TQL：通过防止注意力坍塌使用Transformer扩展Q函数

Perry Dong [*] **Kuo-Han Hung** [*] **Alexander Swerdlow Dorsa Sadigh Chelsea Finn**斯坦福大学https://pd-perry.github.io/transformer-q-learning/

摘要

尽管规模扩展推动了机器学习领域近期的诸多重大进展，强化学习（RL）方法仍主要使用小型价值函数。对价值函数进行简单扩展——包括采用已知具有高度可扩展性的Transformer 架构——往往会导致学习不稳定，且性能更差。在这项工作中，我们探究是什么阻碍了 Transformer 有效扩展价值函数。通过实证分析，我们发现这种扩展中的关键失效模式：随着容量增加，注意力分数会坍塌。我们的核心洞见是，通过控制注意力分数的熵，可以有效防止这种坍塌并稳定训练，从而能够使用更大的模型。为此，我们提出 Transformer Q 学习（TQL），这是一种能够释放 Transformer 在强化学习中学习价值函数时扩展潜力的方法。从最小网络规模扩展到最大网络规模时，我们的方法最多可将性能提升 43%，而先前的方法则会出现性能下降。

1. 引言

深度学习的最新进展表明，扩展模型容量能够显著提升自然语言处理（Ope-nAIet al.，2024年）、计算机视觉（Siméoniet al.，2025年）和机器人学（Physical Intelligenceet al.，2025b；Teamet al.，2025年）等领域的性能。然而，与通常通过增加参数便能提升性能的监督学习不同，强化学习（RL）历来面临着显著的扩展挑战。更大的网络往往无法更好地捕捉数据分布，反而会导致训练不稳定，从而降低性能。近期研究探索了在强化学习中使用更大的策略网络。本文聚焦于扩展价值函数。由于策略最终是从价值函数推导而来的，增加价值函数的参数化规模理应显著提升强化学习的性能；然而，使用更大的价值函数网络往往会导致更差的性能，如何有效扩展这类网络仍是一个悬而未决的问题。在传统监督学习中，Transformer 已被证明能够在众多领域实现良好的扩展。因此，我们要探究的是：是什么阻碍了 Transformer 在强化学习价值学习中实现有效扩展?

在强化学习中扩展网络的先前流行方法采用周期性网络重置，以在在线训练期间于更高的更新数据比（UTD）下保持可塑性（Nauman 等人，2024年；Schwarzer 等人，2023年）。虽然这些方法保留了学习容量，但由于网络需要从头重新训练，它们的计算开销很大，并且通过参数重新初始化存在灾难性遗忘的风险，这可能使在线学习变得不安全。替代方法探索了归一化技术（Lee 等人，2025年）或替代架构（Obando-Ceron 等人，2024年），这些方法通常需要特定的架构改动，因而难以应用于所有场景，例如预训练网络的情况。在本工作中，我们着重理解为何 Transformer 无法在强化学习的价值函数学习中实现扩展，并提出一个通用、最小化的框架来解决这些问题并实现扩展。

为此，我们提出 Transformer Q 学习（TQL），这是一种用于强化学习中可扩展价值函数训练的简单框架。TQL 利用 Transformer 架构在价值函数训练中的可扩展性，同时尽可能减少架构改动。将 Transformer 直接应用于价值学习会产生严重的病理现象（第5.2节）。通过实证分析，我们识别出扩展 Transformer 价值函数训练时的关键失效模式：随着容量增加，Transformer 的注意力权重会发生坍缩，主要关注少数词元而忽略其余词元，从而导致更大的

[*]同等贡献。通讯作者：Perry Dong <per- ryd@stanford.edu>, Kuo-Han Hung <khhung@stanford.edu>。

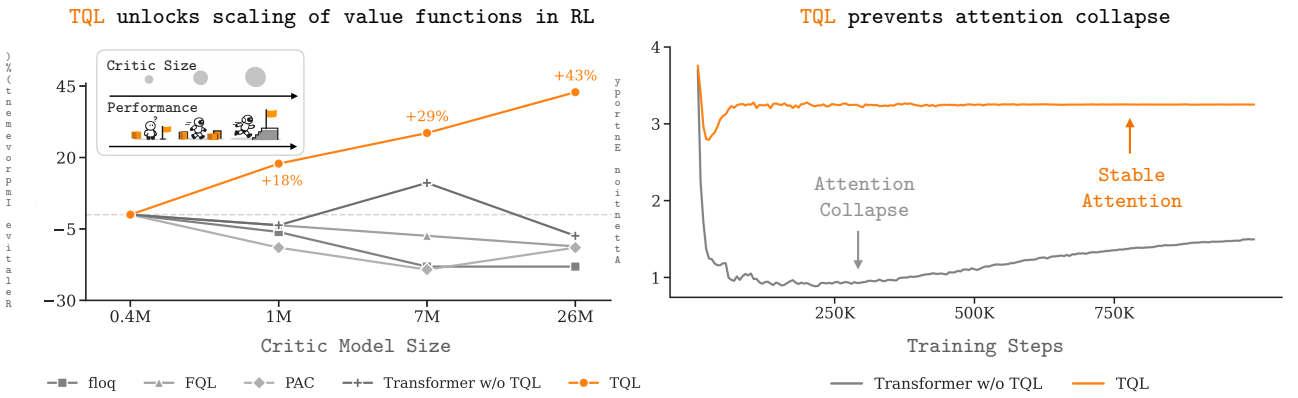

图*1.* **TQL 通过防止注意力坍塌，解锁了强化学习中价值函数的扩展。** 左：TQL 与采用不同生成模型骨干网络的先前方法相比的扩展结果。结果以相对于最小模型规模的提升幅度呈现，并在评估套件中的全部 25 项任务上取平均值，跨随机种子求平均。随着网络规模增大，TQL 的性能持续提升，而先前方法无法有效利用额外容量，性能反而下降。右：采用和未采用 TQL 时的注意力熵，在 25 项任务上取平均值。TQL 通过防止扩展 Transformer 架构进行价值函数训练时发生的注意力坍塌，实现有效扩展。

模型难以在价值自举过程中学习适当的注意力模式。作为一种简单的补救方法，我们提出对注意力分数的熵进行逐层控制。通过将注意力分数的熵调整至目标值，TQL 确保模型有效地将注意力分配到所有输入词元，从而通过防止坍缩实现稳定训练。这使得更大的模型能够利用其容量拟合有用信号，并解锁 Transformer 在价值函数训练中的扩展（图1）。

我们的主要贡献是 TQL，这是一个通过引导注意力熵来实现稳定且有效的价值函数训练、并使其随模型容量扩展的框架。我们在离线强化学习场景中分析 TQL，并在 OGBench 基准（Park 等人，2025a）中的高难度连续控制任务上进行评估，结果表明，与先前方法相比，TQL 在扩展价值函数方面有效。

2. 相关工作

离线强化学习。离线强化学习（Levine等人，2020）的目标是在离线数据集上训练策略，而无需与环境交互。离线强化学习的核心问题是处理数据集分布之外的动作，因为这些动作的价值估计并不可靠。解决这一问题主要有两种方法。一种方法是采用保守性，学习悲观价值函数，以惩罚分布外动作（Tarasov等人，2023；Kumar等，2020；Wu 等人，2019；2021；Yu等人，2020）。另一种方法是施加策略约束，使策略动作接近行为分布（Dong 等人，2025d；Park 等人，2025b；Hansen-Estruch等人，2023；Kostrikov等，2021；Fujimoto和Gu，2021；Nair等人，2021；Kidambi等人，2021；Fujimoto等人，2019）。近期研究还探索了将现代生成式架构用于价值函数学习（Dong等人，2025d；Agrawalla等人，2025年）。我们的工作同样提出了一种新颖的价值函数设计，但明确着眼于可扩展性，以及随着模型容量增加而实现有效的性能提升。我们引入的技术与这些先前工作在很大程度上相互正交，因为原则上，TQL可以与这些不同的算法和分布选择结合使用。

强化学习中的 Transformer。Transformer 已在轨迹建模（Wu 等人，2023年；Zheng 等人，2022；Furuta 等人，2022；Chen 等人，2021；Janner 等人，2021）、世界建模（Cheng 等人，2025年）以及联合策略-价值架构（Springenberg 等人，2024）中取得了显著成功。相比之下，我们关注的是如何专门针对价值函数有效扩展 Transformer。尽管已有一些工作探索了基于 Transformer 的价值函数设计（Obando-Ceron 等人，2024；Chebotar 等人，2023），但 Transformer 无法直接扩展到价值函数的原因仍不明确，我们也缺乏大规模基于 Transformer 的价值函数训练的可靠方案。Physical Intelligence 等人（2025a）探索了使用蒙特卡罗回报训练价值函数，而我们则探索如何通过自举训练此类价值函数；事实证明，这种方法效果更好，但通常会带来更多训练不稳定性。其他许多工作研究了在自然语言处理等领域扩展 Transformer 时产生的训练不稳定性（翟等人，2023；Wortsman 等人，2023年；Zhuo 等人，2025年）。与这些方法不同，我们专门关注使用 Transformer 进行价值函数建模，研究能够在价值函数训练环境中实现稳定且有效扩展的架构选择。

强化学习中的参数扩展。 与监督学习领域相比，扩大强化学习中的参数规模仍然颇具挑战。现有方法主要通过

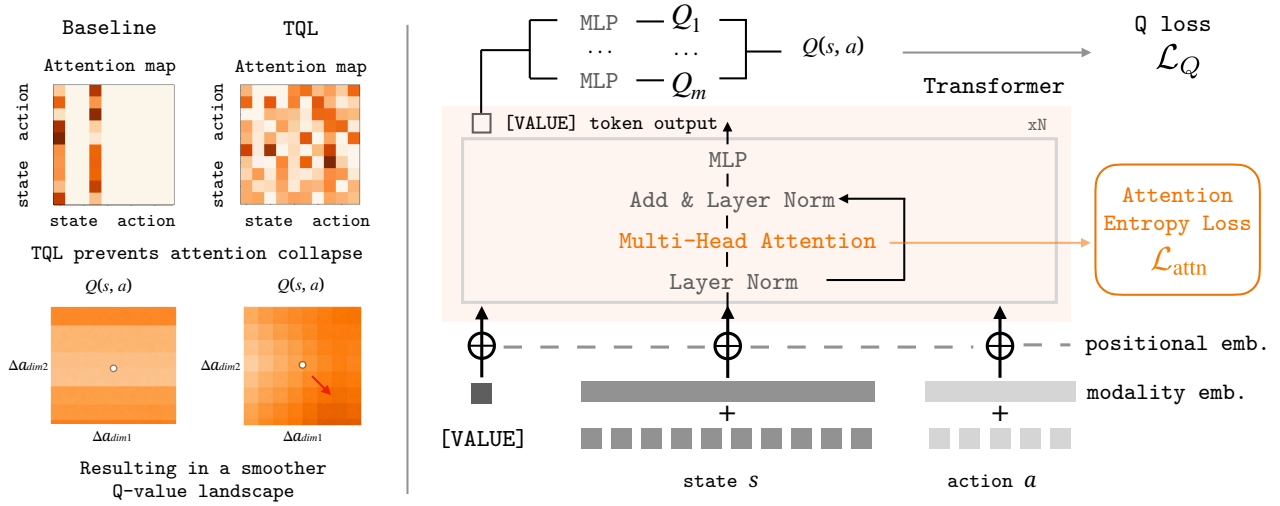

图 2. **使用 TQL 扩展价值网络。** TQL 通过控制注意力分数的目标熵（右图），防止较大网络中的注意力坍塌。与未经正则化的模型相比，TQL 的注意力分布更加均匀，价值景观也更加平滑（左图）。

正则化。一类方法采用定期重置网络的方式，在训练过程中保持可塑性（Nauman 等人，2024年；Schwarzer 等人，2023年），但这种做法计算开销很大，因为网络需要定期从头开始重新训练，而且还可能导致先前学会的动作发生灾难性遗忘，使现实世界中的在线训练变得不安全。其他方法则探索通过归一化技术修改架构和目标函数，以稳定更大规模网络的训练（Lee 等人，2025年），采用混合专家或多跳残差连接等特定架构（Obando-Ceron 等人，2024年；Castanyer 等人，2025年），或者在学习价值函数时使用分类损失，而非回归（Farebrother 等人，2024年）。近期研究考察了如何扩展行为者和评论家网络的深度（Wang 等人，2025年），并在自监督对比强化学习设置中展现了性能提升。尽管这些方法应对了特定的扩展挑战，但它们往往存在局限，需要特定设置或架构改动，因此无法适用于所有情况，例如预训练模型的情形。在这项工作中，我们着重设计一种简单通用的方法，以扩展用于强化学习价值函数训练的 Transformer。

3. 预备知识

我们考虑由元组 $\{\mathcal{S}, \mathcal{A}, \rho, r, \gamma, T\}$ 定义的马尔可夫决策过程（MDP），其中 $\mathcal{S}$ 是状态空间，$\mathcal{A}$ 是动作空间，$\rho(s)$ 是初始状态分布，$r: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 是定义奖励的函数，$\gamma \in [0, 1]$ 是折扣因子，而 $T(s'|s, a)$ 指定了转移概率。强化学习的目标是学习一种策略，使折扣回报的期望总和最大化 $\mathbb{E}_\pi[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)]$。本文考虑离线强化学习问题，该问题还假设存在一个固定数据集 $\mathcal{D} = \{(s, a, r, s')\}$，其中包含预先收集的转移数据。离线强化学习旨在通过从固定数据集中学习，在不进行环境交互的情况下最大化回报。

