频游戏的人工智能中，训练性能与样本生成效率密切相关，而样本生成效率则受到 RL 策略网络和价值网络计算成本的限制（Ye 等人，2020a；Berner 等人，2019）。在本文中，我们试图对 TransformRL 进行全面概述，包括对当前方法和挑战进行分类。我们还讨论了未来前景，因为我们相信，TransformRL 领域将在释放 RL 潜在影响方面发挥重要作用，本调查可为那些希望利用其潜力的人提供一个起点。

本文结构如下。第 2 节介绍了 RL 和 Transformer 的背景，随后简要介绍了如何将两者结合在一起。在第 3 节中，我们介绍了 RL 中网络架构的演变，以及阻碍 Transformer 架构在 RL 中长期广泛应用的挑战。在第 4 节中，我们对 RL 中的变换器进行了分类，并讨论了具有代表性的现有方法。最后，我们在第 5 节中总结并指出了潜在的未来发展方向。

2 问题范围

2.1 强化学习

一般来说，强化学习（RL）考虑的是马尔可夫决策过程（MDP）中的学习 M = □*S,* A*, P, r, γ, ρ*$_0$ □，其中 S 和 A 分别表示状态空间和行动空间，*P* (*s*′ |*s*，*a*) 是过渡动力学，*r*(*s*，*a*) 是奖励函数，$\gamma \in (0, 1)$ 是贴现因子，ρ_0 是初始状态分布。通常，RL 的目标是学习一个策略 $\pi(a|s)$ ，以最大化预期贴现收益 $J(\pi) = \mathsf{E}_{\pi,P,\rho}\, 0\, [\, _t \gamma^t \mathsf{r}(s_t\, ,\, a_t\,)]$ 。要解决 RL 问题，我们需要处理两个不同的部分：学习表示状态和学习行动。学习表示状态

第一部分可受益于归纳偏差（例如，基于图像状态的 CNN 和用于非马尔可夫任务的 RNN）。第二部分可以通过行为克隆（BC）、无模型或基于模型的 RL 来解决。在下一部分中，我们将介绍与 RL 中的变压器进展相关的几个具体 RL 问题。

离线 RL。在离线 RL（Levine 等人，2020 年）中，代理在训练过程中不能与环境交互。相反，它只能访问由任意策略收集的静态离线数据集 D = {(*s, a, s*′ *, r*)} 。如果不进行探索，现代离线 RL 方法（

Fujimoto 等人，2019 年；Kumar 等人，2020 年；Yu 等人，2021 年 b）就会受到限制。
学习到的策略应接近数据分布，以避免可能导致高估的分布外行动。最近，与典型的基于值的方法并行，离线 RL 的一个流行趋势是通过监督学习的 RL（RvS）（Emmons 等人，2021 年）。

目标条件 RL。目标条件 RL（GCRL）将标准 RL 问题扩展到了目标增强设置，即代理旨在学习一个能达到多个目标的目标条件策略 $\pi(a|s,g)$。先前的研究提出使用各种技术，如事后重新标注（Andrychowicz et al. 等人，2015）和自我模仿学习（Ghosh 等人，2019），以提高 GCRL 的泛化和样本效率。GCRL 相当灵活，因为目标的选择多种多样。读者可参阅（Liu et al.

基于模型的 RL。与直接学习策略和价值函数的无模型 RL 不同，基于模型的 RL 学习环境的辅助动态模型。这种模型可直接用于规划（Schrit- twieser 等人，2020 年），或生成假想轨迹以扩大任何无模型算法的训练数据（Hafner 等人，2019 年）。学习模型并非易事，尤其是在大型环境或部分观测环境中，我们首先需要构建状态的表征。最近的一些方法建议使用潜在动态模型（Hafner 等人，2019 年）或价值模型（Schrittwieser 等人，2020 年）来应对这些挑战，并提高 RL 的采样效率。

2.2 变形金刚

Transformer（Vaswani 等人，2017 年）是对顺序数据建模最有效、可扩展的神经网络之一。Transformer 的关键理念是融入*自我关注*机制，从而捕捉到序列数据中的依赖关系。 $\in \mathrm{R}^{\}}$ 以高效的方式处理长序列。形式上，给定一个包含 n 个标记的序列输入 $)\ \mathbf{x}_i$ $^{d\ n}{}_{\mathrm{i}=1}$ 其中 d 是嵌入维度，自我关注层将每个令牌 $\mathbf{x}_i$ 映射为一个查询 $\mathbf{q}_i \in \mathrm{R}^d$ q，一个密钥 $\mathbf{k}_i \in \mathrm{R}^d$ k，以及一个值 $\mathbf{v}_i \in \mathrm{R}^d$ v 通过线性变换，其中 $d_q = d_k$。 让输入、查询、键和值的序列为 $\mathbf{X} \in \mathrm{R}^{n\times d}$，$\mathbf{Q} \in \mathrm{R}^{n\times d}$ q，$\mathbf{K} \in \mathrm{R}^{n\times d}$ k，和 $\mathbf{V} \in \mathrm{R}^{n\times d}$ v，分别。自注意层的输出 $\mathbf{Z} \in \mathrm{R}$ $\mathrm{v}^{n\times d}$ 是所有值的加权总和：

$$\mathbf{Z} = \text{软最大值}\frac{\mathrm{A}\ \ \mathrm{B}}{\sqrt{d\alpha}}\ \ \mathbf{V}$$

通过自我注意机制（Bahdanau 等人，2014 年）以及其他技术，如多头注意和残余连接（He 等人，2016 年），变形金刚可以学习表达性表征并建立长期互动模型。

由于具有强大的表示能力和出色的可扩展性，Transformer 架构在各种有监督和无监督学习任务中表现出了优于 CNN 和 RNN 的性能。因此，一个自然而然的问题是：我们能否使用变形器来解决 RL 中的问题（即学习表示状态和学习行动）？

2.3 变压器和 RL 的组合

我们注意到，越来越多的研究正在寻求以不同的方式将变换器与 RL 结合起来。一般来说，变换器可用作 RL 算法的一个组件，例如表示模块或动态模型。变换器还可以作为一个整体的顺序决策制定器。图 1 简要介绍了变形金刚在 RL 中的不同作用。

3 RL 中的网络结构

在介绍 TransformRL 当前方法的分类之前，我们首先回顾了 RL 中网络架构设计的早期进展，并总结了其面临的挑战。我们之所以这样做，是因为 Transformer 本身就是一种先进的神经网络，而设计适当的神经网络有助于 DRL 的成功。

transformers和rl结合的类别

3.1 函数近似器架构

自开创性的 Deep Q-Network 工作（Mnih 等人，2015 年）以来，人们在为 DRL 代理开发网络架构方面做出

了许多努力。RL 中网络架构的改进主要可分为两类。第一类是设计一种结合 RL 归纳偏差的新结构，以减轻训练策略或价值函数的难度。例如，Wang 等人（2016 年）提出了决斗网络架构，其中一个用于状态价值函数，另一个用于与状态相关的行动优势函数。这种架构的选择包含了归纳偏差，可将学习泛化到不同的行动中。其他例子还包括价值分解网络，它被用于学习单个代理的局部 Q 值（Sunehag 等人，2017 年）或子奖励（Lin 等人，2019 年）。第二类是研究神经网络的一般技术（如正则化、跳转连接、批量归一化）能否应用于 RL。仅举几例，Ota 等人（2020 年）发现，在使用在线特征提取器的同时增加输入维度，可以促进状态表示学习，从而提高性能和效率。

