arXiv:2602.06476v1 [cs.MA] 6 Feb 2026

Prism：面向多智能体强化学习的谱参数共享

Kyungbeom Kim[1] , **Seungwon Oh**[1] 与**Kyung-Joong Kim**[11]光州科学技术院（GIST）{kyungbeom8, sw980907}@gm.gist.ac.kr,kjkim@gist.ac.kr

摘要

参数共享是多智能体强化学习（MARL）中提升可扩展性的关键策略，但传统的完全共享架构往往会退化为同质化行为。近期方法通过聚类、剪枝或掩码引入多样性，但通常会牺牲资源效率。我们提出**Prism**，一种参数共享框架，它通过奇异值分解（SVD）在谱域中表示共享网络，从而诱导智能体间多样性。所有智能体共享奇异向量方向，同时在奇异值上学习不同的谱掩码。该机制促进智能体间多样性并保持可扩展性。在同质（LBF、SMACv2）与异质（MaMuJoCo）基准上的大量实验表明，**Prism**在具备更优资源效率的同时取得了有竞争力的性能。

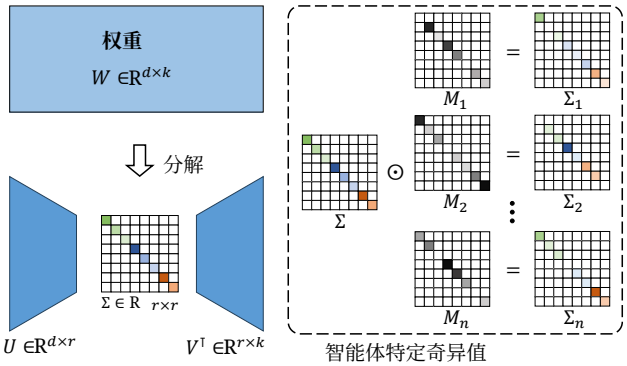


图1：**Prism框架概述**。我们将共享权重 $W \in \mathbb{R}^{d\times k}$ 分解为 $W = U\Sigma V^\top$，并在 SVD参数化空间中对其进行学习。左/右奇异向量 U 和 V由智能体共享，而智能体 i 则对奇异值应用可学习的谱掩码，从而获得智能体特定谱Σ_i，在保留参数共享的同时实现多样性。

1 引言

多智能体强化学习（MARL）已成为解决复杂现实问题的强大范式，例如交通信号控制 [Kolat 等， 2023],机器人学 [Tang 等， 2025], 以及自动驾驶 [Can- dela等， 2022]。这些应用通常需要数十到数百个智能体学习协调行为，因此可扩展性成为一项核心挑战。集中训练与分散执行（CTDE） [Lowe 等，2017] 通过在训练期间利用全局信息，缓解了非平稳性和信用分配问题，但并未从根本上解决可扩展性问题，因为模型规模、计算成本和内存使用量仍会随着智能体数量增加而迅速增长。为解决这一问题，参数共享已得到广泛采用，使所有智能体能够使用一组共享参数进行学习，从而显著降低训练和内存开销。然而，这种效率是以表达能力为代价的：单一共享策略往往无法捕捉智能体异质性，限制了其学习最优行为的能力 [Christianos 等，2020；Fu等， 2022]。尽管有人提出通过融入智能体身份来诱导策略异质性，但此类方

对于需要较高程度智能体间

异质性的任务，这些方法仍显不足，甚至可能阻碍实际学习 [Christianos 等,2021; ab Tessera 等，2025]。

为克服这些局限性，近期研究探索了部分参数共享框架，通过基于早期数据对智能体进行聚类，或在共享网络中通过节点级或边级剪枝和掩码来引入智能体特定多样性，从而缓解策略同质性过强的问题 [Christianos 等人, 2021; Li 等人, 2024a; Kim和Sung, 2023; Li 等人, 2024b]。然而，基于聚类的方法对初始数据分布高度敏感，在非平稳环境中面临重大挑战。与此同时，基于剪枝和掩码的方法需要显式维护节点和边的表示，随着智能体数量增加，资源效率会迅速下降。这些局限凸显了参数共享多智能体强化学习框架中可扩展性与策略异质性之间长期存在的权衡。

在本文中，我们提出了 **Prism**，一种新颖的参数共享方法，它在谱空间而非权重空间中运行。具体而言，我们在谱空间中采用可学习掩码，为每个智能体自适应地分配合适的子空间。这种设计使我们的方法能够处理复杂的多智能体强化学习任务，不受初始数据分布的限制，同时所需资源也比

以往的节点级和边级掩码方法更少。通过大量实验，我们证明，谱参数共享能够在不同的多智能体强化学习基准上同时实现策略异质性和内存效率。

主要贡献如下：

1. 我们提出了一种谱参数共享框架，通过在共享网络中分配不同的谱子空间来实现智能体差异化。2. 我们仅在谱空间中实现智能体差异化，从而避免冗余参数并显著提升资源效率，缓解了可扩展性与策略异质性之间固有的权衡。3. 我们通过在多种多智能体强化学习基准上进行大量实验，实证验证了所提方法的有效性。结果表明，与现有方法相比，该方法取得了有竞争力或更优的性能，同时显著降低了内存开销。

2 背景

我们将多智能体强化学习问题表述为一个分散式部分可观测马尔可夫决策过程（Dec-POMDP）[Oliehoek等人</style}, 2016]。一个 Dec-POMDP 定义为一个元组

$$(\mathcal{S}, \{\mathcal{A}_i\}_{i=1}^n, \{\mathcal{O}_i\}_{i=1}^n, P, R, \mathcal{N}, \gamma),$$

其中，$\mathcal{S}$ 为全局状态空间，$\mathcal{A}_i$ 和 $\mathcal{O}_i$ 分别表示智能体 i的动作空间和观测空间，P：$\mathcal{S} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_n \to \Delta(\mathcal{S})$ 是状态转移函数，R：$\mathcal{S} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_n \to \mathbb{R}$ 是共享奖励函数，$\mathcal{N} = \{1, \ldots, n\}$ 是智能体集合，而 $\gamma \in [0, 1)$ 是折扣因子。

在每个时间步，环境都处于状态 $s \in \mathcal{S}$。每个智能体 $i \in \mathcal{N}$ 接收局部观测 $o_i \in \mathcal{O}_i$，并根据其策略 $\pi_i(a_i \mid o_i)$ 选择一个动作 $a_i \in \mathcal{A}_i$。在完全可观测设置中（分散式部分可观测马尔可夫决策过程框架的一种特殊情况），策略也可以等价地将全局状态 s 作为输入，即 $\pi_i(a_i \mid s)$。联合动作 $\mathbf{a} = (a_1, \ldots, a_n)$ 会引发向下一状态 $s' \sim P(\cdot \mid s, \mathbf{a})$ 的转移，并获得共享奖励 $r = R(s, \mathbf{a})$。目标是学习分散式策略 $\{\pi_i\}_{i=1}^n$，使其最大化期望累积折扣奖励：