4. Transformer Q 学习

本节将介绍 TQL 框架，以便利用 Transformer 架构实现可扩展的价值函数训练。我们指出了一个关键挑战：随着网络试图利用额外的容量学习应关注哪些词元，直接扩大 Transformer 的规模会导致注意力权重的熵坍塌（完整分析见第5.2节）。我们的核心洞见是，防止这种熵坍塌能够让 Transformer 在价值函数训练中有效扩展；而实现这一点的方法是直接控制注意力分数的熵，使其趋近于目标值。我们首先介绍使用 Transformer 训练价值函数的设置（第 4.1节），然后介绍控制注意力分数熵的方法（第4.2节），最后讨论实现细节，具体而言，是引入可学习模态嵌入，以帮助较大的模型更好地关注与任务相关的模式，以及策略提取（§4.3节）。完整的 TQL 框架如图2所示，完整算法见算法 1。

4.1. 基于 Transformer 的价值函数

我们采用 Transformer 架构来学习 Q 函数 $Q(s, a)$，该函数将状态-动作对映射为期望回报。我们使用具有完整自注意力的标准 Transformer 解码器来预测价值，其中将状态的每个维度——

以及动作 $a \in \mathbb{R}^{n_a}$ 的每个维度都视为一个词元，并将每个标量投影到隐藏维度。词元与位置编码结合后，再由一组 Transformer层进行处理。我们在序列前添加一个可学习的[价值]词元，并将其最终表示传入多层感知机头，以生成价值预测。我们采用标准的 Q学习目标

$$\mathcal{L}_Q(\phi) = \mathbb{E}[(Q_\phi(s,a) - r - \gamma Q_{\phi'}(s',a'))^2] \quad (1)$$

其中 $Q_{\phi'}$ 是 Q_ϕ的延迟副本。该设计使模型能够通过自注意力机制学习状态词元与动作词元之间的复杂依赖关系，同时保持简洁。如第 5.2 节所示，仅仅扩大这一标准设计的参数容量，就会导致训练不稳定性，实际上还会造成更差的性能，而非提升性能。事实上，我们发现，扩展会导致性能可预测地下降——在一个案例中（图 3），平均成功率从 46% 骤降至仅6%。

4.2. 直接引导注意力分数的熵

我们假设，大型 Transformer 价值网络性能不佳可归因于注意力坍塌。具体而言，第 5.2 节中的实验表明，大型 Q网络的注意力分布会越来越集中于少数几个词元。一旦发生坍塌，模型便无法准确估计预期折扣回报的价值，也无法取得较高的性能。

为应对注意力不稳定带来的挑战，我们可以在训练过程中直接引导注意力分布的熵趋向目标值。对于给定的注意力层ℓ，令 $A^\ell \in \mathbb{R}^{n\times n}$ 表示该层的注意力分数矩阵（经过 Softmax 后），其中 $n = 1 + n_s + n_a$ 为总序列长度， n_s 为状态维度， n_a为动作维度，1 对应于[VALUE］词元。第 i个查询词元的注意力分布熵为：

$$H_i^\ell = -\sum_{j=1}^{n} A_{ij}^\ell \log A_{ij}^\ell \quad (2)$$

受最大熵强化学习（Haarnoja等人，2018年）的启发，我们引入一个可学习的温度参数 α 并对其进行优化，以维持目标熵 $\bar{H}$:

$$\mathcal{L}_{\text{temp}}(\alpha) = \frac{1}{L}\sum_{\ell=1}^{L} \alpha^\ell \left(H^\ell - \bar{H}\right) \quad (3)$$

$$\mathcal{L}_{\text{attn}}(\phi) = -\frac{1}{L}\sum_{\ell=1}^{L} \alpha^\ell H^\ell \quad (4)$$

其中 $H^\ell = \frac{1}{n}\sum_{i=1}^{n} H_i^\ell$。在实践中，我们将 α参数化为指数形式 $\exp(\hat{\alpha})$，并对 $\hat{\alpha}$ 进行优化。通过最大化注意力损失中的注意力熵，网络会受到鼓励，分散注意力以关注所有输入词元，同时温度项会将熵调整至期望水平，从而确保稳定性。温度参数与网络参数通过梯度下降联合优化，以稳定注意力熵：

$$\mathcal{L}_{\text{critic}}(\phi,\alpha) = \mathcal{L}_Q(\phi) + \mathcal{L}_{\text{attn}}(\phi) + \mathcal{L}_{\text{temp}}(\alpha) \quad (5)$$

在应用注意力熵控制时，我们做出了两项重要的设计选择。首先，我们使用逐层温度参数 α^ℓ，而非单一的全局参数。Transformer 中的不同层可以学会关注不同的模式。让每一层都维持各自的目标熵，可以使模型学会适用于价值函数学习的、各层独有的注意力模式。其次，我们使用单独的温度参数 $\alpha^\ell_{\texttt{[VALUE]}}$ 来控制 [VALUE］词元的注意力熵；该词元会汇总其他所有词元的信息，以生成最终的 Q 值。由于 [VALUE]词元负责汇总所有状态词元和动作词元的信息，从而生成准确的价值估计，因此它可以具有不同的注意力熵模式。这些设计选择使每一层和 [VALUE］词元都能独立维持适当的熵水平，从而实现稳定训练。

自动熵控制机制使 TQL 能够通过直接将熵调节到期望值来防止熵坍塌，因为温度系数会自适应地调节注意力分布中的熵，从而从所有词元中提取信息，同时仍允许模型学习哪些词元与价值预测最相关。这样一来，更大的模型就能利用其额外容量学习最有用的信号，而不会发生坍塌。这种¯方法引入了一个额外的超参数 H，用于设定目标熵值。实践中，我们会在不同任务和环境中使用相似的¯取值 H，并在第C.3节提供如何选择该超参数取值的建议。

4.3. 实现细节

可学习模态嵌入。 除了稳定注意力分数的熵之外，我们还希望更有效地利用大型模型的容量。为此，我们引入可学习模态嵌入 e_s 和 e_a，分别添加到状态词元和动作词元中。这些模态嵌入为模型提供了一种简单的机制，使其能够区分状态信息和动作信息。结合位置嵌入，这使大型模型能够根据 Q 值预测所需的相关信息，将注意力集中在状态或动作上。

算法 1 Transformer Q 学习（TQL）

输入：离线数据集 $\mathcal{D}$，目标熵系数 H，层数 L 使用 Transformer 架构初始化 Q 网络 Q_ϕ 使用 $\phi' \leftarrow \phi$初始化目标网络$Q_{\phi'}$ 初始化策略网络 π_ω、π_θ^β，以及温度参数 $\alpha = \{\alpha^\ell, \alpha^\ell_{\texttt{[VALUE]}}\}_{\ell=1}^L$对于每次训练迭代，执行以下操作：采样批次 ($s$、$a$、$r$、$s'$) $\sim \mathcal{D}$▷训练评论家 采样下一步动作 $a' \sim \pi_\theta(a'|s')$ 使用公式 (1) 计算时序差分损失 $\mathcal{L}_Q$ 计算注意力熵 $\{H$ 过 度 标 (7) $\mathcal{L}_{\mathrm{B}}$

策略提取。 原则上，TQL 可应用于任何基于价值的策略提取方案，且没有特定于方法的设计决策。我们选择了一种近期高性能的离线强化学习策略提取方案（Park 等人，2025b）：我们学习一个单步流策略 π_ω，该策略通过向行为克隆（BC）流策略 π^β 进行蒸馏而受到行为约束，同时引导其最大化 Q 值：

$$\mathcal{L}_{\mathrm{BC}}(\theta) = \mathbb{E}_{\substack{s,a=x^1\sim\mathcal{D},\\ x^0\sim\mathcal{N}(0,I_d),\\ t\sim\mathrm{Unif}([0,1])}}\left[\|\pi_\theta^\beta(t,s,x^t) - (x^1 - x^0)\|_2^2\right], \quad (6)$$

$$\mathcal{L}_{\mathrm{OS}}(\omega) = \underbrace{\mathbb{E}_{s\sim\mathcal{D},a^\pi\sim\pi_\omega}[-Q_\phi(s,a^\pi)]}_{\mathrm{Q}} + \underbrace{\alpha\mathcal{L}_{\mathrm{Distill}}(\omega)}_{\mathrm{BC}}, \quad (7)$$

$$\mathcal{L}_{\mathrm{Distill}}(\omega) = \mathbb{E}_{\substack{s\sim\mathcal{D},\\ z\sim\mathcal{N}(0,I_d)}}\left[\|\mu_\omega(s,z) - \mu_\theta(s,z)\|_2^2\right]. \quad (8)$$

其中 α 控制策略上行为约束的强度。这使得策略在仍最大化价值的同时受到行为约束。

5. 实验

本节通过在各种具有挑战性的任务上进行实验，聚焦于扩展性和性能，对比评估 TQL 与以往的强化学习方法。我们的实验旨在回答以下关键问题：

(Q1) 是什么阻碍了 Transformer 在价值函数方面实现良好的扩展？**(Q2)** 与以往方法相比，在价值函数设置下，TQL 是否能提升 Transformer 可扩展性？**(Q3)** TQL 能否从静态离线数据集中有效学习策略？**(Q4)** TQL 最重要的组成部分是什么？

5.1. 实验设置

我们在最近提出的 OGBench 基准（Park 等人，2025a）上评估 TQL。要在 OGBench 上取得成功，离线强化学习算法需要在较长的时间跨度上进行推理，从非结构化数据中有效学习，并在复杂的多阶段交互中处理延迟信用分配，这使其成为评估离线强化学习算法的严格测试平台。由于我们处于标准的奖励最大化强化学习设定中，我们采用基于奖励的单任务变体（记为“singletask”）。我们的评估涵盖五个领域，每个领域包含五个不同的任务：`cube-double`、`cube-triple`、`scene`、`puzzle-3×3`和`puzzle-4×4`，共计 25 个任务。各领域见图14。

比较。 在实验中，我们将 9 种离线强化学习方法作为对比，这些方法涵盖了广泛的算法和建模策略，包括采用基于流匹配和基于 Transformer 的价值函数的方法。对于使用高斯策略的标准离线强化学习对比，我们纳入了行为克隆（BC）、隐式Q学习（IQL）（Kostrikov等人，2021年）和 ReBRAC（Tarasov等人，2023年），作为应用广泛且具有竞争力的代表方法。为涵盖较新的前沿进展，我们还评估了在离线强化学习中使用流策略的方法，包括 FBRAC（它是采用流策略的行为正则化演员-评论家的一种变体）（Wu 等人，2019）和 IFQL（它是采用流策略的隐式扩散Q学习的一种变体）（Hansen-Estruch 等人，2023年），以及流Q学习（FQL）（Park 等人，2025b），后者使用单步流策略来最大化Q值。由于我们的方法侧重于价值学习，因此我们纳入了基于流的价值学习基线（floq（Agrawalla等人，2025年））和基于 Transformer 的价值学习基线（Q-Transformer（Q-T）（Chebotar 等人，2023年）和感知器演员-评论家（PAC）（Springenberg 等人，2024年））。对于 PAC 和 Q-T，我们均沿用其原始价值函数架构和训练设置，同时采用与我们的方法相同的、基于 FQL 的策略提取流程，以确保比较公平且受控。有关基线的具体实现和超参数，详见附录 C.1。

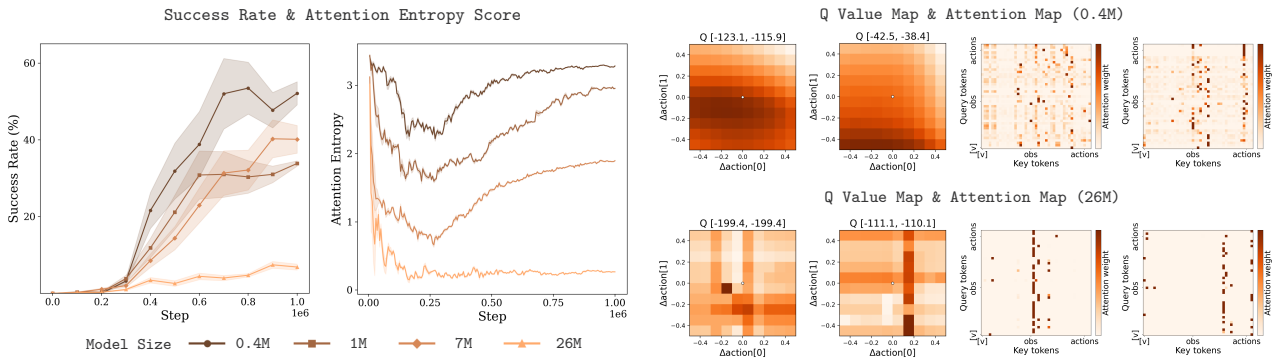

图 *3.* **扩展用于价值函数的 Transformer 会导致熵坍塌和性能下降。** 左：未采用 TQL 的 Transformer 模型在不同模型规模下，于五个立方体双重任务中的成功率和注意力熵可视化结果。右：最小（0.4M）和最大（26M）模型的 Q值景观和注意力图。较大的 Transformer 学到的价值曲面极不平滑，并呈现出其较小对应模型中没有的高频振荡和不连续性。