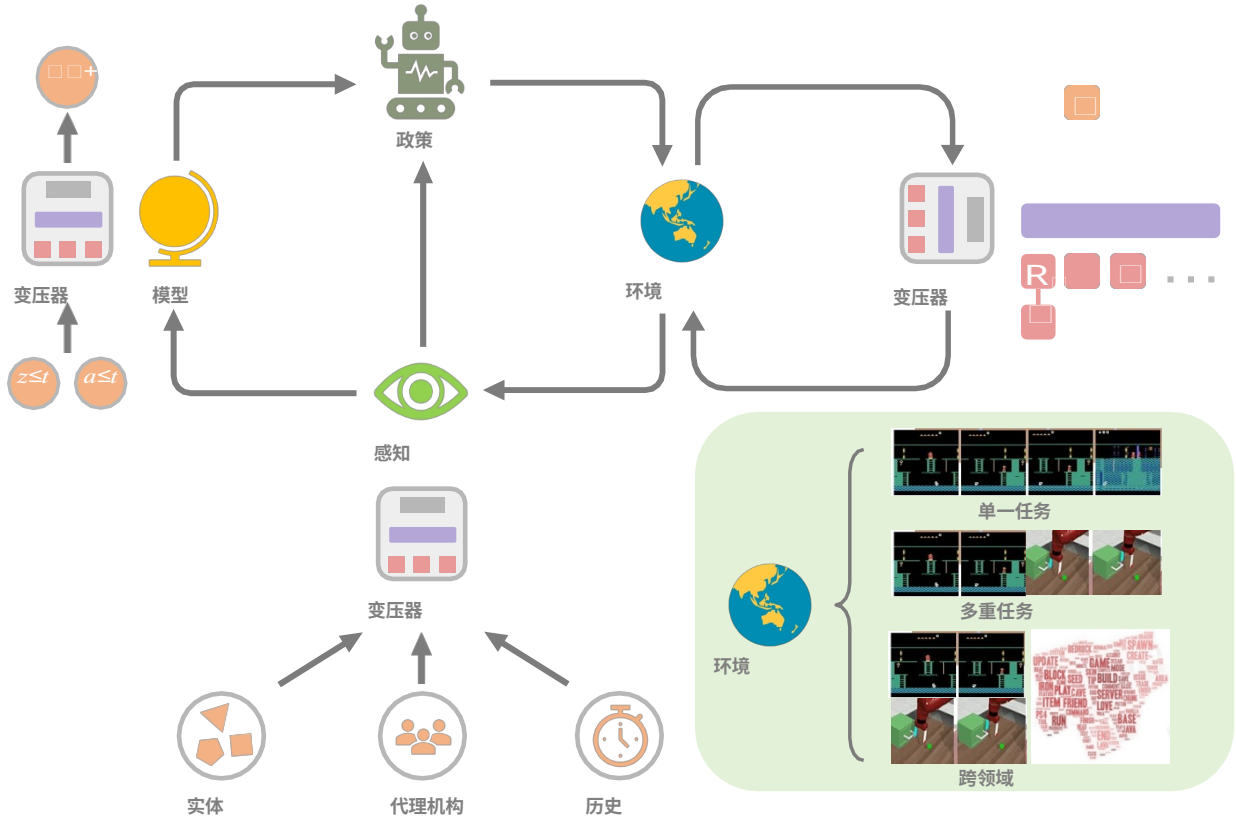

图 1：TransformRL 的示例。一方面，变换器可以作为 RL 的一个组件。特别是，变换器可以编码各种序列，如实体、代理和历史信息堆栈；它也是动态模型的一种表达式预测器。另一方面，Transformers 可以整合 RL 中的所有子程序，并充当顺序决策制定器。总之，Transformers 可以提高 RL 在单任务、多任务和跨领域环境中的学习效率。

DRL 算法的采样效率。Sinha 等人（2020）为 DRL 代理提出了一种深度密集架构，利用跳转连接实现高效学习，并通过归纳偏置来缓解数据处理不平等。Ota 等人（2021 年）使用具有解耦表示学习的 DenseNet（Huang 等人，2017 年）来改善大型网络的信息流和梯度。最近，由于 Transformers 的优越性能，一些研究人员尝试将 Transformers 架构应用于策略优化算法，但发现 vanilla Transformer 设计无法在 RL 任务中实现合理的性能（Parisotto 等人，2020 年）。

3.2 挑战

过去几年，变压器在 SL 领域的应用取得了飞速发展，但在 RL 领域应用变压器却并不简单，面临着以下独特的挑战。

一方面，从 RL 的角度来看，许多研究人员指出，现有的 RL 算法对深度神经网络的架构异常敏感（Henderson 等人，2018 年；Engstrom 等人，2019 年；Andrychowicz 等人，2020 年）。首先，RL 中交替进行数据收集和策略优化（即数据分布转移）的模式会在训练过程中诱发非平稳性。其次，RL 算法通常对训练过程中的设计选择高度敏感。特别是，当与引导和非政策学习相结合时，当值估计变得无限制时，函数近似值的学习可能会出现偏离（即 "致命三联征"）（Van Hasselt 等人，2018 年）。最近，Emmons 等人（2021 年）发现，仔细选择模型架构和正则化对 DRL 代理的性能至关重要。

另一方面，从变换器的角度来看，基于变换器的架构存在内存占用大和延迟高的问题，这阻碍了其高效部署和推理。最近，许多研究人员致力于提高原始 Transformer 的计算和内存效率（Tay 等人，2022 年），但这些工作大多集中在 SL 领域。在 RL 方面，Parisotto & Salakhutdinov（2021 年）提出将学习进度从基于 Transformer 的大型学习者模型提炼到小型行为者模型，以绕过 Transformer 的高推理延迟。然而，这些方法在内存和计算方面仍然很昂贵。迄今为止，RL 界尚未充分探讨高效或轻量级变换器的想法。

4 RL 中的变压器

尽管变换器已成为大多数监督学习研究的基础模型，但由于上述挑战，它在 RL 领域长期以来并未得到广泛应用。实际上，大多数早期的 TransformRL 尝试都是将 Transformer 用于状态表示学习或提供记忆信息，但仍然使用标准的 RL 算法，如代理学习的时差学习和策略优化。因此，这些方法仍然受到传统 RL 框架的挑战。直到最近，离线 RL 才允许从大规模离线数据中学习最优策略。受离线 RL 的启发，最近的研究进一步将 RL 问题视为固定经验上的条件序列建模问题，从而绕过了传统 RL 中引导误差的挑战，使 Transformer 能够释放其强大的序列建模能力。

在这篇调查报告中，我们回顾了 TransformRL 的进展，并提供了一个分类法来介绍当前的方法。我们将现有方法分为四类：表示学习、模型学习、顺序决策和通用代理。图 2 提供了分类简图和相应的子集。

4.1 表征学习的变压器

Transformers 的一种自然用法是将其用作序列编码器。事实上，RL 任务中的各种序列都需要处理，如局部每时间步序列（多实体序列（Vinyals 等人，2019 年；Baker 等人，2019 年）、多代理序列（Wen 等人，2022 年））、时间序列（轨迹序列（Parisotto 等人，2020 年；Banino 等人，2021 年））等等。

4.1.1 本地每时序列编码器

这种方法早期显著的成功之处在于用变换器（Transformers）处理来自分散在代理观察中的数量可变的实体的复杂信息。Zambaldi 等人（2018 年）首次提出要捕捉关系推理

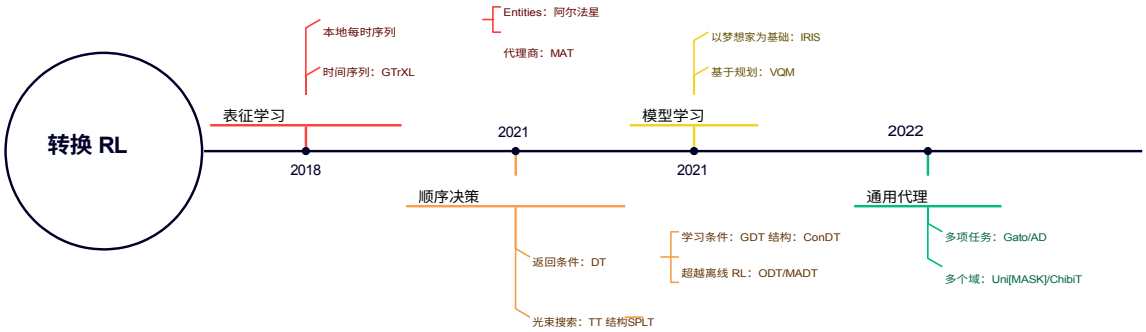

图 2：TransformRL 的分类法。时间轴基于与该分支相关的第一项工作。

随后，AlphaStar（Vinyals 等人，2019 年）将其用于处理具有挑战性的多代理游戏《星际争霸 II》中的多实体观察。拟议的实体转换器将观察结果编码为：

$$\text{Emb} = \text{Transformer}(e_1, \text{---}, e_i, \text{---}),$$

其中，e_i 表示代理对实体 i 的观察结果，可以直接从整个观察结果中切分，也可以由实体标记器给出。

一些后续研究丰富了实体转换器机制。Hu 等人（2020）提出了一种兼容的解耦策略，明确地将行动与各种实体相关联，并利用注意力机制进行策略解释。Wang 等人（2023b）通过迁移学习，在不同类型的游戏中学习具有常识和特征空间无关标记的实体转换器。为了解决具有挑战性的单次视觉模仿问题，Dasari 和 Gupta（2021 年）使用 Transformers 学习一种侧重于特定任务元素的表征。

与分散在观察中的实体类似，一些研究利用变形器处理其他局部每时序列。Tang 和 Ha（2021 年）利用注意力机制来处理感官序列，并构建了一种在输入时不变的策略。在不兼容的多任务 RL 设置中，Transformers 被提出用于提取形态学领域知识（Kurin 等人，2020 年）。关于本地每时观测中存在的多模态信息（如图像和语言），Team 等人（2021 年）利用基于变换器的结构来整合这些多模态信息并表示代理的状态。