$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \mathbf{a}_t)\right],$$

其中，期望是对联合策略 $\pi = \{\pi_i\}_{i=1}^n$ 和转移动态所诱导的轨迹取的。

3 谱参数共享公式

在多智能体强化学习中，每个智能体的策略权重矩阵都可以视为共享基分量的混合。设 $\{\mathbf{B}_1, \ldots, \mathbf{B}_r\}$ 为共享基分量（尺度任意），其中 r 控制表征能力。具体而言，每个智能体都关联一个系数 $\boldsymbol{\alpha}^{(i)} \in \mathbb{R}^r$，其权重矩阵可近似表示为

as

$$\mathbf{W}^{(i)} \approx \sum_{j=1}^{r} \alpha_j^{(i)} \mathbf{B}_j, \qquad \boldsymbol{\alpha}^{(i)} = (\alpha_1^{(i)}, \ldots, \alpha_r^{(i)}).$$

系数 $\boldsymbol{\alpha}^{(i)}$ 决定每个智能体如何利用共享基分量，从而有效地在通用表征和智能体特定表征之间取得平衡。

通过正交性减少干扰与冗余。两个矩阵之间的重叠程度可通过弗罗贝尼乌斯内积 $\langle \mathbf{A}, \mathbf{B} \rangle_F := \mathrm{tr}(\mathbf{A}^\top \mathbf{B})$ 来衡量。当基分量 $\mathbf{B}_j$ 彼此相关时（即 $\langle \mathbf{B}_p, \mathbf{B}_q \rangle_F \neq 0$ 对于 $p \neq q$），它们的表示会通过交叉项 $\langle \mathbf{B}_p, \mathbf{B}_q \rangle_F$ 相互纠缠。这些交叉项导致一个分量的系数取决于其他分量，因此更新或缩放单个基元素可能会无意中改变相邻基元素。这种依赖会导致基分量之间产生干扰，并因允许多个基对重叠信息进行编码而造成冗余。施加正交性，

$$\langle \mathbf{B}_p, \mathbf{B}_q \rangle_F = 0 \quad (p \neq q),$$

消除了此类交叉项，并确保每个基分量都能独立发挥作用。在此条件下，分量 j 的系数可表示为简洁的投影形式，

$$\alpha_j^{(i)} = \frac{\langle \mathbf{W}^{(i)}, \mathbf{B}_j \rangle_F}{\|\mathbf{B}_j\|_F^2},$$

其中每个 $\alpha_j^{(i)}$ 都反映了$\mathbf{B}^j$ 的独立贡献，不受其他分量影响。因此，正交性能够防止基分量之间的干扰并避免冗余表示，从而实现更高效且更易解释的参数化。

奇异值分解作为正交基。为了构造一组正交的秩一基 $\{\mathbf{B}_j\}$，我们引入正交归一矩阵 $\mathbf{U} \in \mathbb{R}^{d \times r}$和 $\mathbf{V} \in \mathbb{R}^{k \times r}$，使其满足 $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_r$ 以及 $\mathbf{V}^\top \mathbf{V} = \mathbf{I}_r$，并定义

$$\mathbf{B}_j := \mathbf{u}_j \mathbf{v}_j^\top, \quad j = 1, \ldots, r,$$

其中$\mathbf{u}^j$ 和$\mathbf{v}^j$ 分别是$\mathbf{U}$和$\mathbf{V}$的第 j 列。每个$\mathbf{B}^j$对应一对左、右奇异向量，在各智能体之间构成相互正交的共享谱基。该基充当通用参数化框架，各智能体的权重表示为

$$\mathbf{W}^{(i)} \approx \mathbf{U}\,\mathrm{diag}\big(\boldsymbol{\sigma} \odot \boldsymbol{\alpha}^{(i)}\big)\,\mathbf{V}^\top,$$

其中，$\boldsymbol{\sigma} \in \mathbb{R}^r_{\geq 0}$ 表示共享奇异值谱，$\boldsymbol{\alpha}^{(i)}$ 表示智能体特定的缩放向量。这种分解为参数共享提供了正交坐标结构，使智能体能够通过选择性缩放不同的谱模态来实现专门化。

4 方法

在本节中，我们提出一种新颖的参数共享方法，通过由 SVD 参数化的低秩分解来建模权重。每个智能体都会对奇异值应用可学习的谱掩码，从而实现自适应且多样化的谱空间。图1展示了所提出方法的概览。

4.1 SVD分解权重

我们使用 SVD分解对共享网络的权重矩阵进行参数化，其中 $\mathbf{W} \in \mathbb{R}^{d\times k}$ 分解为

$$\mathbf{W}_\theta = \mathbf{U}\,\mathrm{diag}(\mathbf{s})\,\mathbf{V}^\top,$$

其中，$\mathbf{U} \in \mathbb{R}^{d\times r}$ 和$\mathbf{V} \in \mathbb{R}^{k\times r}$ 分别表示左、右奇异向量矩阵，$\mathbf{s} \in \mathbb{R}^r$ 表示奇异值，其中 $r = \min(d, k)$。所有分量（$\mathbf{U}, \mathbf{s}, \mathbf{V}$）均视为可学习参数，从而能够在 SVD参数化空间中进行端到端训练。这种分解为在谱空间中引入多样性奠定了基础，下一小节将通过谱空间掩码对此进行介绍。

4.2 谱空间掩码

为了在 SVD参数化的共享网络中引入智能体特定多样性，我们提出谱空间掩码。每个智能体 i 都关联一个作用于奇异值的可学习掩码$\mathbf{m}_i \in \mathbb{R}^r$ 。在这一设计中，所有智能体共享奇异向量矩阵$\mathbf{U}$和$\mathbf{V}$ ，以保持公共基；多样性则仅通过奇异值$\mathbf{s}$引入，并利用可学习谱掩码为每个智能体进行自适应区分。形式上，智能体特定的谱值定义如下

$$\mathbf{s}_i = \mathrm{concat}(\mathbf{s}_{\text{common}}, (\mathbf{s}_{\text{separate}} \odot \mathbf{m}_i)),$$

其中$\mathbf{s}_{\text{common}}$ 表示共享谱分量，$\mathbf{s}^{\text{separate}}$ 表示经掩码$\mathbf{m}_i$缩放的智能体特定部分，而 $\odot$ 表示哈达玛积（见图1）。掩码参数化为

$$\mathbf{m}_i = \mathrm{ReLU}\Big(\mathbf{s}_{\text{norm}} - \sigma\Big(\mathbf{t}_\Psi^{(i)}\Big)\Big),$$

其中$\mathbf{t}_\Psi^{(i)} \in \mathbb{R}^r$ 表示智能体 i的可学习阈值参数，$\sigma(\cdot)$ 表示 Sigmoid 激活，$\mathbf{s}_{\text{norm}} = \mathbf{s}_{\text{separate}} / \max(\mathbf{s}^{\text{separate}})$。

这种形式能够将谱空间灵活地分解为共有分量和智能体特定分量，并通过一个超参数控制二者的相对比例。因此，智能体可以自适应地分配到不同的谱子空间，从而实现智能体差异化，同时保留参数共享的优势。