	MLP Q						Flow Q	Transformer Q		
	BC	IQL	ReBRAC	FBRAC	IFQL	FQL	floq	Q-T	PAC	TQL
立方体双人游戏单任务任务1-v0	8 ± 3	27 ± 5	45 ± 6	47 ± 11	35 ± 9	61 ± 9	45 ± 23	2 ± 1	28 ± 16	$\mathbf{95 \pm 2}$
立方体双人游戏单任务任务2-v0	0 ± 0	1 ± 1	7 ± 3	22 ± 12	9 ± 5	36 ± 6	42 ± 23	0 ± 0	36 ± 16	$\mathbf{84 \pm 8}$
立方体双人游戏单任务任务3-v0	0 ± 0	0 ± 0	4 ± 1	4 ± 2	8 ± 5	22 ± 5	$\mathbf{59 \pm 15}$	0 ± 0	50 ± 11	$\mathbf{60 \pm 27}$
立方体双人游戏单任务任务4-v0	0 ± 0	0 ± 0	1 ± 1	0 ± 1	1 ± 1	5 ± 2	13 ± 8	0 ± 0	10 ± 6	$\mathbf{14 \pm 6}$
立方体双人游戏单任务任务5-v0	0 ± 0	4 ± 3	4 ± 2	2 ± 2	17 ± 6	19 ± 10	46 ± 11	0 ± 0	25 ± 21	$\mathbf{84 \pm 9}$
立方体双人游戏（平均值）	2 ± 1	6 ± 2	12 ± 3	15 ± 6	14 ± 5	29 ± 6	41 ± 16	0 ± 0	30 ± 14	$\mathbf{67 \pm 10}$
场景游戏单任务任务1-v0	19 ± 6	94 ± 3	$\mathbf{95 \pm 2}$	$\mathbf{96 \pm 8}$	$\mathbf{98 \pm 3}$	$\mathbf{100 \pm 0}$	$\mathbf{100 \pm 0}$	27 ± 6	$\mathbf{100 \pm 0}$	$\mathbf{100 \pm 0}$
场景游戏单任务任务2-v0	1 ± 1	12 ± 3	50 ± 13	46 ± 10	0 ± 0	76 ± 9	$\mathbf{92 \pm 9}$	6 ± 3	54 ± 16	81 ± 14
场景游戏单任务任务3-v0	1 ± 1	32 ± 7	55 ± 16	78 ± 4	54 ± 19	$\mathbf{98 \pm 1}$	$\mathbf{97 \pm 4}$	1 ± 1	$\mathbf{98 \pm 2}$	$\mathbf{98 \pm 2}$
场景游戏单任务任务4-v0	2 ± 2	0 ± 1	3 ± 3	4 ± 4	0 ± 0	5 ± 1	4 ± 4	1 ± 1	4 ± 3	$\mathbf{25 \pm 17}$
场景游戏单任务任务5-v0	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 1}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$
scene-play（平均值）	5 ± 2	28 ± 3	41 ± 7	45 ± 5	30 ± 4	56 ± 2	$\mathbf{59 \pm 3}$	7 ± 2	51 ± 4	$\mathbf{61 \pm 7}$
3×3拼图单任务任务1-v0	5 ± 2	33 ± 6	$\mathbf{97 \pm 4}$	63 ± 19	$\mathbf{94 \pm 3}$	90 ± 4	$\mathbf{94 \pm 5}$	3 ± 2	92 ± 4	$\mathbf{96 \pm 4}$
3×3拼图单任务任务2-v0	1 ± 1	4 ± 3	1 ± 1	2 ± 2	1 ± 2	16 ± 5	$\mathbf{22 \pm 9}$	1 ± 1	17 ± 12	19 ± 12
3×3拼图单任务任务3-v0	1 ± 1	3 ± 2	3 ± 1	1 ± 1	0 ± 0	10 ± 3	$\mathbf{16 \pm 5}$	0 ± 1	8 ± 4	7 ± 5
3×3拼图单任务任务4-v0	1 ± 1	2 ± 1	2 ± 1	2 ± 2	0 ± 0	$\mathbf{16 \pm 5}$	$\mathbf{16 \pm 10}$	0 ± 0	14 ± 3	10 ± 4
3×3拼图单任务任务5-v0	1 ± 0	3 ± 2	5 ± 3	2 ± 2	0 ± 0	16 ± 3	25 ± 7	0 ± 1	9 ± 3	$\mathbf{30 \pm 10}$
3×3拼图（平均值）	2 ± 1	9 ± 3	22 ± 2	14 ± 5	19 ± 1	30 ± 4	$\mathbf{35 \pm 7}$	1 ± 1	28 ± 5	32 ± 7
4×4拼图单任务任务1-v0	1 ± 1	12 ± 2	26 ± 4	32 ± 9	49 ± 9	34 ± 8	$\mathbf{60 \pm 8}$	0 ± 0	14 ± 7	$\mathbf{59 \pm 11}$
4×4拼图单任务任务2-v0	0 ± 0	7 ± 4	12 ± 4	5 ± 3	4 ± 4	16 ± 5	$\mathbf{22 \pm 5}$	0 ± 1	13 ± 5	19 ± 6
4×4拼图单任务任务3-v0	0 ± 0	9 ± 3	15 ± 3	20 ± 10	$\mathbf{50 \pm 14}$	18 ± 5	36 ± 7	0 ± 0	8 ± 7	43 ± 5
4×4拼图单任务任务4-v0	0 ± 0	5 ± 2	10 ± 3	5 ± 1	$\mathbf{21 \pm 11}$	11 ± 3	19 ± 4	0 ± 0	3 ± 2	$\mathbf{20 \pm 3}$
4×4拼图单任务任务5-v0	0 ± 0	4 ± 1	7 ± 3	2 ± 2	2 ± 2	7 ± 3	14 ± 3	0 ± 0	5 ± 3	$\mathbf{16 \pm 5}$
4×4拼图（平均值）	0 ± 0	7 ± 2	14 ± 3	13 ± 5	25 ± 8	17 ± 5	$\mathbf{30 \pm 5}$	0 ± 0	9 ± 5	$\mathbf{31 \pm 6}$
三阶魔方单任务任务1-v0	1 ± 1	4 ± 4	1 ± 2	0 ± 0	2 ± 2	20 ± 6	17 ± 11	0 ± 0	1 ± 1	$\mathbf{29 \pm 16}$
cube-triple-play-singletask-task2-v0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	$\mathbf{1 \pm 2}$	0 ± 0	0 ± 0	0 ± 0	0 ± 1
cube-triple-play-singletask-task3-v0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	1 ± 1	0 ± 0	0 ± 0	$\mathbf{5 \pm 4}$
cube-triple-play-singletask-task4-v0	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$
cube-triple-play-singletask-task5-v0	0 ± 0	$\mathbf{1 \pm 1}$	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0
cube-triple-play（平均值）	0 ± 0	1 ± 1	0 ± 0	0 ± 0	0 ± 0	4 ± 2	4 ± 2	0 ± 0	0 ± 0	$\mathbf{7 \pm 4}$
所有任务的平均值	2 ± 1	10 ± 2	18 ± 3	17 ± 4	18 ± 4	27 ± 4	34 ± 7	2 ± 1	24 ± 6	$\mathbf{40 \pm 7}$

表 *1.* **OGBench 评估结果。** TQL 在 5 个领域中的 4 个领域取得了最高的平均性能，并且在全部 25 项任务上的平均性能也是最佳。报告的分数是 3 个随机种子的平均值，其中粗体数值表示该任务上的性能达到最佳结果的 95%。

5.2. 是什么阻碍 Transformer 在价值函数方面有效扩展?

尽管 Transformer 已在众多领域展现出可扩展性，但其在价值函数学习中的应用仍然有限。我们首先探究具体的 bot-

通过实证分析，探究阻碍基于 Transformer 的价值函数有效扩展的瓶颈。我们使用第 `cube-double` 领域中的 5 个任务开展扩展实验，采用第 4.1节所述的 Transformer 架构，在保持架构不变的同时改变模型容量。不同参数数量对应的训练曲线.....

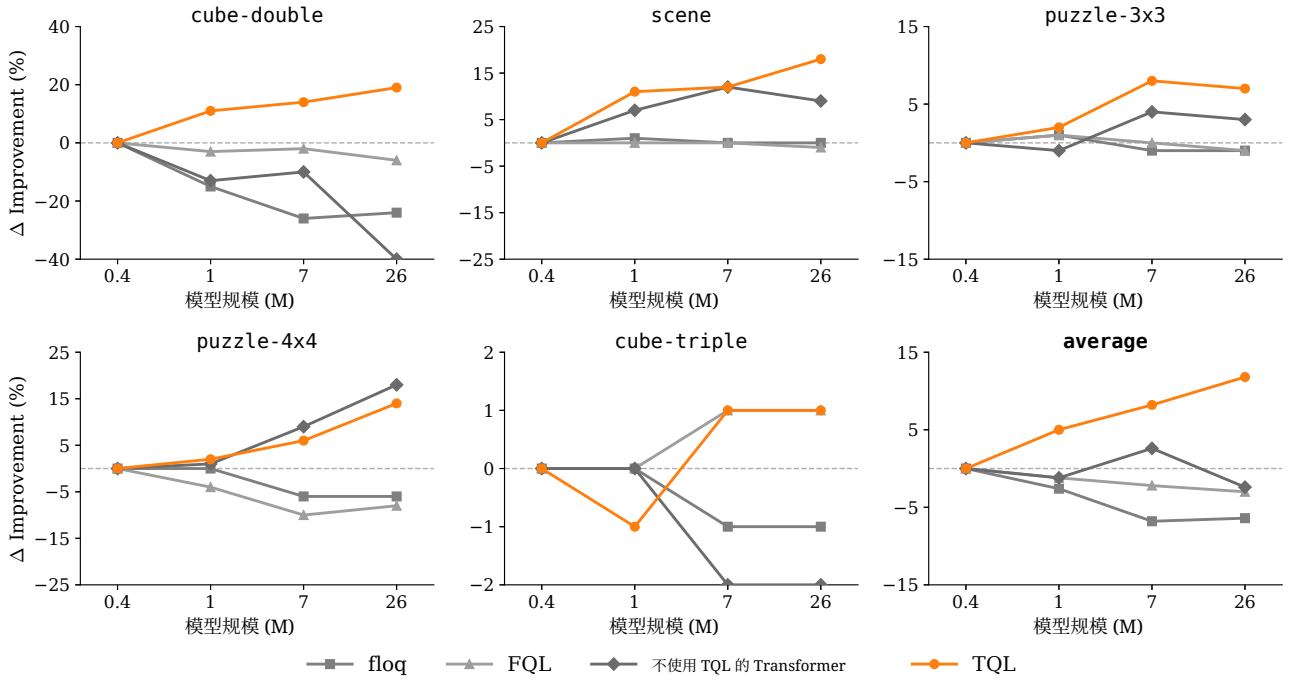

图 4。 **扩展结果**。 我们比较了 TQL 与以往的离线强化学习方法在评论家规模从 0.4M 到 26M 参数范围内的表现。该图报告了各方法相对于最小模型（0.4M）的平均成功率差异。尽管基线方法在规模增大时性能下降，TQL 在所有环境中都能持续良好扩展，并以较大幅度优于以往方法。

如图 3 （左）所示。与语言建模和视觉任务中观察到的典型扩展趋势相反，我们观察到明显的负向扩展模式：随着模型规模增大，性能反而下降，最大模型的表现很差。

为诊断这种失效模式，我们在图3 （中右）中通过等高线可视化分析 Q 值景观。我们观察到，网络越大，价值曲面就越不平滑，并呈现出高频振荡和不连续性，而较小的网络中则没有这些现象。这表明，较大的网络会出现不稳定性，过大的容量似乎有害。我们推测，这种性能退化源于 Transformer 中注意力机制的训练不稳定性。为检验这一假设，我们在图3（右）中考察了不同网络规模下的注意力分数分布。与假设一致，我们发现注意力熵随模型规模增大而显著降低，这表明较大的模型会学习到越来越尖锐且脆弱的注意力模式。这种熵坍塌与不平滑的价值景观及任务表现下降均高度相关，说明在价值函数学习场景中，若不加以改进便扩大规模，注意力机制就无法恰当地泛化。更多关于不同环境下不使用 TQL 的 Transformer 的注意力熵图和可视化结果，请参见第 A.1节和第 A.2节；这些结果呈现出一致的趋势。