此外，最近的 RL 算法正试图将视觉归纳偏差纳入策略学习。例如，视觉转换器（ViT）使用补丁序列来处理视觉领域的图像，可用于 RL 中的再现学习。Tao 等人（2022 年）测试了 ViT 及其变体与各种自监督技术（Data2vec、MAE 和动量对比学习）相结合在视觉控制任务中的有效性。然而，在他们对不太复杂的任务进行的实验中，并未显示出明显的性能提升。另一方面，Kalantari 等人（2022 年）使用 ViT 架构学习视觉输入的 Q 值，显示出其提高 RL 算法采样效率的潜力。此外，Seo 等人（2022a）将 ViT 与改进的特征掩码 MAE 结合起来，学习更适合动态的图像特征，这将有利于决策和控制。

4.1.2 时序编码器

同时，用变换器处理时序也是合理的。这样的时序编码器可作为存储模块使用：

$$\text{Emb}_{0:t} = \text{Transformer}(o_0, \text{---}, o_t),$$

其中，o_t 表示代理在时间步 *t 的观测值*，$\text{Emb}_{0:t}$ 表示从初始观测值到当前观测值的历史观测值的嵌入。

在早期的研究中，Mishra 等人（2018）未能用 vanilla Transformers 处理时序，并发现在某些任务下它甚至不如随机策略。门控变换器-XL（GTrXL）（Parisotto 等人，2020 年）是第一个使用变换器作为存储器的有效方案。GTrXL 通过身份映射重新排序，提供了一个门控 "跳过 "路径，以

从一开始就稳定训练程序。这种架构还可以结合语言指令来加速元 RL（Bing 等人，2022 年）和多任务 RL（Guhur 等人，2023 年）。此外，Loynd 等人（2020 年）提出了一种利用内存向量实现长期依赖性的快捷机制，而 Irie 等人（2021 年）则将线性变换器与快速加权编程器相结合，以获得更好的性能。此外，Melo（2022）提出使用自我注意机制来模仿记忆恢复，以实现基于记忆的元 RL。

虽然随着记忆范围的增长和参数规模的扩大，变换器的性能优于 LSTM/RNN，但它在 RL 信号方面的数据效率较低。后续研究利用一些辅助（自我）监督任务来提高学习效率（Banino 等人，2021 年），或使用预先训练好的 Transformer 作为时序编码器（Li 等人，2022 年；Fan 等人，2022 年）。

4.2 用于模型学习的变压器

除了使用变换器作为序列嵌入的编码器外，变换器结构还可作为基于模型算法的世界模型的支柱。有别于以单步目标和行动为条件的预测，变换器使世界模型能够以历史信息为条件预测过渡。

实际上，Dreamer 及其后续算法（Hafner 等人，2020；2021；2023；Seo 等人，2022b）的成功证明了以历史为条件的世界模型在部分可观测环境或需要记忆机制的任务中的优势。以历史为条件的世界模型包括一个用于捕捉抽象信息的观测编码器和一个用于学习潜空间过渡的过渡模型：

$$z_t \sim P_{\text{enc}}(z_t | o_t)、$$
$$z\hat{}t+1, r\hat{}t+1, \gamma\hat{}t+1 \sim \text{Ptrans}(z\hat{}t+1, r\hat{}t+1, \gamma\hat{}t+1 | z_{\leq t}, a_{\leq t})、$$

其中，z_t 表示观测值 o_t 的潜在嵌入，P_{enc} , P_{trans} 分别表示观测值编码器和转换模型。

在以往的研究中，有不少人尝试用变换器架构取代 RNN，建立以历史为条件的世界模型。具体来说，Chen 等人（2022 年）用基于变换器的模型（变换器状态空间模型，TSSM）取代了 Dreamer 中基于 RNN 的循环状态空间模型（RSSM）。IRIS（Imagination with auto- Regression over an Inner Speech）（Micheli等人，2022年）和TWM（Transformer-based World Model）（Robine等人，2023年）通过对推出经验的自动回归学习来学习基于Transformer的世界模型，而无需像Dreamer那样进行KL平衡，并在Atari（Bellemare等人，2013年）100k基准测试中取得了可观的成绩。

此外，一些研究还尝试将基于 Transformer 的世界模型与规划相结合。Ozair 等人（2021 年）验证了使用基于 Transformer 的世界模型进行规划的有效性，该模型可用于处理需要长时间战术前瞻的随机任务。Sun 等人（2022 年）提出了一种基于目标条件的 Transformer 世界模型，该模型在程序任务的可视化规划中非常有效。

诚然，RNN 和 Transformer 都能兼容以历史信息为条件的世界模型。然而，Micheli 等人（2022 年）和 Chen 等人（2022 年）发现，与 Dreamer 相比，Transformer 是一种数据效率更高的世界模型。事实上，虽然基于模型的方法具有数据效率高的特点，但它们存在着复合预测误差随模型展开长度增加而增加的问题，这极大地影响了性能并限制了模型展开长度（Janner 等人，2019 年）。基于变换器的世界模型可以帮助

减轻较长序列的预测误差。

4.3 用于顺序决策的变压器

除了作为一种可插入传统 RL 算法组件的富有表现力的架构之外，Trans- former 本身还可以作为一种直接进行序列决策的模型。这是因为 RL 可以被视为一个条件序列建模问题--生成一连串能产生高回报的行动。

4.3.1 变形金刚》是离线 RL 的里程碑

变压器在 RL 中广泛应用所面临的一个挑战是，训练过程中的非稳态性可能会阻碍变压器的优化。然而，近年来离线 RL 的蓬勃发展促使越来越多的研究工作聚焦于

方法	设置	后见之明信息	推论	附加结构/用途
DT （Chen 等人，2021 年）	离线	返程	调理	变压器基本结构
TT（Janner 等人，2021 年）	IL/GCRL/ 离线	返程	光束搜索	变压器基本结构
BeT (Shafiullah et al., 2022)	不列颠哥伦比亚省	无	调理	变压器基本结构
BooT（Wang 等人，2022 年）	离线	返程	光束搜索	数据扩增
全球数据传输（古田等人，2021 年）	HIM	任意	调理	反因果聚合器
ESPER（Paster 等人，2022 年）	离线（随机）	预期回报	调理	对抗性聚类
DoC (Yang et al., 2022a)	离线（随机）	学习到的表征	调理	附加潜值函数
QDT （Yamagata 等人，2022 年）	离线	重新命名的返程	调理	附加 Q 功能。
StARformer（Shang 等人，2022 年）	伊利诺伊州/离线	返程/奖励	调理	步进变压器和序列变压器
TIT（Mao 等人，2022 年）	在线/离线	返回/无	调理	内变压器和外变压器
ConDT （科南等人，2022 年）	离线	学习到的表征	调理	依赖回报的转变
SPLT （Villaflor 等人，2022 年）	离线	无	最小搜索	世界和政策的不同模式
德福格（Hu 等人，2023 年）	离线	返程	调理	下拉跨度嵌入
ODT （Zheng 等人，2022 年）	在线微调	返程	调理	基于轨迹的熵
MADT（Meng 等人，2021 年）	在线微调（多代理）	无	调理	演员和评论家分立模式

表 1：用于顺序决策的转换器汇总。

在离线数据上训练变换器模型，以达到最先进的性能。Decision Transformer（DT）（Chen 等人，2021 年）首先应用了这一思想，将 RL 建模为自回归生成问题，以生成所需的轨迹：

$$\tau = 1 R\hat{}1, s_1, a_{1\ ,}\ R\hat{}2, s_2, a_2, \ldots R\hat{}T, s_T, a_T^{\ 2}\ 、$$

其中 $R_t = \sum_{t=T}^{\ t'} r(s', a')$ 是返回值。在第一个时间步中，以适当的目标返回值为条件、DT 无需明确的 TD 学习或动态编程即可生成所需的行动。与此同时，轨迹变换器（TT）（Janner 等人，2021 年）采用了类似的变换器结构，但在执行过程中建议使用波束搜索进行规划。实证结果表明，TT 在长视距预判方面表现出色。此外，TT 还表明，只要对 vanilla beam search 稍作调整，TT 就能在同一框架下执行模仿学习、目标条件 RL 和离线 RL。关于行为克隆设置，Behavior Transformer（BeT）（Shafiullah等人，2022年）提出了与TT类似的Transformer结构，用于从多模态数据集中学习。

鉴于 Transformer 在序列预测方面的卓越准确性，Bootstrapped Transformer（BooT）（Wang 等人，2022 年）提出在优化 Transformer 以进行序列决策的同时，引导 Transformer 生成数据。通过引导变换器进行数据扩充，可以扩大离线数据集的数量和覆盖范围，从而提高性能。更具体地说，BooT 比较了不同的数据生成方案和引导方案，以分析 BooT 如何有利于政策学习。结果表明，它可以生成与底层 MDP 一致的数据，而无需额外的显式保守约束。