4.3 多样性与正交正则化

虽然谱空间掩码能够实现智能体特定的参数化，但仍需额外的正则化，以确保智能体间具有足够的多样性并实现稳定训练。我们引入了两种互补的正则化项：多样性正则化和正交正则化。

多样性正则化。为显式鼓励智能体在谱空间中的差异，我们对其掩码之间的相似性进行惩罚。训练期间，使用连续掩码 $\mathbf{m}_i$ 以保持可微性，而二值掩码 $\tilde{\mathbf{m}}_i$ 则作为计算多样性损失时活跃谱区域的概念性近似。形式上，二值掩码定义为

$$\tilde{\mathbf{m}}_i = \mathbb{I}[\mathbf{m}_i > 0],$$

其中 $\mathbb{I}[\cdot]$表示指示函数。多样性正则化项可表示为

$$J^{\text{div}}(s) = \sum_{i=1,\dots,n} \sum_{j=1, j\neq i}^{n} \left\|\mathbf{s}_{\text{separate}} \odot (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}}_j)\right\|_1,$$

其中，$\odot$ 表示Hadamard积。由于二元掩码操作不可微，我们采用直通估计器（STE）[Bengio等人， 2013]来近似反向传播过程中的梯度。该目标引导智能体关注不同的谱分量，从而减少冗余并促进异质表征。

正交正则化。为了保持SVD分解的稳定性，我们还对奇异向量矩阵$\mathbf{U}$和$\mathbf{V}$ 进行正交正则化：

$$\mathcal{L}_{\text{ortho}} = \left\|\mathbf{U}^\top\mathbf{U} - \mathbf{I}_r\right\|_F^2 + \left\|\mathbf{V}\mathbf{V}^\top - \mathbf{I}_r\right\|_F^2,$$

其中，$\|\cdot\|_F$表示弗罗贝尼乌斯范数，$\mathbf{I}_r$ 是大小为 r的单位矩阵。该正则化可缓解奇异向量坍缩，并在SVD分解下稳定端到端训练。

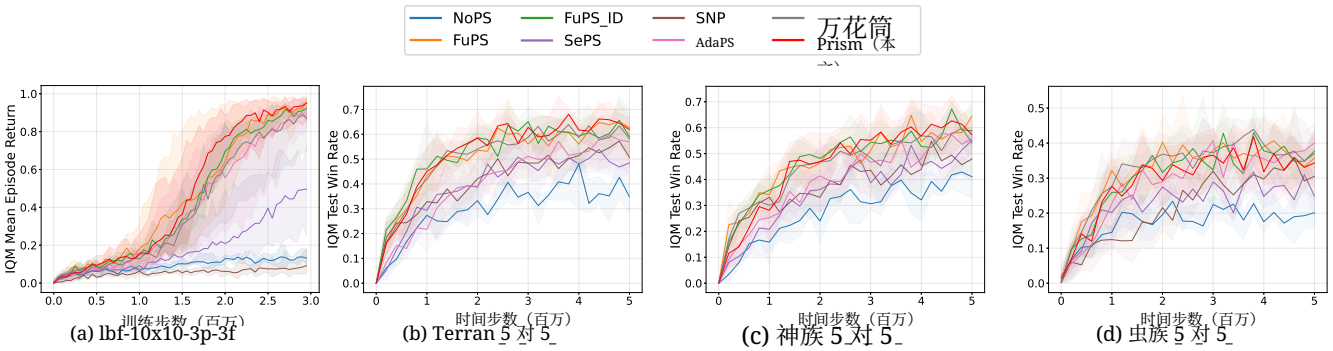
5 实验

本节通过实证评估所提出方法在多智能体强化学习中的有效性。首先，我们将所提出的方法与现有基线进行比较，以评估在谱空间中引入多样性是否能带来稳定的性能提升（第5.2节）。其次，我们从两个方面评估**Prism**的资源效率：参数预算受限时的性能稳健性，以及智能体数量增加时资源消耗的增长情况（第5.3节）。第三，我们开展消融研究，以探究**Prism** 中各个组成部分对整体性能的贡献（第5.4节）。最后，我们对谱空间中的多样性进行可视化，展示训练过程中智能体特定谱空间的形成与分化（第5.5节）。

5.1 实验设置

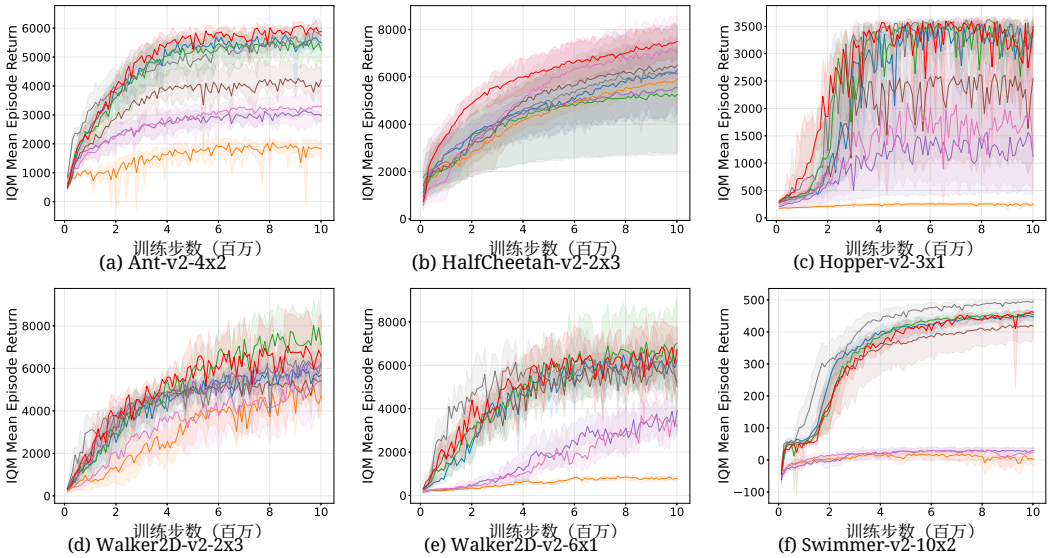
我们将提出的**Prism**与多种参数共享策略进行了比较，涵盖从不共享到自适应和结构化方法：

- **NoPS**：为每个智能体配置独立网络，作为一种强大但参数效率低下的基线。
- **FuPS/ FuPS+ID**：所有智能体共享相同的网络参数。在 FuPS+ ID 中，每个智能体的输入都会增加独热智能体标识符，以实现有限的专门化。
- **SePS** [Christianos 等，2021]：基于通过变分自编码器学习的潜在表征进行簇级参数共享，同一簇中的智能体共享相同的网络。
- **SNP** [Kim和Sung， 2023]：对一个共同的稠密网络进行节点级结构化剪枝，以形成智能体专属子网络，通过稀疏性实现多样性。



(a) lbf-10x10-3p-3f (b) Terran_5_对_5_ (c) 神族_5_对_5_ (d) 虫族_5_对_5_

图2：同质环境（LBF、SMACv2）上的性能评估。



(a) Ant-v2-4x2 (b) HalfCheetah-v2-2x3 (c) Hopper-v2-3x1

(d) Walker2D-v2-2x3 (e) Walker2D-v2-6x1 (f) Swimmer-v2-10x2

图 3：异构环境（MaMuJoCo）上的性能评估 .