5.3. 使用参数规模扩展 TQL

基于第 5.2节的实证分析，我们开展扩展实验，以考察 TQL 是否能缓解注意力熵坍塌，并实现 Q 函数的有效扩展。我们将 TQL 与离线强化学习场景中具有代表性的 Q 函数学习方法进行比较，涵盖多种生成模型骨干网络，包括 FQL （Park 等人，2025b）（多层感知机）、floq （Agrawalla等人，2025年）（流匹配）和 PAC （Springenberg 等人，2024年）（Transformer）。此外，我们还纳入了第 4.1节所述的 Transformer 基线，其架构与 TQL 相同，但不包含注意力熵正则化和模态标记。所有方法均在多种评论家网络规模下进行评估，参数量从 0.4M 到 26M 不等（0.4M、1M、7M 和 26M）。值得注意的是，传统上，强化学习方法使用 ∼1M 参数的网络来训练 Q 函数，既有研究也证明了这一点（Dong 等人，2025d；Agrawalla等人，2025年；Kostrikov等人，2021年；Fujimoto&Gu，2021年；Kumar 等人，2020年），我们的扩展实验结果也显示，基线方法在这些规模下表现最佳。

我们在图 4中展示了结果。从图中可以看出，在所有生成模型骨干网络（多层感知机、流匹配、Transformer）中，随着容量增加，先前的方法扩展效果不佳，从最小设置到最大设置，性能平均下降了 10.6%。这表明，现有的离线强化学习方法即使在离线强化学习基准测试上表现出色，也都难以从更大的价值函数容量中受益——

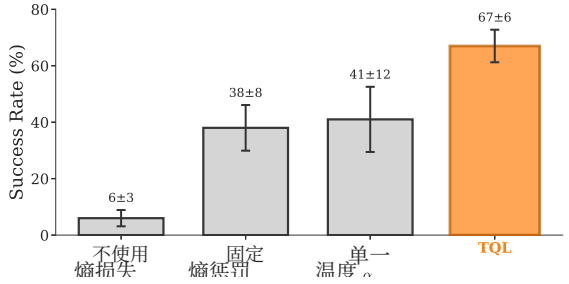

图*5.* **TQL 最重要组件的消融实验。** 我们在 cube-double 环境中将 TQL 与四种消融变体进行比较：(1) 不使用注意力熵引导的 Transformer 基线，(2) 使用固定熵惩罚，以及 (3) 使用单一温度，而非逐层和逐令牌温度。结果表示 cube-double 环境中 5 个任务和 3 个随机种子上的平均性能和标准误。

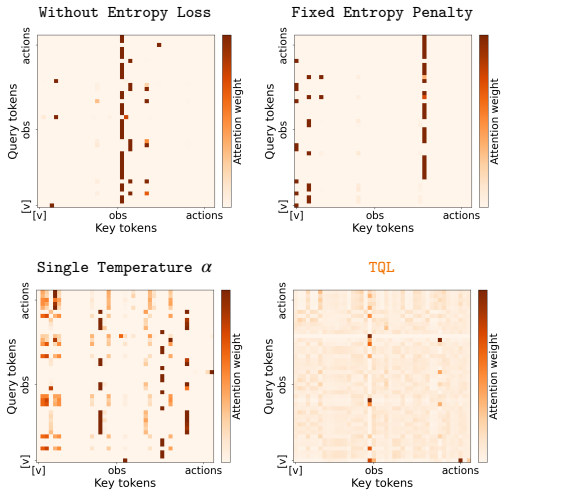

图 *6.* **消融实验的注意力图。** 我们发现，TQL 在防止注意力熵坍塌方面最为有效；而采用固定熵惩罚，以及在各层和词元间使用单一温度，都会导致一定程度的注意力过度专门化，并使性能变差。

与 Transformer基线相比，基线模型在扩展到 700 万个参数时，虽然在某些任务上的性能有所提升，但在达到 2600 万个参数时却有所下降。这种表现与注意力坍塌模式相符，因为 2600 万参数的模型出现了最严重的坍塌，且其 Q值景观高度不平滑。与先前的方法不同，TQL 缓解了这一失效模式，实现了稳定且持续的扩展；从最小模型到最大模型，性能提升了 43%，凸显出 TQL 能够有效利用更大的容量来提升性能。

5.4. 与先前离线强化学习方法的比较

在展示了 TQL 的扩展能力后，我们进一步在一组基准任务上进行评估，并与先前的方法进行比较，以考察 TQL 从离线数据集中学习策略的效果。我们将 TQL 与一系列全面的离线强化学习基线进行比较：对于在相同基准测试上评估的方法，报告其原论文中的数值；对于其他方法，则报告经过调优的超参数设置下的数值。更多详情请参阅第C.4 节。表 1总结了结果：TQL 在 5 个领域中的 4 个领域取得最佳性能，并且在全部 25 个任务上取得最佳平均性能，展现了其在广泛环境中的稳定提升。这些结果凸显了 TQL 通过防止 Transformer 中的注意力坍塌，在具有挑战性的任务中实现出色性能的有效性；因此，该方法既具备良好的扩展性，也达到了最先进的性能。

5.5. TQL 的哪些组件最重要?

在确认了扩展能力和性能之后，我们分析 TQL 的关键组成部分及其对整体性能的影响。我们考虑以下设置：(1) 不使用熵损失，即移除 $\mathcal{L}_{\text{attn}}$ 公式（5）中的损失项；(2) 固定熵惩罚，即用最大熵损失取代熵调整机制，以防止注意力坍塌；(3) 单一温度 α，即在所有层和词元中使用相同的温度。对于每种变体，我们报告其在 cube-double环境中的 5 个任务上的平均性能，使用的模型为 26M。

结果如图 5所示。首先，我们发现，完全移除熵引导会导致性能大幅下降，这是注意力坍塌造成的，正如我们在第5.2节所示。所有形式的熵调节都能通过防止注意力坍塌显著提升性能。我们发现，相较于将熵引导至目标值，采用最大熵损失来施加固定熵惩罚会导致性能显著变差，这可能是因为最大熵损失未能将熵引导至目标值，造成训练过程不稳定。此外，逐层和逐词元熵调节对于实现更稳定的训练也很重要，因为单一温度项难以将所有层的熵引导至同一目标值。图 6展示了消融实验的注意力图。从中可以看出，TQL 在所有状态词元和动作词元上的注意力最为均衡，因此扩展效果最佳，性能也最高。

6. 讨论与局限性

我们提出了 TQL，这是一种控制注意力熵的方法，使基于 Transformer 的价值函数能够在不发生坍塌的情况下扩展。我们的研究发现，注意力熵坍塌在基于 Transformer 的价值函数中很常见，并且这种坍塌与任务表现呈强烈负相关。

尽管我们的方法展现出有效的扩展能力，并在基准测试中取得了出色的表现，但它也存在局限性。在

实现复杂度方面，我们的方法增加了一个额外的损失项，以及一个用于目标熵的额外超参数。虽然这增加了一些复杂度和调参工作，但正如第C.3节所述，我们发现同一组超参数在实验中表现良好。与其他训练稳定性技巧不同，我们的方法不需要进行任何架构更改，因此无需付出高昂的重新训练成本即可使用预训练骨干网络。

7. 致谢

本研究部分得到了RAI研究所、美国海军研究办公室资助项目 N00014-22-1-2293、美国海军研究办公室资助项目 N00014-22-1-2621、美国国家科学基金会 #1941722 以及美国国家科学基金会 CAREER 奖的支持。

参考文献

Agrawalla, B., Nauman, M., Agrawal, K., and Kumar, A. floq：通过流匹配训练评论家网络，以扩展基于价值的强化学习中的计算规模，2025年。URLhttps://arxiv.org/abs/2509.06863。

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jackson, T., Jesmonth, S., Joshi, N. J., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, K.-H., Levine, S., Lu, Y., Malla, U., Manjunath, D., Mordatch, I., Nachum, O., Parada, C., Peralta, J., Perez, E., Pertsch, K., Quiambao, J., Rao, K., Ryoo, M., Salazar, G., Sanketi, P., Sayed, K., Singh, J., Sontakke, S., Stone, A., Tan, C., Tran, H., Vanhoucke, V., Vega, S., Vuong, Q., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., and Zitkovich, B. RT-1：用于大规模真实世界控制的机器人Transformer，2023年。URL https://arxiv.org/abs/2212.06817。

Castanyer, R. C., Obando-Ceron, J., Li, L., Bacon, P.-L., Berseth, G., Courville, A., and Castro, P. S. 深度强化学习中用于大规模稳定学习的稳定梯度，2025年。URLhttps://arxiv.org/abs/2506.15544。

Chebotar, Y., Vuong, Q., Irpan, A., Hausman, K., Xia, F., Lu, Y., Kumar, A., Yu, T., Herzog, A., Pertsch, K., Gopalakr- ishnan, K., Ibarz, J., Nachum, O., Sontakke, S., Salazar, G., Tran, H. T., Peralta, J., Tan, C., Manjunath, D., Singht, J., Zitkovich, B., Jackson, T., Rao, K., Finn, C., and Levine, S. Q-Transformer：通过自回归Q函数实现可扩展的离线强化学习，2023年。URLhttps://arxiv.org/abs/2309.10150。

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A. 和 Mordatch, I. 决策 Transformer：通过序列建模实现强化学习，2021年。URLhttps://arxiv.org/abs/2106.01345。

Cheng, J., Qiao, R., Ma, Y., Li, B., Xiong, G., Miao, Q., Li, Y. 和 Lv, Y. 通过联合优化的世界 - 动作模型预训练扩展离线基于模型的强化学习，2025年。URLhttps://arxiv.org/abs/2410.00564。

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A. P., Caron, M., Geirhos, R., Alabdulmohsin, I. 等。将视觉变换器扩展至220亿参数。载于国际机器学习会议，第7480–7512页。PMLR，2023年。

Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H. 和 Tang, J. CogView：通过 Transformer 掌握文本到图像生成，2021年。网址https://arxiv.org/abs/2105.13290。

Dong, P., Lessing, A. M., Chen, A. S. 和 Finn, C. 通过隐式模仿引导进行强化学习，2025a。网址 https://arxiv.org/abs/2506.07505。

Dong, P., Li, Q., Sadigh, D. 和 Finn, C. Expo：采用表达性策略的稳定强化学习，2025b。网址 https://arxiv.org/abs/2507.07986。

Dong, P., Mirchandani, S., Sadigh, D. 和 Finn, C. 机器人学中的批量在线强化学习，哪些因素最重要？，2025c。网址https://arxiv.org/abs/2505.08078。

Dong, P., Zheng, C., Finn, C., Sadigh, D., and Eysenbach, B. 价值流，2025d。网址 https://arxiv.org/abs/2510.07650。

Farebrother, J., Orbay, J., Vuong, Q., Taïga, A. A., Chebotar, Y., Xiao, T., Irpan, A., Levine, S., Castro, P. S., Faust, A., Kumar, A., and Agarwal, R. 停止回归：通过分类训练价值函数，以实现可扩展的深度强化学习，2024年。网址https://arxiv.org/abs/2403.03950。

Fujimoto, S. and Gu, S. S. 离线强化学习的极简主义方法，2021年。网址 https://arxiv.org/abs/2106.06860。

Fujimoto, S., Meger, D., and Precup, D. 无需探索的离策略深度强化学习，2019。网址 https://arxiv.org/abs/1812.02900。

Furuta, H.、Matsuo, Y. 和 Gu, S. S. 用于离线事后信息匹配的广义决策 Transformer，2022 年。网址：https://arxiv.org/abs/2111.10364。

Haarnoja, T., Zhou, A., Abbeel, P. 和 Levine, S. Soft actor-critic：使用随机策略进行离策略最大熵深度强化学习，2018年。链接https://arxiv.org/abs/1801.01290。

Hansen-Estruch, P., Kostrikov, I., Janner, M., Kuba, J. G. 和 Levine, S. IDQL：将隐式Q学习作为结合扩散策略的行为者-评论家方法，2023年。链接 https://arxiv.org/abs/2304.10573。

Janner, M., Li, Q. 和 Levine, S. 将离线强化学习视为一个大型序列建模问题，2021年。网址 https://arxiv.org/abs/2106.02039。

Kidambi, R., Rajeswaran, A., Netrapalli, P. 和 J oachims, T. Morel：基于模型的离线强化学-习，2021年。网址 https://arxiv.org/abs/2005.05951。