4.3.2 不同的调节方式

虽然以 "返程 "为条件是纳入未来轨迹信息的实用选择，但一个自然的问题是，其他类型的后见之明信息是否有利于顺序决策。为此，Furuta 等人（2021 年）提出了 "后见信息匹配"（Hindsight Information Matching，简称 HIM），这是一个统一的框架，可以制定后见 RL 问题的变体。更具体地说，HIM 将后视 RL 转换为将未来轨迹信息的任何预定义统计数据与所学条件策略诱导的分布进行匹配。此外，本研究还提出了适用于任意统计数据选择的广义 DT（GDT），并展示了其在两个 HIM 问题中的应用：离线多任务状态边际

匹配和模仿学习。

具体来说，在随机环境中，以 "从返回到出发 "为条件的一个缺点是会导致次优行动，因为训练数据可能包含次优行动，而这些行动由于过渡的随机性而幸运地获得了高回报。Paster 等人（2022 年）指出了一般 RvS 方法的这一局限性。他们进一步将 RvS 表述为一个 HIM 问题，并发现如果信息统计与过渡的随机性无关，RvS 策略可以在一致性上实现目标。基于这一含义，他们提出了与环境随机性无关的表示法（ESPER），这是一种首先对轨迹进行聚类并估计每个聚类的平均收益，然后以预期收益为条件训练策略的算法。另外，控制二分法（DoC）（Yang 等人，2022a）建议通过以下方法学习一种与环境中的随机转换和回报无关的表征

互信息最小化。在推理过程中，DoC 会选择具有最高值的表示，并将其输入条件策略。

Yang等人（2022b）提出了一种方法，即在训练集中标注特定任务的程序观察序列，并利用这些序列生成决策行动。这种方法能让代理在事先规划的基础上学习决策，这对需要多步骤预见的任务是有益的。

除了探索不同的后见之明信息外，另一种增强返回-去向调节的方法是增强数据集。Q-learning DT（QDT）（Yamagata 等人，2022 年）建议使用保守值函数来重新标注数据集中的返回-去向，从而将 DT 与动态编程结合起来，提高其缝合能力。

4.3.3 改进变压器的结构

除了研究不同的条件信息外，还有一些工作旨在改进 DT 的结构。为了解决视觉输入任务，State-Action-Reward Transformer（StARformer）（Shang等人，2022年）提出学习一个额外的Step Transformer，用于局部每时表示，并使用该表示进行序列建模。Transformer 中的 Transformer（TIT）（Mao 等人，2022 年）利用两个类似的 Transformer 结构分别处理观察和历史，并将它们串联起来作为在线和离线 RL 的骨干。Konan 等人（2022 年）认为，不同的子任务对应不同级别的返回，需要不同的标记化。因此，他们提出了对比决策转换器（Contrastive Decision Transformer，ConDT）结构，即在将状态和行动嵌入到因果转换器之前，对其进行依赖于返回的转换。与返回相关的变换直观地捕捉了当前子任务的特定特征，并通过辅助对比损失进行学习，以加强变换与返回之间的相关性。Villaflor 等人（2022 年）分析了在与 TT 相同的模型中实施模型预测和策略网络的一个不利因素。在具有长期规划的安全关键场景中，预测未来状态与做出行动决策之间的偏好往往是矛盾的。具体来说，需要在最坏的未来找到最佳行动，而这在一个模型中很难完成。因此，他们提出了 "潜在轨迹变换器"（SeParated Latent Trajectory Transformer，简称 SPLT 变换器），它由世界模型和策略模型两个独立的基于变换器的 CVAE 结构组成，以轨迹为条件。与最小搜索程序类似，SPLT 变换器在规划过程中搜索潜变量空间，以最小化世界模型中的收益-去向，最大化策略模型中的收益-去向。Hu 等人（2023 年）考虑了实际应用场景中可能出现的丢帧问题，即由于丢帧，某些时间步的状态和奖励无法使用，而之前时间步的信息则可以重复使用。他们提出了随机丢帧下的决策转换器（DeFog），通过引入丢帧跨度嵌入来扩展 DT 中的时间步嵌入。Kang 等人（2023 年）在决策转换器中引入了内部工作记忆模块，以解决隐式记忆机制导致的遗忘问题。他们还加入了低阶适应（LoRA）（Hu 等人，2022 年）参数，以适应未见任务。

4.3.4 将 DT 扩展到离线 RL 之外

尽管围绕用于顺序决策的变换器所做的大部分工作都集中在离线环境下，但也有一些人尝试将这一范例应用到在线和多代理环境中。在线决策转换器（ODT）（Zheng 等人，2022 年）将 DT 中的确定性策略替换为随机策略，并定义了轨迹级策略熵，以帮助在线微调过程中的探索。此外，这种两阶段范式（离线预训练与在线微调）也被应用于多代理决策转换器（MADT）（Meng 等人，2021 年），即从单个代理的角度，

用离线数据预训练十级化 DT，并用 MAPPO 作为在线微调的策略网络（Yu 等人，2021a）。

4.4 通用代理的变形金刚

鉴于决策转换器已经在离线数据的各种任务中大显身手，一些作品转而考虑转换器是否能让通用代理解决多种任务或问题，就像在 CV 和 NLP 领域一样。

4.4.1 适用于多种任务

一些研究借鉴了在 CV 和 NLP 大规模数据集上进行预训练的思路，试图从大规模多任务数据集中抽象出一种通用策略。多游戏决策转换器（MGDT）（Lee 等人，2022 年）是 DT 的一种变体，它在由专家数据和非专家数据组成的多样化数据集上学习 DT，并以单组参数在多个 Atari 游戏上实现了接近人类的性能。为了在包含非专家经验的数据集上获得专家级性能，MGDT 包含专家级行动推理机制，该机制可根据返回到目标的先验分布计算专家级返回到目标的后验分布，并根据贝叶斯公式计算预设的专家级返回到目标的似然比例。同样，Switch Trajectory Transformer（SwitchTT）（Lin 等人，2022 年）是 TT 的多任务扩展，它利用稀疏激活模型，用专家混合层取代 FFN 层，实现高效的多任务离线学习。此外，它还采用了分布式轨迹值估计器来模拟值估计的不确定性。有了这两项增强功能，SwitchTT 在多个任务的性能和训练速度方面都比 TT 有了提高。MGDT 和 SwitchTT 利用从多个任务和各种性能级策略中收集的经验来学习通用策略。Zhu 等人（2023 年）建立了一个多目标离线 RL 数据集，并将 DT 扩展到偏好和回报条件学习。然而，构建大规模多任务数据集并非易事。与 CV 或 NLP 中的大规模数据集不同，这些数据集通常是通过互联网上的海量公共数据和简单的人工标注构建的，而行动信息在公共的顺序决策数据中总是缺失的，并且不易标注。因此，Baker 等人（2022 年）提出了一种半监督方案，通过在一小部分有动作标签的数据上学习基于变换器的逆动态模型（IDM），来利用没有动作信息的大规模在线数据。IDM 是在包含人工标注动作的小规模数据集上学习的，其准确性足以为视频提供动作标签，从而实现有效的行为克隆和微调。此外，Venuto 等人（2022 年）在实验中使用其他任务的动作标签数据来训练 IDM，从而减少了对目标任务专用动作标签数据的需求。

提示（Brown 等人，2020 年）在适应新任务方面的功效已在 NLP 领域的许多先前研究中得到证实。秉承这一理念，一些研究成果旨在利用提示技术来实现基于 DT 方法的快速适应。基于提示的决策转换器（Prompt-DT）（Xu 等人，2022 年）从少量演示数据集中采样一系列转换作为提示，并证明它可以在离线元 RL 任务中实现少量策略泛化。Reed 等人（2022 年）进一步利用基于提示的架构，在一个涵盖自然语言、图像、时间决策和多模态数据的超大规模数据集上，通过自动回归序列建模学习了一个通用代理（Gato）。Gato 能够胜任不同领域的一系列任务，包括文本生成和决策。具体来说，Gato 将多模态序列统一在一个共享的标记化空间中，并在部署中调整基于提示的推理，以生成特定任务的序列。与 Gato 类似，RT-1（Brohan 等人，2022 年）和 PaLM-E（Driess 等人，2023 年）利用大规模多模态数据集来训练 Transformers，从而在下游任务中实现高性能。VIMA（Jiang 等人，2022 年）将文本和视觉标记作为多模态提示结合起来，建立了一个可扩展的模型，能在机器人操纵任务中很好地推广。

尽管有效，但拉斯金等人（2022 年）指出，基于提示的框架的一个局限性是，提示是从一个行为良好的策略中演示出来的，因为在这两项工作中，上下文不足以捕捉策略改进。受 Transformer 的上下文学习能力（Alayrac 等人，2022 年）的启发，他们提出了算法蒸馏（AD）（Laskin 等人，2022 年），即根据单任务 RL 算法学习进度的 跨集序列来训练 Transformer。因此，即使在新任务中，变换器也能在自动递归生成过程中学习逐步改进其策略。