• **AdaPS** [Li等，2024a]：将簇级分组与结构化剪枝相结合，在利用智能体间相似性的同时增强簇内多样性。

• **万花筒** [Li 等, 2024b]：直接在权重空间中应用可学习的边级掩码，使智能体能够端到端地发现不同的空间。

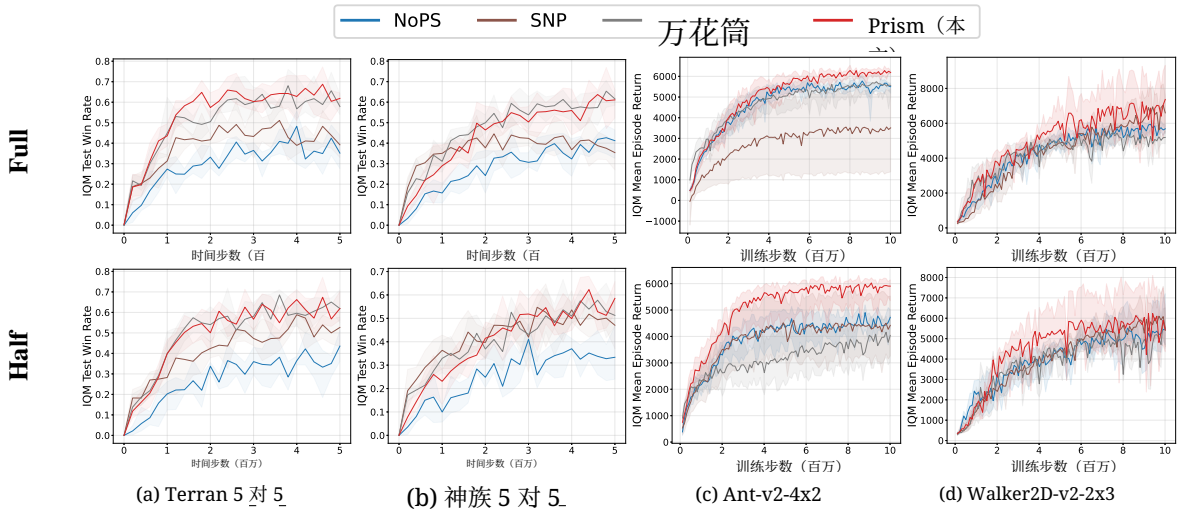
我们在同质和异构多智能体环境中评估**Prism**的性能。具体而言，我们采用三个标准基准测试：用于部分观测下离散控制的 **基于等级的觅食（LBF）** [Albrecht 和 Ramamoorthy， 2013],**SMACv2** [Ellis 等， 2023] 以及用于完全观测下连续控制的**MaMuJoCo** [Peng 等， 2021] 。我们按照 [Li等（2024b）]的网络配置，在 LBF 和 SMACv2 中采用 **QMIX** [Rashid等， 2018] 在 MaMuJoCo 中采用**MATD3** [Ackermann等， 2019]。所有结果均报告五个随机种子下平均回合回报的四分位均值（IQM），并给出 95% 自助法置信区间。

我们在 SMACv2 中使用 3 个簇和单位类型掩码，在 LBF 和 MaMuJoCo 中使用 2 个簇和智能体级掩码，二者的剪枝率均为 0.1。对于 **Prism**，我们将通用比率、正交正则化系数和多样性正则化系数分别设置为 (0.6, 0.01, 5.0)、(0.5, 0.01, 5.0) 和 (0.5, 0.01, 0.1)，对应 LBF、SMACv2 和 MaMuJoCo。

5.2 性能评估

在考察资源效率之前，我们首先评估**Prism** 在标准训练条件下的性能。这为理解在未明确考虑网络容量时，不同参数共享策略的表现提供了一个参照点。性能曲线展示于图2（LBF、SMACv2）和图 3（MaMuJoCo）。

同质化场景。 在 LBF 和 SMACv2 中，最优团队策略在很大程度上是同质的，因为智能体通常会采取相似的行为，例如共同攻击附近的目标，或集中火力攻击最近的敌人。因此，在随机且具有部分可观测动态的环境中，共享策略的学习效率至关重要。如图2所示，FuPS 的表现优于 NoPS，这得益于其在共享参数下更高的样本效率。SePS 等基于聚类的方法在此场景中的表现较差，因为它们依赖初始数据，从而限制了对随机环境的适应性。尽管 AdaPS 也是一种基于聚类的方法，但它采用共享自适应网络，提高了样本效率，进而带来逐步的性能提升。



(a) Terran 5 对 5　(b) 神族 5 对 5　(c) Ant-v2-4x2　(d) Walker2D-v2-2x3

图 4：**SMACv2 和 MaMuJoCo 在参数预算约束下的性能**。通过调整模型宽度，使用 **完整** （与 NoPS 参数量相同）和**一半** （NoPS 的 50%）预算对各方法进行评估，以确保公平比较。

相比之下，**Prism** 在不同地图上的表现不逊于其他方法，甚至更胜一筹，这得益于其采用单一共享网络，并通过可学习的谱掩码实现高效适应，以应对多样化的智能体行为。

异构设置。 与 LBF 和 SMACv2 不同，MaMuJoCo 将单个机器人拆分为多个智能体，每个智能体在相同的全局观测下控制不同的身体部位，并承担异构目标（例如，*HalfCheetah-v2-2x3*中的前腿和后腿）。因此，智能体间多样性对于实现协同控制至关重要。如图 3 所示，NoPS 的表现优于 FuPS，这凸显了在此类异构任务中，多样性比严格的参数共享更为重要。基于聚类的方法（例如 SePS、AdaPS）的表现因环境而异，这是因为不同任务的最优聚类数量有所不同——例如，AdaPS 在*HalfCheetah-v2-2x3* 上表现良好，但在其他环境中表现不佳。相比之下，**Prism**在各项任务中始终表现出色且具有竞争力，并且往往优于同样对评论家网络采用可学习掩码的万花筒。

5.3 资源效率评估

资源约束下。在大多数现实世界的多智能体系统中，计算能力或内存有限等资源约束不可避免。在这些条件下评估参数共享，可以更真实地衡量内存效率与性能之间的权衡。因此，我们在逐步缩减的参数预算下考察不同方法，分析它们利用共享表示的效率。我们在 SMACv2 的*Terran 5* 对 *5*和神族 *5* 对 *5* 场景，以及 MaMuJoCo 的 *Ant-v2-4x2* 和 *Walker2D-v2-2x3* 任务上评估我们的方法。