Kostrikov, I., Nair, A. 和 Levine, S. 使用隐式Q学习进行离线强化学习，2021年。网址 https://arxiv.org/abs/2110.06169。 Kumar, A., Zhou, A., Tucker, G. 和 Levine, S. 用于离线强化学习的保守Q学习，2020年。网址 https://arxiv.org/abs/2006.04779。 Lee, H., Hwang, D., Kim, D., Kim, H., Tai, J. J., Subramanian, K., Wurman, P. R., Choo, J., Stone, P. 和 Seno, T. Si mba：深度强化学习中用于扩展参数规模的简约偏差，2025年。网址https://arxiv.org/abs/2410.09754。 Levine, S., Kumar, A., Tucker, G. 和 Fu, J. 离线强化学习：教程、综述及开放问题展望，2020年。网址 https://arxiv.org/abs/2005.01643。 Nair, A., Gupta, A., Dalal, M. 和 Levine, S. AWAC：利用离线数据集加速在线强化学习，2021年。网址 https://arxiv.org/abs/2006.09359。 Nauman, M., Ostaszewski, M., Jankowski, K., Miło´s, P. 和 Cygan, M. 更大、更规范化、更乐观：面向计算效率与样本效率连续控制的扩展，2024年。网址 https://arxiv.org/abs/2405.16158。 Obando-Ceron, J., Sokar, G., Willi, T., Lyle, C., Farebrother, J., Foerster, J., Dziugaite, G. K., Precup, D. 和 Castro, P. S. 专家混合为深度强化学习解锁参数扩展，2024年。网址 https://arxiv.org/abs/2402.08609。 OpenAI, :, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., M.adry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Paino, A., Renzin, A., Passos, A. T., Kirillov, A., Christakis, A., Conneau, A., Kamali, A., Jabri, A., Moyer, A., Tam, A., Crookes, A., Tootoochian, A., Tootoonchian, A., Kumar, A., Vallone, A., Karpathy, A., Braunstein, A., Cann, A., Codispoti, A., Galu, A., Kon- drich, A., Tulloch, A., Mishchenko, A., Baek, A., Jiang, A., Pelisse, A., Woodford, A., Gosalia, A., Dhar, A., Pan- tuliano, A., Nayak, A., Oliver, A., Zoph, B., Ghorbani, B., Leimberger, B., Rossen, B., Sokolowsky, B., Wang, B., Zweig, B., Hoover, B., Samic, B., McGrew, B., Spero, B., Giertler, B., Cheng, B., Lightcap, B., Walkin, B., Quinn, B., Guarraci, B., Hsu, B., Kellogg, B., Eastman, B., Lu- garesi, C., Wainwright, C., Bassin, C., Hudson, C., Chu, C., Nelson, C., Li, C., Shern, C. J., Conger, C., Barette, C., Voss, C., Ding, C., Lu, C., Zhang, C., Beaumont, C., Hallacy, C., Koch, C., Gibson, C., Kim, C., Choi, C., McLeavey, C., Hesse, C., Fischer, C., Winter, C., Czar- necki, C., Jarvis, C., Wei, C., Koumouzelis, C., Sherburn, D., Kappler, D., Levin, D., Levy, D., Carr, D., Farhi, D., Mely, D., Robinson, D., Sasaki, D., Jin, D., Valladares, D., Tsipras, D., Li, D., Nguyen, D. P., Findlay, D., Oiwoh, E., Wong, E., Asdar, E., Proehl, E., Yang, E., Antonow, E., Kramer, E., Peterson, E., Sigler, E., Wallace, E., Brevdo, E., Mays, E., Khorasani, F., Such, F. P., Raso, F., Zhang, F., von Lohmann, F., Sulit, F., Goh, G., Oden, G., Salmon, G., Starace, G., Brockman, G., Salman, H., Bao, H., Hu, H., Wong, H., Wang, H., Schmidt, H., Whitney, H., Jun, H., Kirchner, H., de Oliveira Pinto, H. P., Ren, H., Chang, H., Chung, H. W., Kivlichan, I., O'Connell, I., O'Connell, I., Osband, I., Silber, I., Sohl, I., Okuyucu, I., Lan, I., Kostrikov, I., Sutskever, I., Kanitscheider, I., Gulrajani, I., Coxon, J., Menick, J., Pachocki, J., Aung, J., Betker, J., Crooks, J., Lennon, J., Kiros, J., Leike, J., Park, J., Kwon, J., Phang, J., Teplitz, J., Wei, J., Wolfe, J., Chen, J., Harris, J., Varavva, J., Lee, J. G., Shieh, J., Lin, J., Yu, J., Weng, J., Tang, J., Yu, J., Jang, J., Candela, J. Q., Beut- ler, J., Landers, J., Parish, J., Heidecke, J., Schulman, J., Lachman, J., McKay, J., Uesato, J., Ward, J., Kim, J. W., Huizinga, J., Sitkin, J., Kraaijeveld, J., Gross, J., Ka- plan, J., Snyder, J., Achiam, J., Jiao, J., Lee, J., Zhuang, J., Harriman, J., Fricke, K., Hayashi, K., Singhal, K., Shi, K., Karthik, K., Wood, K., Rimbach, K., Hsu, K., Nguyen, K., Gu-Lemberg, K., Button, K., Liu, K., Howe, K., Muthukumar, K., Luther, K., Ahmad, L., Kai, L., Itow, L., Workman, L., Pathak, L., Chen, L., Jing, L., Guy, L., Fedus, L., Zhou, L., Mamitsuka, L., Weng, L., McCal- lum, L., Held, L., Ouyang, L., Feuvrier, L., Zhang, L., Kondraciuk, L., Kaiser, L., Hewitt, L., Metz, L., Doshi, L., Aflak, M., Simens, M., Boyd, M., Thompson, M., Dukhan, M., Chen, M.,Gray,M., Hudnall,M., Zhang, M., Aljubeh, M., Litwin, M., Zeng, M., Johnson, M., Shetty, M., Gupta, M., Shah, M., Yatbaz, M., Yang, M. J., Zhong, M., Glaese, M., Chen, M., Janner, M., Lampe, M., Petrov, M., Wu, M., Wang, M., Fradin, M., Pokrass, M., Castro, M., de Castro, M. O. T., Pavlov, M., Brundage, M., Wang, M., Khan, M., Murati, M., Bavarian, M., Lin, M., Yesil- dal, M., Soto, N., Gimelshein, N., Cone, N., Staudacher, N., Summers, N., LaFontaine, N., Chowdhury, N., Ryder,

N., Stathas, N., Turley, N., Tezak, N., Felix, N., Kudige, N., Keskar, N., Deutsch, N., Bundick, N., Puckett, N., Nachum, O., Okelola, O., Boiko, O., Murk, O., Jaffe, O., Watkins, O., Godement, O., Campbell-Moore, O., Chao, P., McMillan, P., Belov, P., Su, P., Bak, P., Bakkum, P., Deng, P., Dolan, P., Hoeschele, P., Welinder, P., Tillet, P., Pronin, P., Tillet, P., Dhariwal, P., Yuan, Q., Dias, R., Lim, R., Arora, R., Troll, R., Lin, R., Lopes, R. G., Puri, R., Miyara, R., Leike, R., Gaubert, R., Zamani, R., Wang, R., Donnelly, R., Honsby, R., Smith, R., Sahai, R., Ramchandani, R., Huet, R., Carmichael, R., Zellers, R., Chen, R., Chen, R., Nigmatullin, R., Cheu, R., J ain, S., Altman, S., Schoenholz, S., Toizer, S., Miserendino, S., Agarwal, S., Culver, S., Ethersmith, S., Gray, S., Grove, S., Metzger, S., Hermani, S., Jain, S., Zhao, S., Wu, S., Jomoto, S., Wu, S., Shuaiqi, Xia, Phene, S., Papay, S., Narayanan, S., Coffey, S., Lee, S., Hall, S., Balaji, S., Broda, T., Stramer, T., Xu, T., Gogineni, T., Christian- son, T., Sanders, T., Patwardhan, T., Cunninghman, T., Degry, T., Dimson, T., Raoux, T., Shadwell, T., Zheng, T., Underwood, T., Markov, T., Sherbakov, T., Rubin, T., Stasi, T., Kaftan, T., Heywood, T., Peterson, T., Walters, T., Eloundou, T., Qi, V., Moeller, V., Monaco, V., Kuo, V., Fomenko, V., Chang, W., Zheng, W., Zhou, W., Man- assra, W., Sheu, W., Zaremba, W., Patil, Y., Qian, Y., Kim, Y., Cheng, Y., Zhang, Y., He, Y., Zhang, Y., Jin, Y., Dai, Y., and Malkov, Y. GPT-4o系统卡片，2024年。URLhttps://arxiv.org/abs/2410.21276。

Park, S., Frans, K., Eysenbach, B., 和 Levine, S. OGBench：离线目标条件强化学习基准测试，2025a。网址 https://arxiv.org/abs/2410.20092。

Park, S.、Li, Q. 和 Levine, S. 流Q学习。*arXiv* 预印本 *arXiv:2502.02538*，2025b。

Physical Intelligence, Amin, A., Aniceto, R., Balakrishna, A., Black, K., Conley, K., Connors, G., Darpinian, J., Dhabalia, K., DiCarlo, J., Driess, D., Equi, M., Esmail, A., Fang, Y., Finn, C., Glossop, C., Godden, T., Gory- achev, I., Groom, L., Hancock, H., Hausman, K., Hussein, G., Ichter, B., Jakubczak, S., Jen, R., Jones, T., Katz, B., Ke, L., Kuchi, C., Lamb, M., LeBlanc, D., Levine, S., Li-Bell, A., Lu, Y., Mano, V., Mothukuri, M., Nair, S., Pertsch, K., Ren, A. Z., Sharma, C., Shi, L. X., Smith, L., Springenberg, J. T., Stachowicz, K., Stoeckle, W., Swerd- low, A., Tanner, J., Torne, M., Vuong, Q., Walling, A., Wang,H., Williams, B., Yoo, S., Yu, L., Zhilinsky, U., 和 Zhou, Z. $\pi^*_{0.6}$：一种从经验中学习的视觉-语言-动作系统，2025a。网址 https://arxiv.org/abs/2511.14759。

Physical Intelligence、Black, K.、Brown, N.、Darpinian, J.、Dhabalia, K.、Driess, D.、Esmail, A.、Equi, M.、Finn, C.、Fusai, N.、Galliker, M. Y.、Ghosh, D.、Groom, L.、Hausman, K.、Ichter, B.、Jakubczak, S.、Jones, T.、Ke, L.、LeBlanc, D.、Levine,S.、Li-Bell, A.、Mothukuri, M.、Nair, S.、Pertsch, K.、Ren, A. Z.、Shi, L. X.、Smith, L.、Springenberg, J. T.、Stachowicz, K.、Tanner, J.、Vuong, Q.、Walke, H.、Walling, A.、Wang, H.、Yu, L. 和 Zhilinsky, U. $\pi_{0.5}$：一种具备开放世界泛化能力的视觉-语言-动作模型，2025b。网址 https://arxiv.org/abs/2504.16054。Rybakov, O.、Chrzanowski, M.、Dykas, P.、Xue, J. 和 Lanir, B. 改进大型语言模型训练稳定性的方法，2024年。网址 https://arxiv.org/abs/2410.16682。

Schwarzer, M.、Obando-Ceron, J.、Courville, A.、Bellemare, M.、Agarwal, R.与Castro, P. S.。更大、更好、更快：以人类水平的效率实现人类水平的雅达利，2023年。URL https://arxiv.org/abs/2305.19452。

Shazeer, N.。Glu变体改进Transformer，2020年。URLhttps://arxiv.org/abs/2002.05202。

Siméoni, O.、Vo, H. V.、Seitzer, M.、Baldassarre, F.、Oquab, M.、Jose, C.、Khalidov, V.、Szafraniec, M.、Yi, S.、Rama- monjisoa, M.、Massa, F.、Haziza, D.、Wehrstedt, L.、Wang, J.、Darcet, T.、Moutakanni, T.、Sentana, L.、Roberts, C.、Vedaldi, A.、Tolan, J.、Brandt, J.、Couprie, C.、Mairal, J.、Jégou, H.、Labatut, P.与Bojanowski, P.。Dinov3，2025年。URL https://arxiv.org/abs/2508.10104。

Springenberg, J. T., Abdolmaleki, A., Zhang,J., Groth, O., Bloesch, M., Lampe, T., Brakel, P., Bechtle, S., Kap- turowski, S., Hafner, R., Heess, N., and Riedmiller, M. 离线演员-评论家强化学习可扩展至大型模型，2024年。网址https://arxiv.org/abs/2402.05546。

Tarasov, D., Kurenkov, V., Nikulin, A., and Kolesnikov,S. 重访离线强化学习的极简主义方法，2023年。网址 https://arxiv.org/abs/2305.09836。

Team, G. R., Abdolmaleki, A., Abeyruwan, S., Ainslie, J., Alayrac, J.-B., Arenas, M. G., Balakrishna, A., Batchelor, N., Bewley, A., Bingham, J., Bloesch, M., Bousmalis, K., Brakel, P., Brohan, A., Buschmann, T., Byravan, A., Cabi, S., Caluwaerts, K., Casarini, F., Chan, C., Chang, O., Chappellet-Volpini, L., Chen, J. E., Chen, X., Chiang, H.-T. L., Choromanski, K., Collister, A., D'Ambrosio, D. B., Dasari, S., Davchev, T., Dave, M. K., Devin, C., Palo, N. D., Ding, T., Doersch, C., Dostmohamed, A., Du, Y., Dwibedi, D., Egambaram, S. T., Elabd, M., Erez, T., Fang, X., Fantacci, C., Fong, C., Frey, E., Fu, C., Gao, R., Giustina, M., Gopalakrishnan, K., Graesser, L., Groth, O., Gupta, A., Hafner, R., Hansen, S., Hasenclever, L., Haves, S., Heess, N., Hernaez, B., Hofer, A., Hsu, J., Huang, L., Huang, S. H., Iscen, A., Jacob, M. G., Jain, D., Jesmonth, S., Jindal, A., Julian, R., Kalashnikov, D., Karagozler, M. E., Karp, S., Kecman, M., Kew, J.C., Kim,