4.4.2 推广到多个领域

除了适用于多种任务之外，Transformer 还是一个强大的 "通用 "模型，可以统一一系列与顺序决策相关的领域。受 NLP 中掩码语言建模（Devlin 等人，2018 年）技术进步的推动，Carroll 等人（2022 年）提出了 Uni[MASK]，它将各种常用研究领域（包括行为克隆、离线 RL、GCRL、过去/未来推理和动态预测）统一为一个掩码推理问题。Uni[MASK] 比较了不同的掩码方案，包括特定任务掩码、随机掩码和微调变体。结果表明，使用随机掩码训练的单个变换器可以解决任意推理任务。更令人惊讶的是，与特定任务的对应方案相比，随机掩码在单任务设置中仍能提高性能。

除了统一 RL 领域的顺序推理问题外，Reid 等人（2022 年）还发现，在语言数据集或包含语言模态的多模态数据集上预训练 Transformer，有利于微调 DT。具体来说，Reid 等人（2022 年）发现，用语言数据预训练 Transformer，同时鼓励语言和基于 RL 的表征之间的相似性，有助于提高 DT 的性能和收敛速度。这一发现意味着，即使是来自非 RL 领域的知识，也能通过 Transformer 从 RL 训练中获益。Li 等人（2022 年）的研究进一步表明，使用预训练语言模型初始化的策略可以进行微调，以适应交互决策中的不同任务和目标。此外，一些研究（Huang 等人，2022a;b；Raman 等人，2022；Yao 等人，2022；Ahn 等人，2022；Wang 等人，2023c；Du 等人，2023；Wang 等人，2023a）发现，预训练的大规模语言模型能够生成合理的高级计划，以完成复杂任务，而无需进一步微调。然而，即使有了良好的低级策略，在特定场景或环境中，直接应用大型语言模型执行也是低效的，甚至是不可行的。因此，这些研究提出了利用负担能力指导（Ahn 等人，2022 年）、自回归校正（Huang 等人，2022 年a）、校正再提示（Raman 等人，2022 年）、交错推理和行动轨迹生成（Yao 等人，2022 年）以及描述反馈（Huang 等人，2022 年b；Wang 等人，2023 年c；Du 等人，2023 年；Wang 等人，2023 年a）来生成有效的行动序列。即使没有 RL 模块，Wu 等人（2023 年）也发现，GPT-4（OpenAI，2023 年）在相关 RL 论文和思维链的提示下，也能在 RL 基准任务中取得有竞争力的表现。

5 总结与未来展望

本文简要回顾了 RL 变换器方面的进展。我们对这些进展进行了分类：*a)* 变换器可以作为 RL 的一个强大模块，例如，充当表示模块或世界模型；*b)* 变换器可以充当顺序决策制定器；*c)* 变换器有利于跨任务和跨领域的泛化。虽然我们讨论的是这一主题的代表性作品，但变换器在 RL 中的应用并不局限于我们的讨论。鉴于变换器在更广泛的人工智能领域的蓬勃发展，我们认为将变换器与 RL 结合起来是一个大有可为的趋势。最后，我们将讨论这一方向的未来前景和有待解决的问题。

强化学习与（自我）监督学习相结合。追溯 Trans- formRL 的发展历程，训练方法涉及强化学习和（自我）监督学习。在传统的 RL 框架下进行训练时，Transformer 架构通常无法进行优化（Parisotto 等人，2020 年）。当使用变形器通过序列建模解决决策问题时，由于采用了（自我）监督学习范式，"致命的三元组问题"（Van Hasselt 等人，2018 年）得以消除。在这种框架下，策略的性能深受离线数据质量的制约。因此，当我们在 Transformer 学习中结合 RL 和（自我）监督学习时，可能会学到更好的策略。一些研究（Zheng 等人，2022 年；Meng 等人，2021 年）尝试了监督预训练和 RL 参与微调的方案。然而，相对固定的策略会限制探索（Nair 等人，2020 年），这是需要解决的瓶颈之一。此外，沿着这一思路，用于性能评估的任务也相对简单。值得进一步探讨的是，Transformers 能否将这种（自我）监督学习扩展到更大的数据集、更复杂的环境和真实世界的应用中。此外，我们期待未来的工作能提供更多理论和经验上的见解，以说明在哪些条件下这种（自我）监督学习有望取得良好效果（Brandfonbrener 等人，2022 年；Siebenborn 等人，2022 年；Takagi，2022 年）。

通过 Transformers 架起在线和离线学习的桥梁。进入离线 RL 是 TransformRL 的一个里程碑。实际上，利用变换器捕捉决策序列中的依赖关系并抽象出政策，主要离不开大量离线数据的支持。然而，在实际应用中，某些决策任务要摆脱在线框架是不可行的。一方面，在某些任务中获取专家数据并不那么容易。另一方面，有些环境是开放式的（如 Minecraft），这就意味着策略必须不断调整，以应对在线交互过程中的未知任务。因此，我们认为有必要在在线学习和离线学习之间架起一座桥梁。然而，继决策转换器（DT）之后的大多数研究进展都集中在离线学习框架上。有几项研究试图采用离线预训练和在线微调的范式（Xie 等人，2022 年）。然而，与离线 RL 算法一样，在线微调的分布偏移仍然存在，因此我们期待对 DT 进行一些特殊设计来解决这一问题。此外，如何从零开始训练一个性能良好的在线 DT 也是一个有趣的开放性问题。

为决策问题量身定制的变换器结构。目前基于 DT 的方法中的变换器结构主要是 vanilla 变换器，这种变换器最初是为文本序列设计的，可能不适合决策问题的性质。例如，对轨迹序列采用 vanilla 自注意机制是否合适？在位置嵌入时，是否需要区分决策序列中的不同元素或同一元素的不同部分？此外，由于在基于 DT 的不同算法中，将轨迹表示为序列的变体很多，如何从中选择仍缺乏系统研究。例如，在行业中部署此类算法时，如何选择稳健的后视信息？此外，Vanilla Transformer 是一种计算成本很高的结构，这使得它在训练和推理阶段都很昂贵，而且内存占用率也很高，这限制了它捕捉依赖关系的长度。为了缓解这些问题，NLP 领域的一些工作（Zhou 等人，2021 年）从这些方面对结构进行了改进，在决策问题中能否使用类似的结构也值得探讨。

通过变形金刚实现更多通用代理。我们的综述显示了变换器作为通用策略的潜力（第 4.4 节）。事实上，Transformers 的设计允许使用类似的处理模块处理多种模式（如图像、视频、文本和语音），并对大容量网络和海量数据集具有出色的可扩展性。最近的研究在训练能够执行多任务和跨领域任务的代理方面取得了重大进展。然而，鉴于这些代理是在海量数据的基础上训练出来的，它们是否只是记住了数据集，是否能进行有效的泛化，这些都还是未知数。因此，如何在没有强假设的情况下学习一种能泛化到未见任务的代理是非常值得研究的（Boustati 等人，2021 年）。此外，我们也很好奇，Transformer 是否足够强大，能针对不同任务和场景学习通用世界模型（Schubert 等人，2023 年）。

用于变形金刚的 RL。虽然我们已经讨论了如何利用 RL 从变换器的使用中获益，但反过来说，即利用 RL 从变换器的训练中获益，却是一个令人感兴趣的开放性问题，但探索得还比较少。我们看到，一些研究通过离线 RL 设置来模拟语言/对话生成任务，并通过重新标记（Snell 等人，2022b）或价值函数（Verma 等人，2022；Snell 等人，2022a；Jang 等人，2022）来学习生成策略。最近，来自人类反馈的强化学习（RLHF）（欧阳等人，2022 年）学习了一个奖励模型，并使用 RL 算法对 Transformer 进行微调，使语言模型与人类意图保持一致（Nakano 等人，2021 年；OpenAI，2023 年）。未来，我们相信 RL 可以成为进一步完善 Transformer 在其他领域性能的有用工具。

参考资料

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al：*ArXiv preprint arXiv:2204.01691*, 2022.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Men- sch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *神经信息处理系统进展》*，35:23716-23736，2022 年。

Marcin Andrychowicz、 Filip Wolski、 Alex Ray、 Jonas Schneider、 Rachel Fong、 Peter Welinder、 Bob McGrew、Josh Tobin、OpenAI Pieter Abbeel 和 Wojciech Zaremba。后见之明的经验重放。*神经信息处理系统进展》*，2017 年第 30 期。

Marcin Andrychowicz, Anton Raichuk, Piotr Stan´czyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Leonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, et al. What matters for on-policy deep actor-critic methods? A large-scale study.*国际学习表征会议*，2020 年。

Dzmitry Bahdanau、Kyunghyun Cho 和 Yoshua Bengio.通过联合学习对齐和翻译的神经机器翻译》，*arXiv preprint arXiv:1409.0473*, 2014.