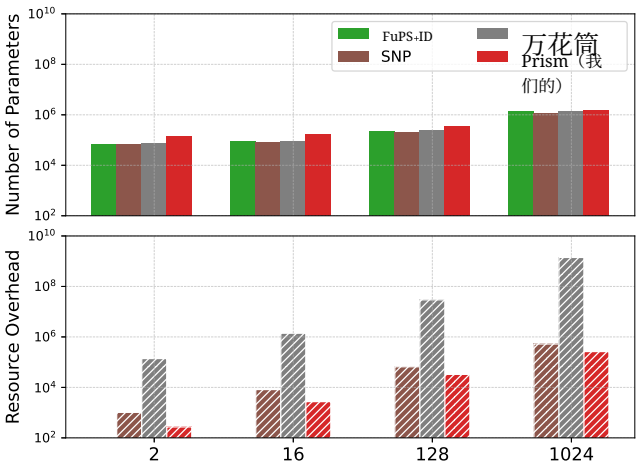
为模拟资源约束，我们将参数预算定义为相对于 NoPS 的值。我们考虑两种设置：**完整**，参数预算与 NoPS 相同；以及**一半**，预算设为 NoPS 参数数量的一半。对于每种方法，我们都会调整模型宽度以满足指定预算，确保在参数数量方面进行公平比较。

如图 4 所示，所提出的方法在不同参数预算下始终保持稳定且具有竞争力的性能。这些结果表明，谱空间掩码能够实现更有效的参数共享，并在资源受限的多智能体系统中提供明显优势。

智能体扩展下的资源效率 为进一步研究参数共享框架的可扩展性，我们使用 MaMuJoCo（多智能体游泳者）分析内存资源如何随智能体数量变化。如图5所示，由于所有方法都依赖参数共享，各方法的参数总数均仅随智能体数量略有增长。相比之下，资源开销的增长率要高得多，因为节点级或边级掩码等逐智能体组件会为每个智能体单独实例化。因此，降低资源开销对于多智能体系统的可扩展性至关重要。

万花筒会造成显著的资源开销，因为每个智能体都要维护一个与权重空间对应的完整边掩码，导致资源增长远超共享参数的增幅。SNP 为每个智能体分配节点级掩码，与万花筒相比降低了这一开销，但随着智能体数量增加，资源使用量仍会迅速上升。

相比之下，**Prism**借助更紧凑的谱空间大幅提升了资源效率，其中谱掩码取代了原始权重空间中节点级或边级的智能体专属掩码。这一设



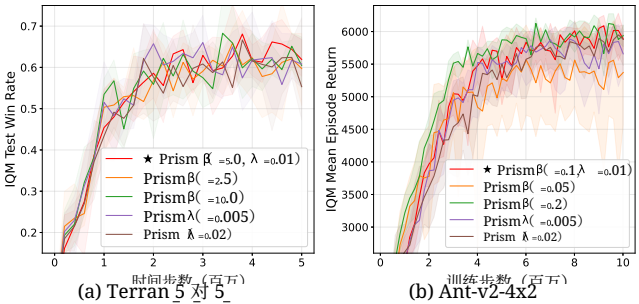
智能体	FuPS+ID	SNP	万花筒	Prism
2	0.0000	0.0145	0.6618	**0.0020**
16	0.0000	0.0892	0.9402	**0.0166**
128	0.0000	0.2488	0.9922	**0.0829**
1024	0.0000	0.3204	0.9990	**0.1483**

图5：智能体数量对资源效率的影响。**上图：** 随着智能体数量增加，各方法所需的参数总数和额外资源开销。**下图：** 归一化资源开销，定义为额外资源与模型总大小之比（资源 / (参数 + 资源)）。

计显著缓解了资源开销的增长，同时保持了可扩展性，使**Prism** 更适用于大规模多智能体系统。定量而言，Prism始终表现出低于SNP和万花筒的归一化资源开销（图5，下）。

5.4 消融研究

我们在两个代表性环境中进行消融研究，以考察各项设计选择的作用：SMACv2（*Terran 5* 对 *5*）和 MaMuJoCo（*Ant-v2-4x2*）。我们的分析重点关注所提出的正则化项的稳健性以及多样性的有效性。



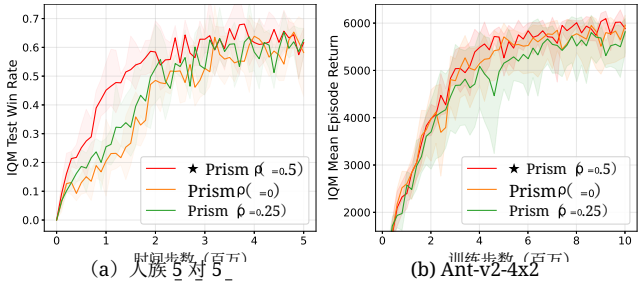
(a) Terran 5 对 5 (b) Ant-v2-4x2

图6：多样性（β）和正交性（λ）正则化强度的消融研究。

正则化鲁棒性。 我们首先考察正交正则化和多样性正则化强度对性能的影响。

具体而言，我们将系数设为第5.2节所用默认超参数值的一半和两倍，并进行实验。如图9所示，正则化强度对多样性正则化和正交正则化的影响有所不同。尤其是在 MaMuJoCo 等异构环境中，多样性正则化的影响更为显著。将多样性系数减至默认值的一半会导致性能下降，而将系数加倍则能提升最终性能。这表明，在智能体异质性发挥关键作用的环境中，增强多样性正则化可能有益。相比之下，在所有环境和系数设置下，正交正则化的性能差异都微乎其微，表明其对强度不敏感。总的来说，这些结果表明**Prism**总体上对正则化超参数具有鲁棒性，同时也说明更强的多样性正则化可能带来