D., Kim, F., Kim, J., Kipf, T., Kirmani, S., Konyushkova, K., Ku, L. Y., Kuang, Y., Lampe, T., Laurens, A., Le, T. A., Leal, I., Lee, A. X., Lee, T.-W. E., Lever, G., Liang, J., Lin, L.-H., Liu, F., Long, S., Lu, C., Maddineni, S., Majumdar, A., Maninis, K.-K., Marmon, A., Martinez, S., Michaely, A. H., Milonopoulos, N., Moore, J., Moreno, R., Neunert, M., Nori, F., Ortiz, J., Oslund, K., Parada, C., Parisotto, E., Paryag, A., Pooley, A., Power, T., Quaglino, A., Qureshi, H., Raju, R. V., Ran, H., Rao, D., Rao, K., Reid, I., Rendleman, D., Rey mann, K., Rivas, M., Romano, F., Rubanova, Y., Sampedro, P. P., Sanketi, P. R., Shah, D., Sharma, M., Shea, K., Shridhar, M., Shu, C., Sindhwani, V., Singh, S., Soricut, R., Sterneck, R., Storz, I., Surdulescu, R., Tan, J., Tompson, J., Tunyasuvunakool, S., Varley, J., Vesom, G., Vezzani, G., Villalonga, M. B., Vinyals, O., Wagner, R., Wahid, A., Welker, S., Wohlhart, P., Wu, C., Wulfmeier, M., Xia, F., Xiao, T., Xie, A., Xie, J., Xu, P., Xu, S., Xu, Y., Xu, Z., Yan, J., Yang, S., Yang, S., Yang, Y., Yu, H. H., Yu, W., Yuan, W., Yuan, Y., Zhang, J., Zhang, T., Zhang, Z., Zhou, A., Zhou, G., and Zhou, Y. Gemini robotics 1.5：借助先进的具身推理、思考与运动迁移，推动通用型机器人的发展前沿，2025年。URLhttps://arxiv.org/abs/2510.03342。

Wagenmaker, A., Dong, P., Tsao, R., Finn, C., 和Levine, S. 后验行为克隆：预训练行为克隆策略以实现高效的强化学习微调，2025年。URL https://arxiv.org/abs/2512.16911。

Wang, K.,Javali, I., Bortkiewicz, M., Trzci´nski, T., 和 Eysenbach, B. 用于自监督强化学习的1000 层网络：扩展深度可以带来新的目标达成能力，2025年。URL https://arxiv.org/abs/2503.14858。

Wortsman, M., Liu, P. J., Xiao, L., Everett, K., Alemi, A., Adlam, B., Co-Reyes, J. D., Gur, I., Kumar, A., Novak, R., Pennington, J., Sohl-dickstein, J., Xu, K., Lee, J., Gilmer, J., 和 Kornblith, S. 大规模Transformer 训练不稳定性的小规模代理，2023年。URLhttps://arxiv.org/abs/2309.14322。

Wu, Y., Tucker, G., 和 Nachum, O. 行为正则化离线强化学习，2019. URL https://arxiv.org/abs/1911.11361.

Wu, Y., Zhai, S., Srivastava, N., Susskind, J., Zhang, J., Salakhutdinov, R., 和 Goh, H. 面向离线强化学习的不确定性加权演员-评论家，2021年. URLhttps://arxiv.org/abs/2105.08140.

Wu, Y.-H., Wang, X., 和 Hamaya, M. 弹性决策 Transformer，2023年。网址：https://arxiv.org/abs/2307.02484。

Yoshida, Y. 和 Miyato, T. 用于提升深度学习泛化能力的谱范数正则化，2017年。网址：https://arxiv.org/abs/1705.10941。

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., 和 Ma, T. Mopo：基于模型的离线策略优化，2020年. URL https://arxiv.org/abs/2005.13239.

Zhai, S., Likhomanenko, T., Littwin, E., Busbridge, D., Ramapuram, J., Zhang, Y., Gu, J., 和 Susskind, J. 通过防止注意力熵坍缩稳定 Transformer 训练，2023年. URLhttps://arxiv.org/abs/2303.06296.

Zhang，B. 和 Sennrich，R. 均方根层归一化，2019。网址https://arxiv.org/abs/1910.07467。

Zheng，Q.、Zhang，A. 和 Grover，A. 在线决策 Transformer，2022。网址 https://arxiv.org/abs/2202.05607。

Zhuo, Z., Zeng, Y., Wang, Y., Zhang, S., Yang, J., Li, X., Zhou, X. 和 Ma, J. Hybridnorm：通过混合归一化实现稳定高效的 Transformer 训练，2025年。网址https://arxiv.org/abs/2503.04598。

A. 完整缩放分析

A.1. 未使用 TQL 的 Transformer 完整注意力熵图

本节中，我们考察了未使用 TQL 的 Transformer 在全部五种环境中的注意力熵，如图 7 所示。我们观察到一个一致的趋势：随着模型规模增大，注意力熵逐渐降低，最大的模型最终坍塌至接近零。此实证证据证实了我们在分析中发现的注意力坍塌问题。

此外，我们发现，这些熵曲线能够很好地预测图 4 中报告的缩放行为。具体而言，图 `puzzle-3x3` 和 `puzzle-4x4` 环境在图 7 中表现出最轻微的熵坍塌。相应地，在图 4 中，未使用 TQL 的 Transformer 能够在这些相同任务上实现轻微的正向缩放。持续的注意力熵与扩展能力之间的强相关性进一步支持了我们的假设：注意力坍塌是模型规模增大时性能的主要瓶颈。

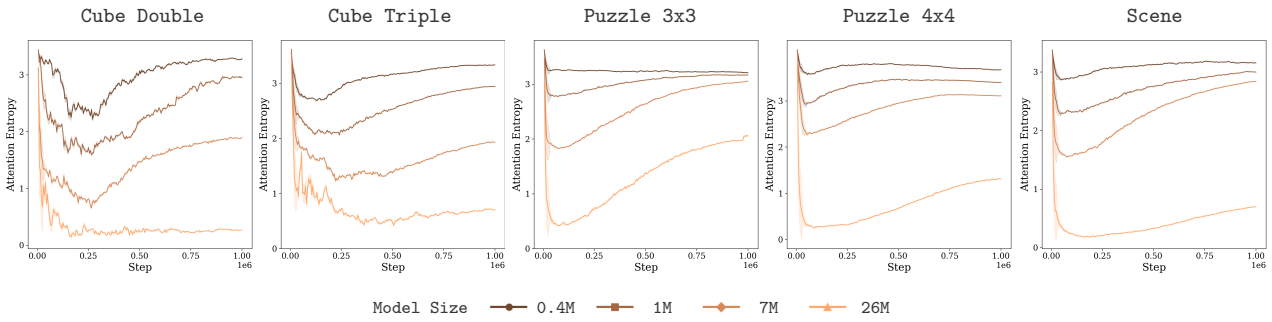

图 7。 **未使用 TQL 的 Transformer 注意力熵分析**。 我们展示了模型规模增大时，5 种环境中的注意力熵变化情况。图中呈现出一致的趋势：模型越大，熵越低，为注意力坍塌现象提供了实证证据。

A.2. 注意力图与Q值可视化

在本节中，我们展示了注意力图和Q值景观的综合可视化对比。我们将拥有2600万个参数的TQL与未采用TQL、且在未进行注意力熵调节的情况下训练的Transformer基线进行比较。为确保公平比较，我们在五个环境中使用从随机成功轨迹中提取的相同状态与动作对。

图 8 展示了未采用TQL的Transformer的结果。我们观察到，注意力模式变得极其稀疏，并高度集中于少数几个词元。这种坍缩导致Q值图呈现出高度不连续且不规则的特征。相反，图 9 表明，我们的训练目标成功缓解了这一问题。注意力机制保持了更广泛的分布，从而形成显著更平滑、更一致的Q值景观。

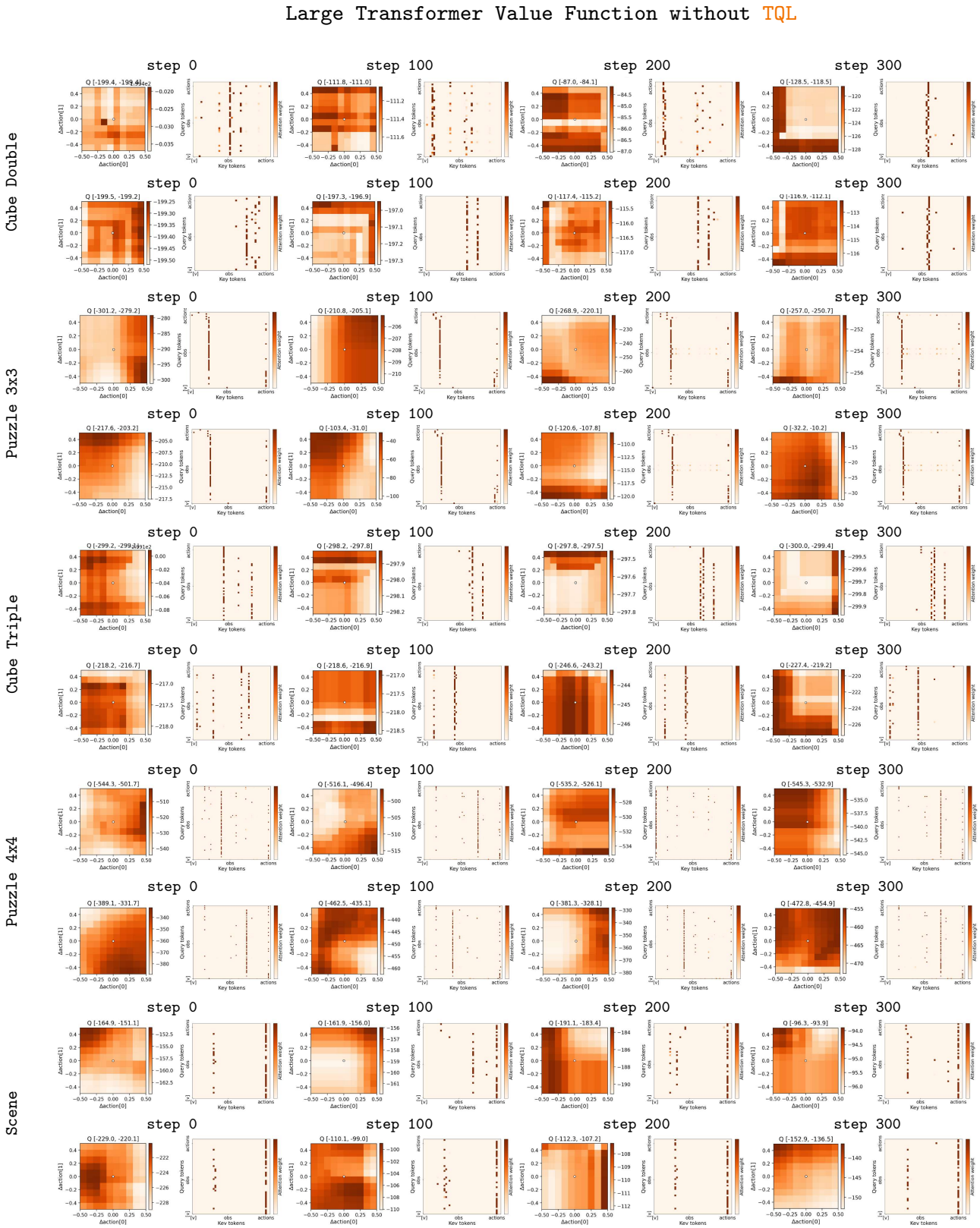

图 *8.* 未采用 TQL 的 Transformer 在全部五个环境中的注意力图和 Q值景观。学得的 Q 值函数呈现出不平滑的景观，并伴随着过拟合的注意力模式，从而阻碍了有效的价值学习。