Bowen Baker、Ingmar Kanitscheider、Todor Markov、Yi Wu、Glenn Powell、Bob McGrew 和 Igor Mordatch 。多机器人自动程序中的新兴工具使用。2019年*学习表征国际会议。*

Bowen Baker、Ilge Akkaya、Peter Zhokov、Joost Huizinga、Jie Tang、Adrien Ecoffet、Brandon Houghton、Raul Sampe- dro 和 Jeff Clune。视频预训练（VPT）：通过观看无标签在线视频学习行动。In Alice H.

Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Andrea Banino、Adria Puigdomenech Badia、Jacob C Walker、Tim Scholtes、Jovana Mitrovic 和 Charles Blundell。科贝尔用于强化学习的对比波特。*学习表征国际会议*，2021 年。

Marc G Bellemare、Yavar Naddaf、Joel Veness 和 Michael Bowling。街机学习环境：通用代理的评估平台。*人工智能研究期刊*》，47：253-279，2013。

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław De˛biak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *ArXiv preprint arXiv:1912.06680*, 2019.

Zhenshan Bing, Alexander Koch, Xiangtong Yao, Fabrice O Morin, Kai Huang, and Alois Knoll.通过语言指令的元强化学习。*arXiv预印本arXiv:2209.04924*，2022。

Rishi Bommasani、Drew A Hudson、Ehsan Adeli、Russ Altman、Simran Arora、Sydney von Arx、Michael S Bernstein、Jeannette Bohg、Antoine Bosselut、Emma Brunskill 等：《基础模型的机遇与风险》，*ArXiv 预印本 arXiv:2108.07258*, 2021 年。

Ayman Boustati, Hana Chockler, and Daniel C McNamee.决策转换器中的因果反事实推理迁移学习。*arXiv 预印本 arXiv:2110.14355*, 2021.

David Brandfonbrener, Alberto Bietti, Jacob Buckman, Romain Laroche, and Joan Bruna.离线强化学习的回归条件监督学习何时起作用？ *arXiv preprint arXiv:2206.01079*, 2022.

Anthony Brohan、Noah Brown、Justice Carbajal、Yevgen Chebotar、Joseph Dabis、Chelsea Finn、Keerthana Gopalakr-ishnan、Karol Hausman、*Alex* Herzog、Jasmine Hsu 等：Rt-1：用于真实世界大规模控制的机器人变压器。

Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。 语言模型是少数学习者。*神经信息处理系统进展*》，33:1877-1901，2020 年。

Micah Carroll, Orr Paradise, Jessy Lin, Raluca Georgescu, Mingfei Sun, David Bignell, Stephanie Milani, Katja Hofmann, Matthew Hausknecht, Anca Dragan, et al. Unimask：*ArXiv preprint arXiv:2211.10869*, 2022.

Chang Chen、Yi-Fu Wu、Jaesik Yoon 和 Sungjin Ahn。Transdreamer：*ArXiv preprint arXiv:2202.09481*, 2022.

陈莉莉、陆凯文、阿拉温德-拉杰斯瓦兰、李基民、阿迪提亚-格罗弗、米沙-拉斯金、彼得-阿贝尔、阿拉温德-斯里尼瓦斯和伊戈尔-莫尔达奇。决策转换器：通过序列建模进行强化学习。*神经信息处理系统进展*》，34:15084-15097，2021 年。

Sudeep Dasari 和 Abhinav Gupta.用于单次视觉模仿的变换器。*机器人学习会议*，第 2071-2084 页。PMLR，2021年。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova.Bert：用于语言理解的深度双向转换器的预训练。*arXiv preprint arXiv:1810.04805*, 2018.

董林浩、徐爽、徐波。Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition.In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp.IEEE, 2018.

Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。图像胜过 16x16 个单词：*ArXiv preprint arXiv:2010.11929*, 2020.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e：*ArXiv preprint arXiv:2303.03378*, 2023.

Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas.用大型语言模型指导强化学习中的预训练。*arXiv 预印本 arXiv:2302.06692*, 2023.

斯科特-埃蒙斯、本杰明-艾森巴赫、伊利亚-科斯特里科夫和谢尔盖-莱文。Rvs：通过监督学习实现离线 RL 的关键是什么？ *arXiv preprint arXiv:2112.10751*, 2021.

Logan Engstrom、 Andrew Ilyas、 Shibani Santurkar、 Dimitris Tsipras、 Firdaus Janoos、 Larry Rudolph 和 Aleksander Madry。深度 RL 中的实施问题：PPO和TRPO案例研究。2019年*学习表征国际会议。*

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar.Menedojo：构建具有互联网规模知识的开放式化身代理。*第三十六届神经信息处理系统数据集与基准会议*，2022年。

Scott Fujimoto、David Meger 和 Doina Precup。无需探索的非策略深度强化学习。在 *机器学习国际会议*，第 2052-2062 页。PMLR，2019 年。

Hiroki Furuta, Yutaka Matsuo, and Shixiang Shane Gu.用于离线后见信息匹配的广义决策变换器。*arXiv 预印本 arXiv:2111.10364*, 2021.

Dibya Ghosh, Abhishek Gupta, Ashwin Reddy, Justin Fu, Coline Devin, Benjamin Eysenbach, and Sergey Levine. 通过迭代监督学习达到目标。*arXiv preprint arXiv:1912.06088*, 2019.

皮埃尔-路易-古尔、陈世哲、里卡多-加西亚-皮内尔、马卡兰德-塔帕斯维、伊万-拉普捷夫和科迪莉亚-施密德。机器人操作的指令驱动历史感知策略。*机器人学习会议*，第 175- 页 187.PMLR, 2023.

达尼亚尔-哈夫纳、蒂莫西-利利克拉普、吉米-巴和穆罕默德-诺鲁兹。从梦想到控制：*arXiv preprint arXiv:1912.01603*, 2019.

达尼亚尔-哈夫纳、蒂莫西-利利克拉普、吉米-巴和穆罕默德-诺鲁兹。从梦想到控制：通过潜在想象力学习行为。2020年*学习表征国际会议。*

达尼亚尔-哈夫纳、蒂莫西-P-利利克拉普、穆罕默德-诺鲁兹和吉米-巴。用离散世界模型掌握 Atari。*国际学习表征会议*，2021年。

达尼亚尔-哈夫纳、尤吉斯-帕苏科尼斯、吉米-巴、蒂莫西-利利克拉普。通过世界模型掌握不同领域 *arXiv preprint arXiv:2301.04104*, 2023.

Matthew Hausknecht 和 Peter Stone.部分可观测 mdps 的深度递归 q-learning.*2015年AAAI秋季系列研讨会*，2015年。

何开明、张翔宇、任少清和孙健。图像识别的深度残差学习。In *Proceed- ings of the IEEE conference on computer vision and pattern recognition*, pp.

Peter Henderson、Riashat Islam、Philip Bachman、Joelle Pineau、Doina Precup 和 David Meger。重要的深度强化学习。*美国人工智能学会会议论文集*》，第 32 卷，2018 年。

Sepp Hochreiter 和 Jürgen Schmidhuber.长短期记忆。*神经计算*》，9（8）：1735-1780，1997。Edward J Hu、沈业龙、菲利普-沃利斯、朱泽元、李远志、王申、王璐、陈伟柱。LoRA：大型语言模型的低等级适应。*国际学习表征会议*、2022.

胡凯哲、郑雷辰、高扬、徐华哲。随机丢帧下的决策变换器》，*arXiv 预印本 arXiv:2303.03391*, 2023.

Siyi Hu、Fengda Zhu、Xiaojun Chang 和 Xiaodan Liang.Updet：通过变压器实现策略解耦的通用多代理 RL。2020年*国际学习表征会议。*

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 密集连接的卷积网络。In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.

黄文龙、Pieter Abbeel、Deepak Pathak 和 Igor Mordatch。语言模型作为零射规划器：*ArXiv preprint arXiv:2201.07207*, 2022a.

黄文龙、夏飞、肖特、Harris Chan、梁杰、Pete Florence、曾安迪、Jonathan Tompson、Igor Mordatch、Yevgen Chebotar 等。 内心独白：*ArXiv preprint arXiv:2207.05608*, 2022b.

Kazuki Irie, Imanol Schlag, Róbert Csordás, and Jürgen Schmidhuber.用快速权重编程器超越线性变换器。*神经信息处理系统进展*，34：7703-7717，2021。

Youngsoo Jang、Jongmin Lee 和 Kee-Eung Kim。Gpt-critic：端到端任务导向对话系统的离线强化学习。*国际学习表征会议*，2022 年。

Michael Janner、Justin Fu、Marvin Zhang 和 Sergey Levine。何时相信你的模型？基于模型的策略优化。*神经信息处理系统进展》*，2019 年第 32 期。

Michael Janner、Qiyang Li 和 Sergey Levine。作为一个大序列建模问题的强化学习。在 *ICML 2021 年无监督强化学习研讨会*，2021 年。

蒋云帆、阿格里姆-古普塔、张子辰、王冠之、窦永强、陈彦君、李菲菲、阿尼玛-阿南德-库马尔、朱玉珂和范林熙。维玛：*ArXiv preprint arXiv:2210.03094*, 2022.