在异质性至关重要的任务中带来额外收益。



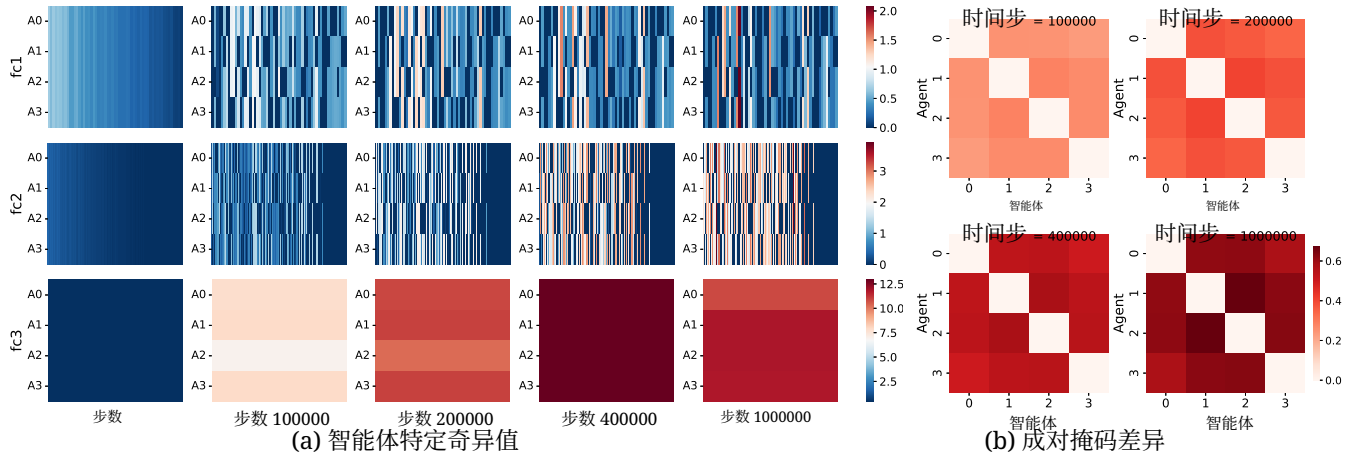
(a) 人族 5 对 5 (b) Ant-v2-4x2

图 7：公共掩码比率 ρ 的消融研究。

谱共享比率。 我们进一步分析**Prism** 对共享谱分量与智能体特定谱分量之比的敏感性。为此，我们将共享分量的比例调整为 0.0、0.25 和 0.5，并在代表性环境中评估最终性能。如图 7 所示，在人族 *5* 对 *5*中，较高比例的共享谱分量能够加快收敛。这一趋势与任务特点相符：该任务对样本效率要求较高，而更高的公共比率能够带来更稳定的梯度流，并更充分地利用共享学习信号。相比之下，*Ant-v2-4x2* 对共享比率的敏感性极低，这表明在高度依赖智能体个体特性的环境中，性能主要取决于智能体特定专业化，而非智能体间谱共享。总体而言，这些发现表明，强调样本效率的环境更能受益于较高比例的共享谱分量；而在异质协调任务中，性能提升主要来自更强的多样性正则化，而非调整共享比率。

5.5 谱空间中的多样性可视化

接下来，我们考察训练过程中多样性的形成。图 8 (a) 以热图形式展示了每个智能体在 *Ant-v2-4×2* 上掩码奇异值的变化。在训练初期，智能体关注相似的谱分量，因此多样性有限。随着学习的推进，



(a) 智能体特定奇异值　　(b) 成对掩码差异

图 8：**训练期间谱空间中的智能体表示可视化（Ant-v2-4x2）**。(a) 不同训练步数下各智能体的掩码奇异值（谱掩码）热力图。(b) 成对掩码差异矩阵，显示智能体间专门化程度随时间不断提高。

谱掩码逐渐分化，为不同智能体分配不同的奇异方向子集。这种专门化表明，智能体在利用共享表征基的同时，会自适应地占据智能体特定谱子空间。定量而言，图 8 (b) 报告了训练过程中智能体特定谱掩码之间的两两L1距离。这些掩码差异幅度的增大表明，智能体逐步为共享奇异分量分配不同的权重，从而使各智能体之间的谱利用日益分化。

6 相关工作

6.1 参数共享方法

参数共享 [Tan， 1993] 在多智能体强化学习中被广泛用于提升可扩展性，但单一的共享策略往往难以表征异质的智能体行为。为解决这一问题，先前的工作通过聚类或结构化稀疏性引入策略异质性。SePS [Christianos等， 2021] 根据潜在表征对智能体进行聚类；SNP-PS [Kim和Sung， 2023] 通过结构化剪枝引入多样性；AdaPS [Li 等， 2024a] 则结合聚类与剪枝，以实现高效的参数共享。近期，万花筒 [Li等，2024b] 采用可学习的权重空间掩码，使共享网络内能够形成智能体特定子空间。尽管取得了这些进展，现有方法仍面临内存效率与异质性之间的权衡。相比之下，在谱空间中引入多样性有望以较低的参数开销缓解这一权衡。

6.2 深度学习中的 SVD 训练

SVD 参数化通过在谱域中运作来控制神经网络层的有效秩。[Yang 等，2020] 采用 SVD 形式对权重进行参数化，并通过鼓励奇异值的稀疏性来促进低秩，从而实现奇异值剪枝。[Khodak 等， 2021] 表明，SVD 指导的初始化和正则化能够使低秩因子化层在不同架构（包括注意力）中可靠地训练。在一种互补的压缩设置中， [Wang 和 Zhang， 2024] 使用截断 SVD 将非结构化剪枝稀疏权重转换为稠密低秩分解，并通过正则化和可学习的奇异值选择来针对每个矩阵调整截断程度。相比之下，我们的工作利用 SVD 训练，并非以稀疏性或低秩压缩为主要目标，而是为了在谱空间中促成智能体差异化，以用于多智能体强化学习。

7 结论

本文提出了**Prism**，这是一种面向多智能体强化学习的新型参数共享框架，通过基于奇异值分解的分解和可学习谱掩码，在谱空间中引入多样性。通过在谱域中表示共享权重，Prism 能够有效地将参数空间分解为共享子空间和智能体特定谱子空间，在不引入冗余参数的情况下兼顾内存效率和异质性。在同质化场景（LBF、SMACv2）和异质场景（MaMuJoCo）基准测试上的大量实验表明，与最先进的参数共享方法相比，**Prism**始终能达到具有竞争力或更优的性能，尤其是在内存效率和异质性都至关重要的资源受限场景中。消融研究进一步表明，调整多样性正则化有利于异质环境，而提高公共比例在同质化场景中更为有效。

尽管**Prism** 有效应对了可扩展性挑战，并在各种环境中展现出很强的通用性，但仍存在一些局限性。首先，该方法引入了额外的超参数，用于控制共享谱分量与智能体特定谱分量之间的比率，这些超参数需要根据任务进行调整。其次，我们的实验完全在离策略设置下进行；将**Prism**扩展到同策略算法（例如 MAPPO）是一个有趣的未来方向，可能进一步提升其在动态环境中的适应性和稳定性。

训练流程

算法 1 Prism训练算法

输入： 智能体数量 N，学习率 α，经验回放缓冲区 $\mathcal{D}$
输出： 更新后的策略 θ

初始化共享SVD参数化权重（$\mathbf{U}$，$\mathbf{s}$，$\mathbf{V}$）
初始化智能体特定阈值 $\{\mathbf{t}_\Psi^{(i)}\}_{i=1}^N$
当 未收敛**时**
  从 $\mathcal{D}$ 中采样一个小批量 (o, a, r, o')
  对每个智能体 $i = 1, \ldots, N$**执行**
    计算掩码 $\mathbf{m}_i \leftarrow \mathrm{ReLU}\left(\mathbf{s}_{\text{norm}} - \sigma(\mathbf{t}_\Psi^{(i)})\right)$
    计算 $\mathbf{s}_i \leftarrow \mathrm{concat}(\mathbf{s}_{\text{common}}, \mathbf{s}_{\text{separate}} \odot \mathbf{m}_i)$
    构造策略参数 $\mathbf{W}_i \leftarrow \mathbf{U}\mathrm{diag}(\mathbf{s}_i)\,\mathbf{V}^\top$
    计算策略损失 $\mathcal{L}_i^{\text{policy}}$
  结束循环
  计算多样性正则化 $\mathcal{J}_{\text{div}}$
  计算正交性损失 $\mathcal{L}_{\text{ortho}}$
  $\theta \leftarrow \theta - \alpha \nabla_\theta \left(\sum_i \mathcal{L}_i^{\text{policy}} - \lambda_{\text{div}} \mathcal{J}_{\text{div}} + \lambda_{\text{ortho}} \mathcal{L}_{\text{ortho}} \right)$
结束循环