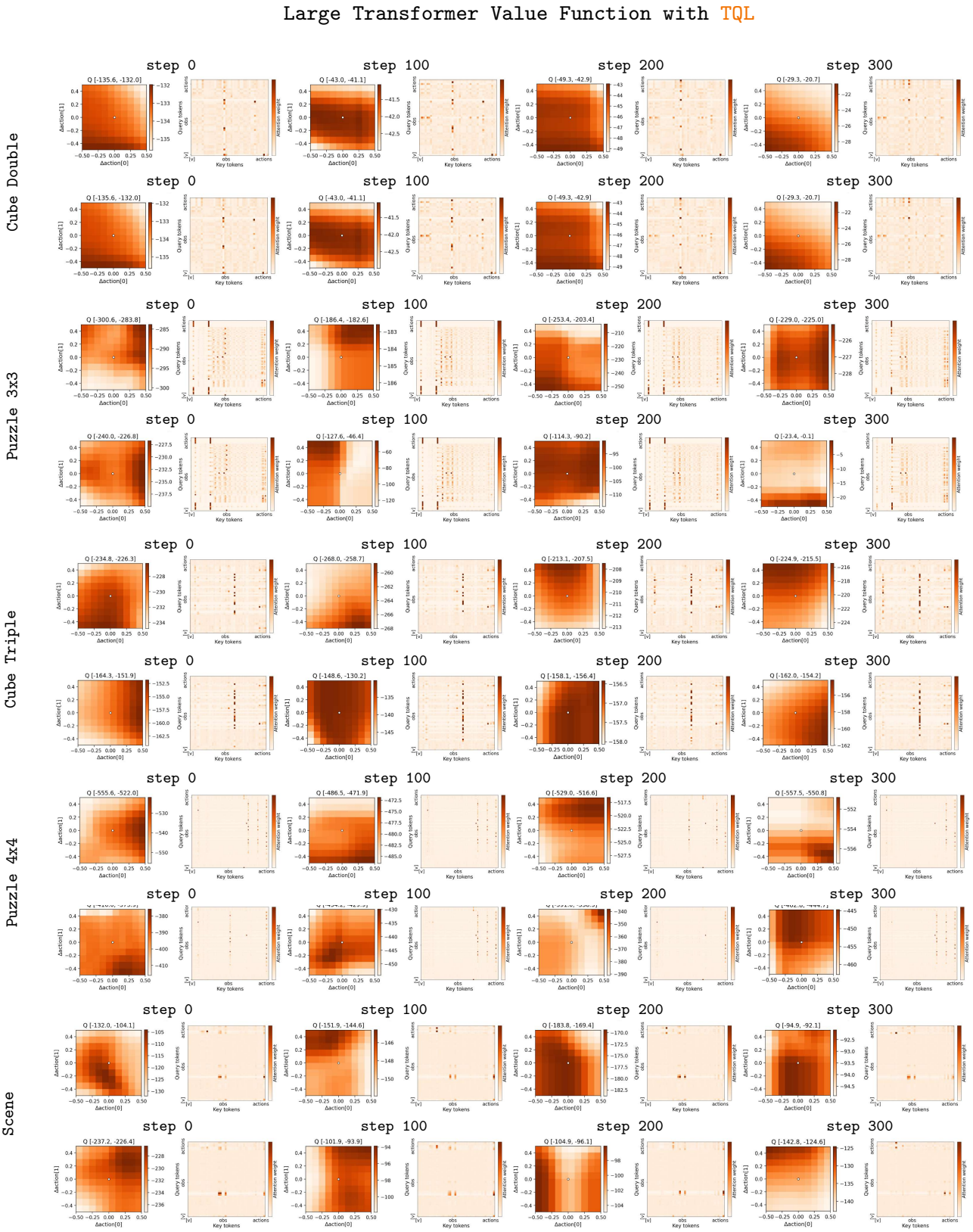

图 *9.* TQL 在全部五个环境中的注意力图和 Q值图。学得的 Q 值函数呈现平滑的景观，并伴随分布更广的注意力模式，从而促进稳定而有效的价值学习。

B. 其他实验

B.1. 增加深度

在我们的主要实验中，我们通过增加隐藏维度来扩展 Transformer 网络。本节将展示另一种常见的扩展方式，适用于第4.1节所述模型以及 TQL：增加网络深度。为此，我们进行了一项参数受控实验，将参数总数固定在最大设置附近，约为 26M。我们将网络深度增加到 4 层，并选择合适的隐藏维度，使参数总数尽可能接近 26M。在第5.3节的主要扩展实验中，我们遵循标准做法，让每个注意力头的隐藏维度保持不变；在我们的实验中，该维度为 32。结果见图10，注意力图见图11。结果表明，对于基线 Transformer，增加深度会导致与增加隐藏维度时相同的熵坍塌和性能不佳；这是预料之中的，因为增加深度与增加宽度类似，也会提高模型容量。与宽度设置下的情况一样，TQL 通过防止熵坍塌来提升性能。事实上，TQL 在 2 层和 4 层时的性能非常相近（67% 对 66%），这表明网络总规模是衡量性能的良好指标。

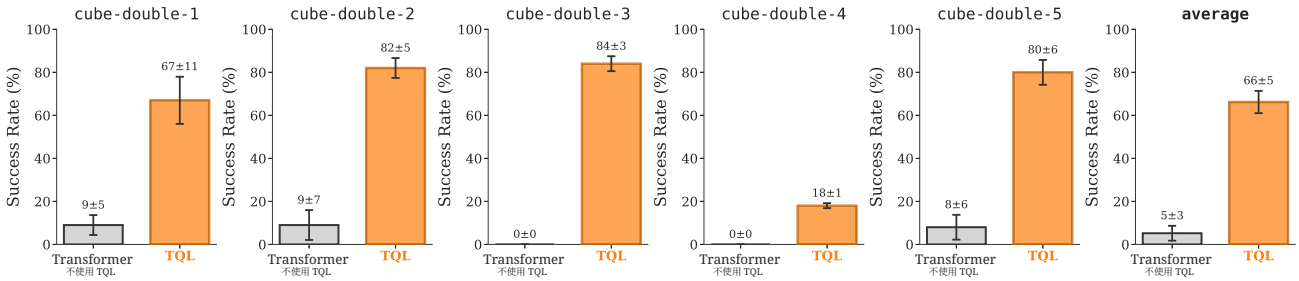

图*10*。 **增加层数后的 TQL 性能**。 不使用 TQL 的 Transformer 会像层数较少时一样发生注意力坍塌，而 TQL 可以解决这一问题并提升性能。我们发现，增加深度后，两种情况下的性能相近。

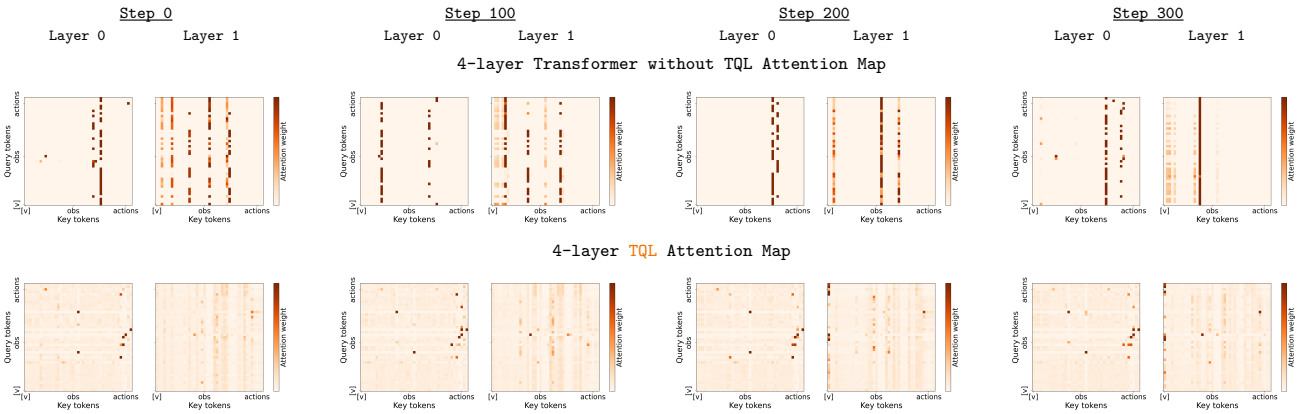

图*11*。 **深度增加后的 TQL 注意力图**。 我们发现，无论深度如何，不使用 TQL 时都会发生注意力坍塌，而使用 TQL 时注意力分数则更加均匀。

B.2. 与监督学习中稳定 Transformer 的方法比较

在监督学习中，人们提出了许多方法来提高大规模 Transformer 训练的稳定性。这些方法通常旨在让从业者突破训练稳定性的边界，例如以更高的学习率进行训练（Rybakov等人，2024年）；而在稳定的训练条件下，这些方法基本无效。

在本节中，我们将分析这些方法在价值函数训练场景下防止熵坍塌及其后续性能下降的效果，重点关注三种主要方法：

- QK 归一化（D Ehghani 等人，2023年），一种在查询和键投影之后都添加层归一化的方法 s

- σReparam（Zhai 等人，2023年），一种基于谱归一化的方法（Yoshida 和 Miyato，2017），旨在稳定注意力熵
- 一种包含用于稳定训练更大模型的一组技术的Transformer实现，其中包括RMS 归一化（Zhang & Sennrich，2019），夹心归一化（Ding 等人，2021年），QK 归一化（Dehghani 等人，2023年），以及SwiGLU（Shazeer，2020年）

我们在图 12 中展示了性能，并在图 13中展示了相应的注意力熵。从图中可以看出，这些方法有助于缓解熵坍塌，并相比Transformer基线获得更高的熵值和性能。然而，这些方法无法始终做到这一点，并且在应用于价值学习时总体表现不如TQL。

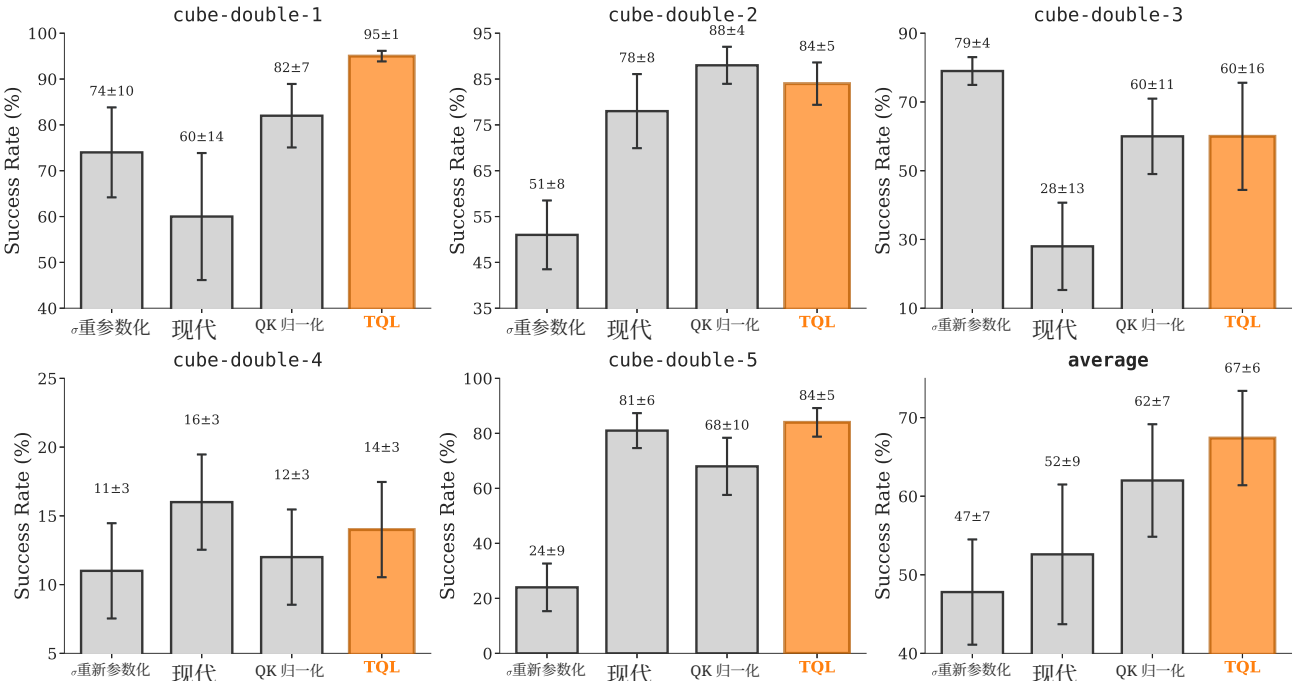

图 *12.* **其他Transformer稳定技术与TQL的性能对比**。与监督学习中用于稳定Transformer 训练的方法相比，TQL取得了更高的性能。

C. 实验细节

C.1. 基准测试与评估

我们在 OGBench 基准套件（Park 等人，2025a）上评估我们的方法。OGBench 是一个为离线目标条件强化学习设计的大规模基准，此外还提供与标准奖励最大化强化学习算法兼容的单任务变体。在这项工作中，我们对所有领域均采用单任务变体。对于每个任务，奖励范围为 $-n_{\text{task}}$ 至 0，具体取决于已完成子任务的数量。

我们对每个领域使用以下 OGBench 数据集：

- cube-double-play-v0
- cube-triple-play-v0
- scene-play-v0
- puzzle-3x3-play-v0
- puzzle-4x4-play-v0