Amir Ardalan Kalantari, Mohammad Amini, Sarath Chandar, and Doina Precup.使用注意力和视觉转换器提高基于值的模型的样本效率》。*arXiv 预印本 arXiv:2202.00710*, 2022.

Jikun Kang、Romain Laroche、Xindi Yuan、Adam Trischler、Xue Liu 和 Jie Fu。先思而后行：具有内部工作记忆的决策转换器。*arXiv预印本arXiv:2305.16338*，2023。

萨尔曼-汗、穆扎玛尔-纳赛尔、穆纳瓦尔-哈亚特、赛义德-瓦卡斯-扎米尔、法哈德-沙巴兹-汗和穆巴拉克-沙阿。视觉中的变形金刚：调查。*ACM计算调查（CSUR）*，54（10s）：1-41，2022.

Sachin G Konan、Esmaeil Seraj 和 Matthew Gombolay。对比决策变换器*第 6 届机器人学习年会*，2022 年。

Aviral Kumar、Aurick Zhou、George Tucker 和 Sergey Levine.用于离线强化学习的保守q-learning。*神经信息处理系统进展》*，33:1179-1191，2020 年。

Vitaly Kurin, Maximilian Igl, Tim Rocktäschel, Wendelin Boehmer, and Shimon Whiteson.我的身体是个笼子：形态学在基于图的不兼容控制中的作用》。*arXiv 预印本 arXiv:2010.01856*, 2020.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, et al. In-context reinforcement learning with algorithm distillation. *ArXiv preprint arXiv:2210.14215*, 2022.

Yann LeCun、Léon Bottou、Yoshua Bengio 和 Patrick Haffner。基于梯度的学习应用于文档识别。*电气和电子工程师学会论文集》*，86（11）：2278-2324，1998 年。

Kuang-Huei Lee, Ofir Nachum, Sherry Yang, Lisa Lee, C. Daniel Freeman, Sergio Guadarrama, Ian Fischer, Winnie Xu, Eric Jang, Henryk Michalewski, and Igor Mordatch.多游戏决策转换器*神经信息处理系统进展》*，2022 年。

Sergey Levine、Aviral Kumar、George Tucker 和 Justin Fu。离线强化学习：*ArXiv preprint arXiv:2005.01643*, 2020.

Shuang Li、Xavier Puig、Chris Paxton、Yilun Du、Clinton Wang、Linxi Fan、Tao Chen、De-An Huang、Ekin Akyürek、Anima Anandkumar 等。 用于交互决策的预训练语言模型。*神经信息处理系统进展》*，35:31199-31212，2022 年。

Qinjie Lin, Han Liu, and Biswa Sengupta.用于多任务强化学习的具有分布值近似的开关轨迹变换器。*ArXiv 预印本 arXiv:2203.07413*, 2022.

林子川、赵立、杨德瑞、秦涛、刘铁岩、杨广文。强化学习的分布式奖励分解。*神经信息处理系统进展》*，32，2019.

Minghuan Liu, Menghui Zhu, and Weinan Zhang.目标条件强化学习：问题与解决方案 *arXiv preprint arXiv:2201.08299*, 2022.

Ricky Loynd, Roland Fernandez, Asli Celikyilmaz, Adith Swaminathan, and Matthew Hausknecht.工作记忆图。*国际机器学习会议*，第 6404-6414 页。PMLR, 2020.

Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch.作为通用计算引擎的预训练变换器。 *arXiv preprint arXiv:2103.05247*, 1, 2021.

毛瀚宇、赵睿、陈浩、郝建业、陈轶群、李东、张俊阁、肖震。变压器中的变压器作为深度强化学习的骨干。*arXiv preprint arXiv:2212.14538*, 2022.

Luckeciano C Melo.变压器是元强化学习器。 In *International Conference on Machine Learn- ing*, pp.PMLR, 2022.

孟令辉、温慕宁、杨耀东、乐晨阳、李喜云、张伟南、温颖、张海峰、王军、徐波。离线预训练多代理决策转换器：*ArXiv preprint arXiv:2112.02845*, 2021.

Vincent Micheli, Eloi Alonso, and François Fleuret. 变压器是样本有效的世界模型。*arXiv 预印本 arXiv:2209.00588*，2022。

Nikhil Mishra、Mostafa Rohaninejad、Xi Chen 和 Pieter Abbeel。一个简单的神经注意元学习器。*国际学习表征大会*，2018。

Volodymyr Mnih、Koray Kavukcuoglu、David Silver、Andrei A Rusu、Joel Veness、Marc G Bellemare、Alex Graves、Martin Riedmiller、Andreas K Fidjeland、Georg Ostrovski 等：《通过深度强化学习实现人级控制》，《*自然*》，518（7540）：529-533，2015。

Ashvin Nair、Abhishek Gupta、Murtaza Dalal 和 Sergey Levine。Awac：*ArXiv preprint arXiv:2006.09359*, 2020.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt：有人工反馈的浏览器辅助问题解答。*arXiv 预印本 arXiv:2112.09332*, 2021.

OpenAI.Gpt-4 技术报告，2023 年。

Kei Ota、Tomoaki Oiki、Devesh Jha、Toshisada Mariyama 和 Daniel Nikovski.增加输入维度能否改善深度强化学习？*国际机器学习大会》*，第 7424-7433 页。PMLR, 2020.

Kei Ota, Devesh K Jha, and Asako Kanezaki.为深度强化学习训练更大的网络。*arXiv preprint arXiv:2102.07920*, 2021.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *ArXiv preprint arXiv:2203.02155*, 2022.

Sherjil Ozair、Yazhe Li、Ali Razavi、Ioannis Antonoglou、Aaron Van Den Oord 和 Oriol Vinyals。用于规划的向量量化模型。*国际机器学习大会*，第 8302-8313 页。PMLR，2021 年。

Emilio Parisotto 和 Ruslan Salakhutdinov.使用行为学习器分解强化学习中的高效转换器》。*ArXiv 预印本 arXiv:2104.01655*, 2021.

Emilio Parisotto, Francis Song, Jack Rae, Razvan Pascanu, Caglar Gulcehre, Siddhant Jayakumar, Max Jaderberg, Raphael Lopez Kaufman, Aidan Clark, Seb Noury, et al. Stabilizing transformers for reinforcement learning.*国际机器学习会议*，第 7487-7498 页。PMLR, 2020.

Keiran Paster、Sheila McIlraith 和 Jimmy Ba.你不能指望运气：为什么决策转换器在随机环境中会失败》，*arXiv preprint arXiv:2205.15967*, 2022.

Shreyas Sundara Raman、Vanya Cohen、Eric Rosen、Ifrah Idrees、David Paulius 和 Stefanie Tellex。通过纠正性再提示使用大型语言模型进行规划。*arXiv 预印本 arXiv:2211.09935*, 2022.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. *A* generalist agent.

Machel Reid、Yutaro Yamada 和 Shixiang Shane Gu.维基百科能帮助离线强化学习吗？ *arXiv preprint arXiv:2201.12122*, 2022.

Jan Robine, Marc Höftmann, Tobias Uelwer, and Stefan Harmeling.基于变换器的世界模型在 100k 次交互中都很快乐。*arXiv 预印本 arXiv:2303.07109*, 2023.

Tom Schaul、Daniel Horgan、Karol Gregor 和 David Silver。通用值函数近似器。*机器学习国际会议*，第 1312-1320 页。PMLR，2015。

Julian Schrittwieser、Ioannis Antonoglou、Thomas Hubert、Karen Simonyan、Laurent Sifre、Simon Schmitt、Arthur Guez、Edward Lockhart、Demis Hassabis、Thore Graepel 等。 通过学习模型规划掌握阿塔里、围棋、国际象棋和将棋。*自然》*，588（7839）：604-609，2020。

Ingmar Schubert, Jingwei Zhang, Jake Bruce, Sarah Bechtle, Emilio Parisotto, Martin Riedmiller, Jost Tobias Springenberg, Arunkumar Byravan, Leonard Hasenclever, and Nicolas Heess.用于控制的通用动力学模型。*arXiv 预印本 arXiv:2305.10912*, 2023.

Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter Abbeel.用于视觉控制的遮蔽世界模型。*机器人学习会议*，第 1332-1344 页。PMLR，2022a。

Younggyo Seo、Kimin Lee、Stephen L James 和 Pieter Abbeel。通过视频进行无动作预训练的强化学习。*国际机器学习会议*，第 19561-19579 页。PMLR，2022b。

努尔-穆罕默德-马希-沙菲乌拉、崔子辰、阿里云图亚-阿尔坦扎亚、勒雷尔-平托。行为转换器：*arXiv preprint arXiv:2206.11251*, 2022.