B 参数效率比较：权重与谱掩码

为了对权重空间掩码与谱空间掩码之间的内存使用进行闭式比较，我们分析了隐藏层权重矩阵为边长 $d \times d$ 的方阵这一常见情形。该设置反映了标准的深度多智能体强化学习架构，其中大多数参数来自方形的隐藏层到隐藏层变换。我们考虑一个由 A 个智能体使用的共享演员网络，其中 A 表示参数共享方案中的智能体数量。

采用权重空间掩码时，每个智能体都会在共享权重矩阵上维护一个逐元素掩码。参数总数变为

$$P_{\text{weight}} = d^2 + Ad^2 = (A+1)d^2.$$

采用谱掩蔽时，共享的奇异值分解参数化表示 $\mathbf{W}_\theta = \mathbf{U}\mathrm{diag}(\mathbf{s})\,\mathbf{V}^\top$ 包含 $2d^2 + d$ 个参数，而每个智能体都会引入一个谱掩码 $\mathbf{m}_i \in \mathbb{R}^d$。因此

$$P_{\text{spectral}} = 2d^2 + d + Ad.$$

因此，参数效率差距为

$$\begin{aligned} \Delta &= P_{\text{weight}} - P_{\text{spectral}} && (1)\\ &= (A+1)d^2 - (2d^2 + d + Ad) && (2)\\ &= A(d^2 - d) - (d^2 + d). && (3) \end{aligned}$$

当 $\Delta > 0$ 时，谱掩蔽的参数效率更高；这恰好在以下条件下成立

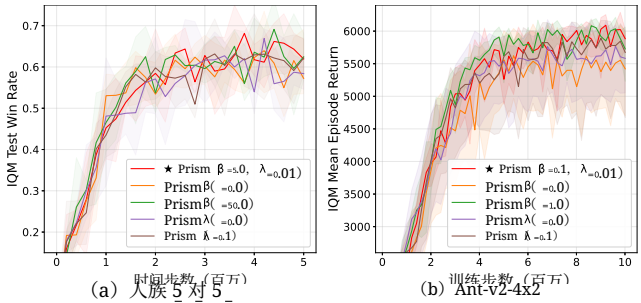
$$A(d-1) > d+1 \iff A > \frac{d+1}{d-1}.$$

对于典型维度 $d \in [32, 512]$，阈值为 $\frac{d+1}{d-1} \approx$ —— 1.01 ～ 1.06。因此，任何至少包含两个智能体（$A \geq 2$）的多智能体系统都满足该条件，这意味着在所有现实的多智能体强化学习场景中，谱掩蔽都比权重空间掩码更节省参数。

上述比较对应于最坏情况：每个智能体都可以单独控制所有 d 个奇异方向。实际上，**Prism** 仅对谱中非共享部分进行掩码，其大小为 $(1-\rho)d$。因此，提高共享比例 ρ 会缩小智能体专属掩码维度，进一步增强谱掩蔽在此处推导的满秩界之外的参数效率优势。

C 其他消融研究

C.1 正则化鲁棒性



(a) 人族 5 对 5_ (b) Ant-v2-4x2

图9：多样性（β）和正交性（λ）正则化强度的消融研究。

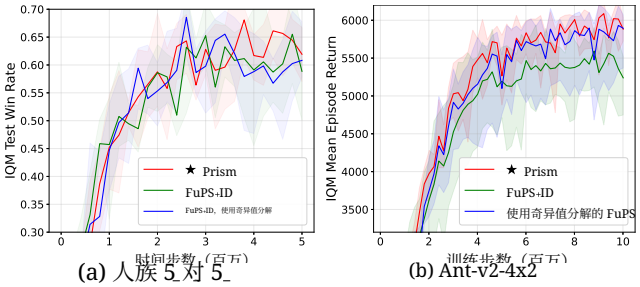
除了正文中展示的消融实验外，我们还通过将多样性和正交性系数按数量级缩放（即将其设为默认值的 0×和 10×），进一步研究 Prism 在极端正则化设置下的鲁棒性。本分析旨在评估所提出的正则化项在第 5.4 节所考察的中等范围之外的稳定性。

与主要结果一致，正则化强度的影响在 MaMuJoCo 等异质环境中尤为显著。完全移除多样性正则化（$\beta = 0$）时，性能会大幅下降，这表明，明确促进智能体间多样性对于异质控制任务中的有效协同至关重要。

同样，禁用正交正则化（$\lambda = 0$）也会导致性能明显下降，这表明，在奇异值分解参数化表示中，保持奇异向量之间的正交性对于稳定训练发挥着重要作用。

总体而言，这些结果证实，多样性正则化和正交正则化都是 Prism 的关键组成部分，移除其中任一项都会显著降低性能，尤其是在智能体异质性至关重要的环境中。

C.2 谱掩蔽



(a) 人族 5 对 5_ (b) Ant-v2-4x2

图10：谱掩蔽的消融研究。

为考察谱参数共享对性能提升的贡献，我们比较了三个变体：(i) 使用智能体标识符的标准参数共享（FuPS+ID）；(ii) 不使用多样性正则化的谱参数共享（FuPS+ID w/ 奇异值分解）；(iii) 使用可学习谱掩码和多样性正则化的谱参数共享（Prism）。此比较能够分离出在谱域中表示共享参数所产生的影响，并进一步评估通过谱掩蔽显式实现智能体差异化的益处。

这种性能差异在异构的 Ant-v2-4×2任务中尤为明显。尽管谱参数共享已经比传统参数共享带来了更好的性能，但引入具有多样性正则化的可学习谱掩码后，性能进一步提升。这一结果表明，明确分配智能体特定谱子空间，对于充分发挥谱参数共享的优势至关重要，尤其是在异构环境中。

D 超参数

常见超参数	值
动作选择器	ε-贪心
ε起始值	1.0
ε终止值	0.05
ε退火时间	5×10^4
运行器	回合
环境数量	1
回放缓冲区大小	5×10^3
批次大小	32
最大步数	3×10^6
目标网络更新间隔	200
智能体数量	3
食物数量	3
隐藏维度	64
混合嵌入维度	32
超网络嵌入维度	64
学习率	0.0005
梯度范数裁剪	10
折扣因子（γ）	0.99
TD Lambda	0.6
Q Lambda	否

表1：LBF 中使用的常见超参数。

常见超参数	值
动作选择器	ε-贪心
ε起始值	1.0
ε终止值	0.05
ε退火时间	1×10^5
运行器	并行
环境数量	4
回放缓冲区大小	5×10^3
批次大小	128
最大步数	5×10^6
目标网络更新间隔	200
友方智能体数量	5
敌方智能体数量	5
单位类型数量	3
隐藏维度	64
混合嵌入维度	32
超网络嵌入维度	64
学习率	0.001
梯度范数裁剪	10
折扣因子（γ）	0.99
TD Lambda	0.6
Q Lambda	否