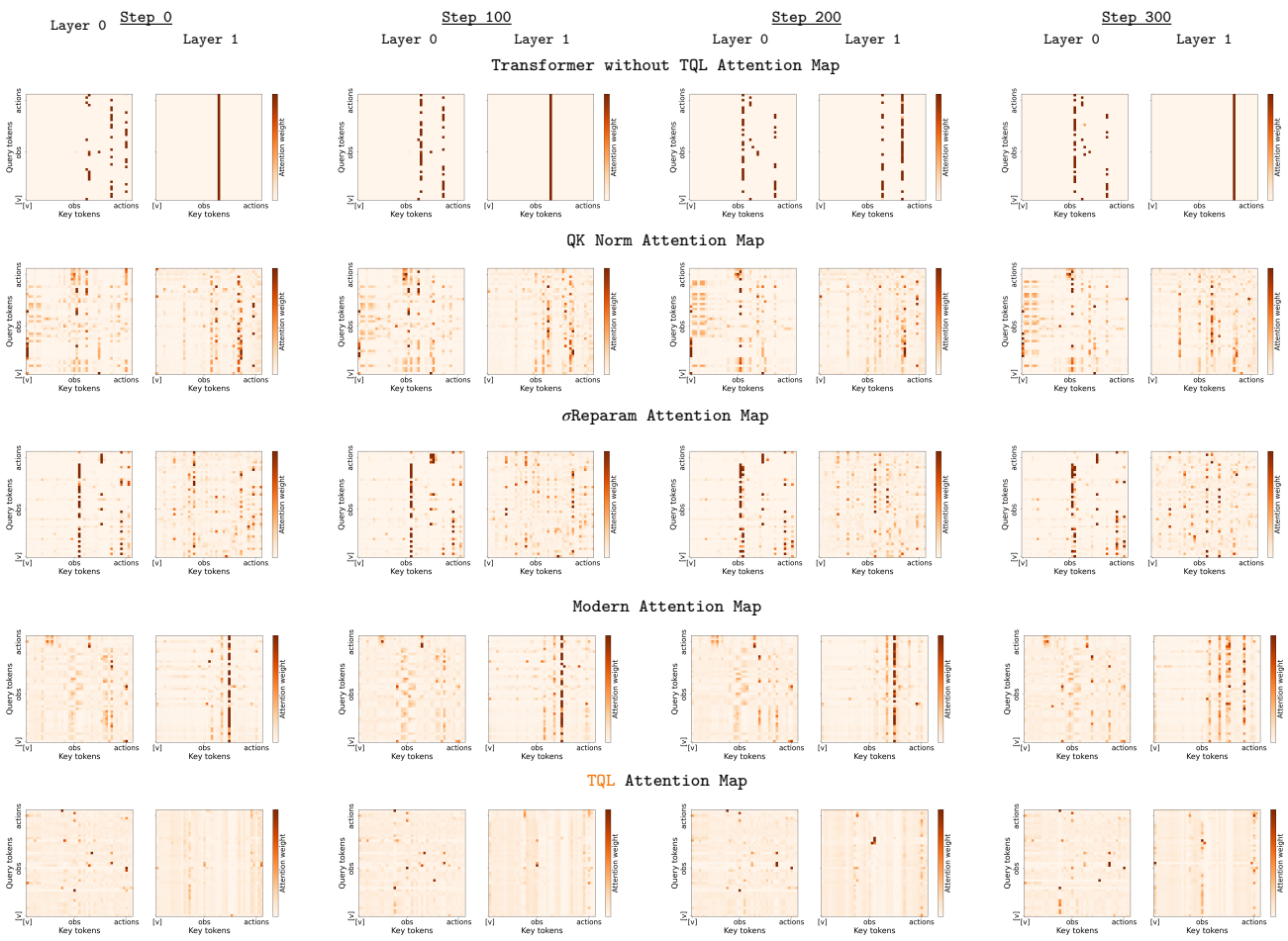

图 *13.* **与 TQL 相比，其他 Transformer 稳定化技术的注意力图。** 我们看到，与原始Transformer相比，监督学习中用于稳定 Transformer 的常见方法确实能提高熵。然而，我们也发现，某些层的熵仍会坍塌，而 TQL 的注意力模式则更加均匀。

该`cube-double` 和`cube-triple` 任务涉及对多个彩色立方体积木进行复杂的拾取与放置操作。 `scene` 任务要求进行长时程推理，并在杂乱环境中与各种物体交互。 `puzzle-3x3` 和`puzzle-4x4` 任务基于*Lights Out* 拼图，并使用机械臂求解，进一步评估智能体在组合状态空间上的泛化能力。所有环境的可视化结果如图14所示。

所有实验均遵循 OGBench（Park 等人，2025a）定义的官方评估流程和指标。对于每种设置，我们使用三个随机种子进行实验，并报告成功率的平均值和标准差，该指标取自 800k、900k 和1M 训练步检查点的平均值。对于柱状图，我们报告相同的平均值，并以标准误作为误差线。

C.2. 模型结构细节

我们使用 Transformer 对 Q 函数 $Q(s,\ a)$ 进行参数化，以词元级方式建模状态和动作各组成部分之间的交互。给定状态 $s \in \mathbb{R}^{n_s}$ 和动作 $a \in \mathbb{R}^{n_a}$，我们将状态和动作的每个标量维度分别视为一个独立词元，从而构造词元序列。具体而言， s 和 a 分别重塑为长度为 n_s 和n_a的序列，并通过学习得到的线性投影，将每个词元独立投影到维度为 H 的共享隐空间中。为区分状态词元和动作词元，我们向相应的词元表示添加可学习模态嵌入。随后，将状态词元序列与动作词元序列拼接，形成长度为 $n_s + n_a$的序列。

可学习的位置嵌入　　将嵌入添加到所有词元中，并在序列前添加一个 [价值] 词元，以聚合　　e

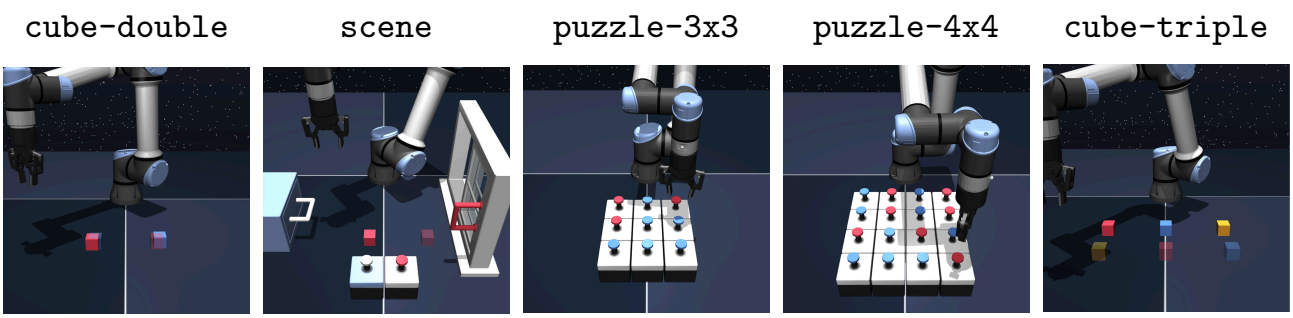

图 *14.* **可视化展示了 我们评估的领域。** OGBench 的 5 个领域各包含 5 项任务，共计 25 项任务 .

全局状态-动作信息。生成的序列由 L 个采用前置层归一化架构的 Transformer 块处理。每个块由多头自注意力和一个包含 GELU 激活的两层前馈网络组成，并在注意力子层和前馈子层之后均应用残差连接和随机失活。遵循常见做法，我们固定头维度，选择 $d_h = 32$，使头数 h 由$h = d_{\text{model}}/32$决定。TQL 的熵计算可以在对各头的注意力分数求平均后进行，也可以先对各头的熵求平均。本文的实验中，我们先对该层各头的注意力分数求平均，再计算熵。经过最后一次层归一化后，将对应于 [价值]词元的表示用作输入的全局摘要。最后，我们使用应用于 [价值] 嵌入的 K 轻量级多层感知机头，预测一个由 K 个 Q 值估计组成的集成，生成 $Q(s,\ a) \in \mathbb{R}^K$。

在扩展实验中，我们将 Transformer 层数设为 $L = 2$，并通过将隐藏维度设为 $\{128, 256, 512, 1024\}$ 来改变网络规模。除非另有说明，我们报告隐藏维度为 1024 的最大模型的最终性能。

C.3. 优化超参数

本节将补充介绍 TQL 的优化与训练细节。我们在实验中使用的所有超参数汇总于表 2。遵循 Park 等人(2025b)，我们使用环境特定的行为克隆系数 α，以适应数据集质量和任务难度的差异。对于学习率调度，我们对行为者和评论家均采用余弦衰减。行为者学习率固定为 5e-4，而评论家学习率设为 1e-4；我们发现，除单独调优的最大模型设置外，这一设置在不同维度下的整体表现最佳。

表 *2.* TQL 的默认训练超参数。

超参数	**价值**
优化器	AdamW（权重衰减 0.01）
训练步数	1 M
批量大小	256
折扣因子（γ）	0.99
目标平滑系数（τ）	0.005
演员学习率	5e-4，采用余弦调度
评论家学习率	1e-4，采用余弦调度
行为克隆系数 (α)	环境特定，参见表3
目标注意力熵 ($\bar{H}$)	环境特定，参见表3
演员流步骤	10
动作样本数量	16

注意力熵目标。 注意力熵目标 H 是 TQL 中一个重要的超参数。注意力熵的尺度主要由任务的输入维度决定，具体来说，是由状态维度 n_s 和动作维度 n_a的数量决定。在我们的实验中，我们发现，最优目标熵取决于所评估的环境，而基本不受模型规模影响；不过，为了获得最佳性能，仍应针对每个问题和设置调整熵目标。

¯为了针对每个环境进行初始化 H，我们首先根据输入维度计算注意力熵的一个粗略上界。具体而言，最大熵为

$$H_{\max} = \ln\left(1 + n_{\mathrm{s}} + n_{\mathrm{a}}\right), \tag{9}$$

并将初始目标熵设为 $0.8 \times H_{\max}$。以此值为起点，我们在 a± 0.5 范围内进行局部超参数搜索。

我们还发现，为输出层设置较低的目标熵，可以使训练更加稳定，并促使预测更具确定性。在我们的实验中，我们为输出层设定一个固定的 -0.5 较低目标熵。所有熵调优实验均使用次大模型配置进行，随后将选定的目标熵值固定下来，并在同一环境中的所有模型规模下重复使用。各环境所用的最终目标注意力熵值见表3。

C.4. 基线实现细节

我们总结了实验中使用的所有基线方法的实现细节。对于 BC、IQL、ReBRAC、FBRAC 和 IFQL，我们直接报告先前工作（Park 等人，2025b；Dong 等人，2025d）中的结果，这些结果是 8 个随机种子的平均值。我们根据 floq 提供的 Weights and Biases 日志报告其结果，这些结果基于 3 个随机种子。所有方法的领域特定超参数汇总于表 3。我们报告了所有基于 Transformer 的方法在 ∼26M 参数网络规模设置下的结果，以确保公平比较。下面，我们将介绍其余基线方法的更多实现细节。

FQL（Park 等人，2025b）。对于 FQL，主要比较中报告原论文的结果。在扩展实验中，我们将评论家的深度固定为四层，同时将隐藏维度设为 {320, 512, 1536, 2944}，以研究模型容量的影响。所有超参数设置均与官方实现保持一致。

floq（Agrawalla等人，2025年）。对于 floq，我们报告其发布的 Weights and Biases 日志中提供的结果在 1M 训练步数时的性能，以确保训练步数与其他基线及我们的方法保持一致。在扩展实验中，我们使用四层训练 floq 评论家，并将隐藏维度设为 $\{320, 512, 1536, 2944\}$。所有超参数设置均与官方实现保持一致。

Q-Transformer（Q-T）（Chebotar 等人，2023年）。对于 Q-Transformer，我们采用Brohanetal.，2023）中描述的因果Transformer解码器架构。我们实现了 QK 归一化（Dehghani 等人，2023），以减少注意力坍塌；由于序列长度固定，我们使用可学习位置嵌入。除因果掩码、QK 归一化以及输入和输出投影外，该架构与我们的基线架构基本一致，因此可以直接比较。我们对动作分箱、学习率、AdamW 权重衰减和保守性权重进行搜索，发现最优的 $N = 64$，$lr = 2e-5$，$\lambda = 0.02$，以及 $\alpha = 0.5$（定义见Chebotar 等人（2023年）中的公式 2）。我们发现，要取得合理的性能，需要采用相对较低的学习率和较高的权重衰减；否则会出现明显的不稳定性。此外，不使用 QK 归一化时，我们观察到明显的熵坍塌。

感知器演员-评论家（PAC）Springenberg 等人，2024年）。对于 PAC，我们遵循原论文指定的评论家架构，采用32 个潜在令牌，并将价值空间离散化为若干分箱。状态和动作表示均采用稠密编码器。对于价值离散化，我们将下界设为可能的最小回报，并将每个任务的上界设为 0 。价值分箱的数量设置为使每个分箱的宽度为 1，这对应于所有评估任务中的最小奖励差异。尽管 PAC 最初旨在联合学习离散动作预测，但我们发现，在当前评估设置中，这一组件会导致训练不稳定，且性能较差。为确保公平比较，我们禁用离散动作预测，并采用与 TQL 相同的策略提取机制（Park etal.，2025b），以专注于价值学习。对于扩展实验，我们训练具有两个隐藏层的 PAC 模型，并将隐藏维度设为不同的值，范围为 $\{80, 160, 320, 640\}$。

表*3.* **TQL 和基线方法的领域特定超参数。对于注意力熵目标** $\bar{H}$，第一个维度对应 Transformer 层。在每个元组中，第一个值指定 值令牌的目标熵，第二个值则适用于所有其他词元。

	IQL	ReBRAC		FBRAC	IFQL	FQL	floq	PAC		TQL
领域或任务	α	α_1	α_2	α	N	α	α	α	α	$\bar{H}$
立方体双人游戏	0.3	0.1	0	100	32	300	300	300	300	((3.0,3.0), (2.5,2.5))
cube-triple-play	10	0.03	0	100	32	300	300	300	300	((3.5, 3.5),(3.0, 3.0))
3×3拼图	10	0.3	0.001	100	32	1000	1000	1000	1000	((3.5,3.5), (3.0,3.0))
4×4拼图	3	0.3	0.01	300	32	1000	1000	1000	1000	((3.5,3.5), (3.0,3.0))
场景播放	10	0.1	0.001	100	32	300	300	300	300	((3.5,3.5), (3.0,3.0))