尚京焕、库马拉-卡哈塔皮蒂亚、李翔、迈克尔-S-柳。Starformer：用于视觉强化学习的具有状态-动作-奖励表示的变换器。In *European Conference on Computer Vision*, pp.

479.Springer, 2022.

Max Siebenborn、Boris Belousov、Junning Huang 和 Jan Peters。变压器在决策变压器中的作用有多大？*arXiv preprint arXiv:2211.14655*, 2022.

David Silver、Aja Huang、Chris J Maddison、Arthur Guez、Laurent Sifre、George Van Den Driessche、Julian Schrit-twieser、Ioannis Antonoglou、Veda Panneershelvam、Marc Lanctot 等：《用深度神经网络和树搜索掌握围棋游戏》，《*自然*》，529(7587):484-489, 2016。

Samarth Sinha、Homanga Bharadhwaj、Aravind Srinivas 和 Animesh Garg。D2rl：强化学习中的深度密集架构。*ArXiv 预印本 arXiv:2010.09163*, 2020.

Charlie Snell、Ilya Kostrikov、Yi Su、Mengjiao Yang 和 Sergey Levine.隐式语言 Q 学习的离线自然语言生成 RL。*arXiv 预印本 arXiv:2206.11871*, 2022a.

Charlie Snell, Sherry Yang, Justin Fu, Yi Su, and Sergey Levine.面向目标对话系统的语境感知语言建模。*arXiv 预印本 arXiv:2204.10198*, 2022b.

孙建凯、黄德安、陆波、刘云辉、周伯磊和 Animesh Garg。盘子：程序任务中使用变压器的视觉基础规划。*IEEE 机器人与自动化通讯*》，7（2）：4924-4930，2022 年。

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi- agent learning. *ArXiv preprint arXiv:1706.05296*, 2017.

高木史郎不同模式的变压器预训练对离线强化学习的影响。
arXiv preprint arXiv:2211.09817, 2022.

唐昱瑾和夏大卫.作为变压器的感觉神经元：用于强化学习的置换不变神经网络。*神经信息处理系统进展*》，34:22574-22587，2021 年。

陶天心、丹尼尔-雷达、米希尔-范德潘内。从像素评估深度强化学习的视觉变换器方法》，*arXiv preprint arXiv:2204.04905*, 2022.

Yi Tay、Mostafa Dehghani、Dara Bahri 和 Donald Metzler。高效变压器：一项调查。*ACM Computing Surveys*，55（6）：1-28，2022.

DeepMind 交互式代理团队、Josh Abramson、Arun Ahuja、Arthur Brussee、Federico Carnevale、Mary Cassin、Felix Fischer、Petko Georgiev、Alex Goldin、Mansi Gupta 等：《利用模仿和自我监督学习创建多模态交互式代理》，*ArXiv 预印本 arXiv:2112.03763*, 2021 年。

Hado Van Hasselt、Yotam Doron、Florian Strub、Matteo Hessel、Nicolas Sonnerat 和 Joseph Modayil.Deep rein- forcement learning and the deadly triad. *arXiv preprint arXiv:1812.02648*, 2018.

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力就是你所需要的一切。*神经信息处理系统进展*》，2017年30期。

David Venuto, Sherry Yang, Pieter Abbeel, Doina Precup, Igor Mordatch, and Ofir Nachum.多环境预训练可转移到行动受限数据集》，*arXiv preprint arXiv:2211.13337*, 2022.

Siddharth Verma、Justin Fu、Mengjiao Yang 和 Sergey Levine。Chai：A chatbot ai for task-oriented dialogue with offline reinforcement learning. *arXiv preprint arXiv:2204.08426*, 2022.

Adam R Villaflor、黄哲、Swapnil Pande、John M Dolan 和 Jeff Schneider。解决强化学习序列建模中的乐观主义偏差。*国际机器学习大会*》，第 22270-22283 页。PMLR, 2022.

Oriol Vinyals、Igor Babuschkin、Wojciech M Czarnecki、Michaël Mathieu、Andrew Dudzik、Junyoung Chung、David H Choi、Richard Powell、Timo Ewalds、Petko Georgiev 等：利用多代理强化学习提高《星际争霸 II》的大师级水平。*自然*》，575（7782）：350-354，2019。

王冠之、谢雨琦、蒋云帆、阿杰-曼德勒卡尔、肖超伟、朱昱珂、范林熙和阿尼玛-阿南德-库马尔。旅行者具有大型语言模型的开放式嵌入式代理。*arXiv预印本arXiv:2305.16291*，2023a。

王克荣、赵汉业、罗旭芳、任侃、张伟南、李东生。用于离线强化学习的引导变换器。*ArXiv 预印本 arXiv:2206.08569*, 2022.

王润东、王伟轩、曾宪汉、王亮、连振杰、高一鸣、刘飞宇、李思勤、王贤良、傅强等 多代理多游戏实体转换器。2023b.

王梓豪、蔡绍飞、刘安吉、马晓健和梁一涛。描述、解释、规划和选择：使用大型语言模型的交互式规划使开放世界多任务代理成为可能。*arXiv预印本arXiv:2302.01560*, 2023c.

王梓瑜、汤姆-肖尔、马特奥-赫塞尔、哈多-哈瑟尔特、马克-兰克托和南多-弗雷塔斯。深度强化学习的对决网络架构。*国际机器学习会议*，第 1995-2003 页。PMLR，2016。

温慕宁、Jakub Grudzien Kuba、林润基、张伟南、温颖、王军、杨耀东。多代理强化学习是一个序列建模问题。*ArXiv 预印本 arXiv:2205.14953*, 2022.

吴越、苏妍敏、Shrimai Prabhumoye、Yonatan Bisk、Ruslan Salakhutdinov、Amos Azaria、Tom Mitchell 和李远志。春天Gpt-4通过研究论文和推理超越RL算法. *arXiv预印本arXiv:2305.15486*, 2023.

谢志辉、林子川、李俊友、李帅、叶德恒。深度强化学习中的预训练：调查。*arXiv preprint arXiv:2211.03959*, 2022.

Mengdi Xu、Yikang Shen、Shun Zhang、Yuchen Lu、Ding Zhao、Joshua Tenenbaum 和 Chuang Gan。用于少量策略泛化的提示决策转换器。*国际机器学习大会》*，第 24631-24645 页。PMLR, 2022.

Taku Yamagata、Ahmed Khalil 和 Raul Santos-Rodriguez.Q-learning decision transformer：在离线 RL 中利用动态编程进行条件序列建模。*arXiv 预印本 arXiv:2209.03993*, 2022.

杨梦娇、Dale Schuurmans、Pieter Abbeel 和 Ofir Nachum。控制的二分法：*ArXiv preprint arXiv:2210.13435*, 2022a.

Mengjiao Sherry Yang、Dale Schuurmans、Pieter Abbeel 和 Ofir Nachum。程序克隆的思维模仿链。*神经信息处理系统进展》*，35：36366-36381，2022b。

姚顺禹、赵杰夫、于琰、杜楠、Izhak Shafran、Karthik Narasimhan 和 Yuan Cao。React：*arXiv preprint arXiv:2210.03629*, 2022.

叶德恒、陈桂斌、张文、陈胜、袁波、刘波、陈佳、刘钊、邱富豪、于洪生等：《用深度强化学习玩全moba游戏》。*神经信息处理系统进展》*，33:621-632，2020a.

叶德恒、刘钊、孙明飞、石蓓、赵培林、吴昊、于洪生、杨少杰、吴锡鹏、郭庆伟等：用深度强化学习掌握moba 游戏中的复杂控制。In *Proceedings of the AAAI Conference on Artificial Intelligence, volume* 34, pp.

Chao Yu, Akash Velu, Eugene Vinitsky, Yu Wang, Alexandre Bayen, and Yi Wu.合作多代理博弈中 PPO 的惊人有效性》。*ArXiv 预印本 arXiv:2103.01955*, 2021a.

余天和、阿维拉尔-库马尔、拉斐尔-拉法洛夫、阿拉温德-拉杰斯瓦兰、谢尔盖-莱文和切尔西-芬恩。组合：基于模型的保守离线策略优化。*神经信息处理系统进展》*，34：28954-28967，2021b。

Vinicius Zambaldi、David Raposo、Adam Santoro、Victor Bapst、Yujia Li、Igor Babuschkin、Karl Tuyls、David Re-ichert、Timothy Lillicrap、Edward Lockhart 等：带有关系归纳偏差的深度强化学习。*国际学习表征会议*，2018。

Qinqing Zheng, Amy Zhang, and Aditya Grover.在线决策变换器。*ArXiv 预印本 arXiv:2202.05607*, 2022.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang.告知器：超越长序列时间序列预测的高效转换器。In *Proceedings of the AAAI Conference on Artificial Intelligence, volume* 35, pp.

Baiting Zhu, Meihua Dang, and Aditya Grover.通过离线多目标 Rl 实现规模化帕累托高效决策 *arXiv preprint arXiv:2305.00567*, 2023.