表 2：SMACv2 中使用的常见超参数。

常见超参数	值
批次大小	1000
回放缓冲区大小	1×10^6
探索噪声	0.1
固定顺序	True
折扣因子（γ）	0.99
n 步回报	5
噪声裁剪	0.5
策略更新频率	2
策略噪声	0.2
Polyak 系数	0.005
评论家数量	5
激活函数	ReLU
最终激活函数	Tanh
演员学习率	5×10^{-4}
评论家学习率	0.001
隐藏维度	256
最大步数	1×10^7
训练间隔	50
预热步数	1×10^4

表 3：MaMuJoCo 中使用的常见超参数。

E 资源约束参数

本附录报告了第5.3节所述资源约束实验中使用的确切参数数量。对于每种方法，我们都会调整网络宽度，使其尽可能接近目标参数预算，从而使各方法的参数数量几乎相同。

比率	NoPS	SNP	万花筒	Prism
1.0	196,535	194,187	197,715	194,997
0.5	99,935	99,851	99,761	100,553

表4：资源受限 SMACv2（神族 *5* 对 *5*）中的参数数量。

- - -

比率	NoPS	SNP	万花筒	Prism
1.0	199,735	195,667	195,137	198,382
0.5	102,135	100,891	102,561	101,453

表5：资源受限 SMACv2（*Terran 5* 对 *5*）中的参数数量。

- - -

比率	NoPS	SNP	万花筒	Prism
1.0	379,912	380,008	377,624	380,536
0.5	190,184	191,150	188,786	190,996

表 6：资源约束 MaMu-JoCo（*Ant-v2-4x2*）中的参数数量。

比率	NoPS	SNP	万花筒	Prism
1.0	142,342	141,969	142,143	143,415
0.5	71,206	71,129	70,719	70,615

表 7：资源约束 MaMu-JoCo（*Walker2D-v2-2x3*）中的参数数量。

参考文献

[ab Tessera等, 2025] Kale ab Tessera、Arrasy Rahman、Amos Storkey 和 Stefano V. Albrecht。HyperMARL：用于多智能体强化学习的自适应超网络。发表于第二届多智能体强化学习中的协同与合作研讨会，2025年。

[Ackermann等， 2019] J. Ackermann、V Gabler、T Osa 和 M Sugiyama。使用双重集中式评论家减少多智能体领域中的高估偏差。arxiv 2019。*arXiv* 预印本 *arXiv:1910.01465*，2019。

[Albrecht 和 Ramamoorthy， 2013] Stefano V. Albrecht 和 Subramanian Ramamoorthy。多智能体系统中临时协调的博弈论模型与最佳响应学习方法。载于 *AAMAS '13 Proceedings ofthe 2013 international conference onAutonomous agentsandmulti-agent systems*，第 1155–1156 页。International Foundation for Autonomous Agents and Multiagent Systems，2013 年。

[Bengio 等， 2013] YoshuaBengio、NicholasL´eonard 和 Aaron Courville。通过随机神经元估计或传播用于条件计算的梯度。*arXiv* 预印本 *arXiv:1308.3432*，2013 年。

[Candela等, 2022] Eduardo Candela、Leandro Parada、Luis Marques、Tiberiu-Andrei Georgescu、Yiannis Demiris 和 Panagiotis Angeloudis. 将多智能体强化学习策略迁移到自主驾驶中，使用仿真到现实。在*2022 IEEE/RSJ*智能机器人与系统国际会议（*IROS*），第 8814–8820 页，2022 年。

[Christianos 等， 2020] Filippos Christianos、Lukas Sch¨afer 和 Stefano Albrecht。用于多智能体强化学习的共享经验演员-评论家。载于 H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin 编，神经信息处理系统进展，第 33 卷，第 10707–10717 页。Curran Associates, Inc.，2020 年。

[Christianos 等, 2021] Filippos Christianos、Georgios Papoudakis、Muhammad A. Rahman 和 Stefano V. Albrecht。《通过选择性参数共享扩展多智能体强化学习》。载于 Marina Meila 和 Tong Zhang（编），第*38*届国际机器学习会议论文集，第139卷，机器学习研究论文集，第1989–1998页。PMLR，2021年7月18日至24日。

[Ellis 等, 2023] Benjamin Ellis、Jonathan Cook、Skander Moalla、Mikayel Samvelyan、Mingfei Sun、Anuj Mahajan、Jakob Nicolaus Foerster 和 Shimon Whiteson。《SMACv2：改进的协作式多智能体强化学习基准》。载于神经信息处理系统数据集与基准专题第三十七届会议，2023年。

[Fu 等，2022] Wei Fu、Chao Yu、Zelai Xu、Jiaqi Yang 和 Yi Wu。《重新审视协作式多智能体强化学习中的一些常见做法》。载于 Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesvari、Gang Niu 和 Sivan Sabato（编），第*39*届国际机器学习会议论文集，第162卷，机器学习研究论文集，第6863–6877页。PMLR，2022年7月17日至23日。

[Khodak 等, 2021] Mikhail Khodak、Neil A. Tenenholtz、Lester Mackey 和 Nicolo Fusi。因子化神经网络层的初始化与正则化。发表于国际学习表征会议，2021年。

[Kim和Sung， 2023] WooJun Kim 和 Youngchul Sung。通过网络剪枝进行参数共享，以实现可扩展的多智能体深度强化学习。发表于第*22*届国际自主智能体与多智能体系统会议（*AAMAS*）。AAMAS，2023年。

[Kolat 等, 2023] M´at´e Kolat、B´alint K˝ov´ari、Tam´as B´ecsi 和 Szil´ard Aradi。用于交通信号控制的多智能体强化学习：一种协作方法。可持续发展，15(4):3479，2023年。

[Li 等，2024a] Dapeng Li、Na Lou、Bin Zhang、Zhi wei Xu 和 Guoliang Fan。面向多智能体强化学习的自适应参数共享。载于*ICASSP 2024- 2024 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第 6035–6039 页。IEEE，2024。

[Li 等，2024b] Xinran Li、Ling Pan 和 Jun Zhang。《万花筒：面向异构多智能体强化学习的可学习掩码》。收录于 A. Globerson、L. Mackey、D. Belgrave、A. Fan、U. Paquet、J. Tomczak 和 C. Zhang,

编者，《神经信息处理系统进展》，第37卷，第22081–22106页。Curran Associates, Inc.，2024年。[Lowe等，2017] Ryan Lowe、Yi I Wu、Aviv Tamar、Jean Harb、OpenAI、Pieter Abbeel 和 Igor Mordatch。《混合合作—竞争环境中的多智能体演员—评论家》。神经信息处理系统进展，第30卷，2017年。[Oliehoek等，2016] Frans A Oliehoek、Christopher Amato 等。《去中心化部分可观测马尔可夫决策过程简明导论》，第1卷。施普林格，2016年。[Peng等，2021] Bei Peng、Tabish Rashid、Christian Schroeder de Witt、Pierre-Alexandre Kamienny、Philip Torr、Wendelin Boehmer 和 Shimon Whiteson。FACMAC：因子化多智能体集中式策略梯度。收录于 A. Beygelzimer、Y. Dauphin、P. Liang 和 J. Wortman Vaughan 编，《神经信息处理系统进展》，2021年。[Rashid等，2018] Tabish Rashid、Mikayel Samvelyan、Christian Schroeder、Gregory Farquhar、Jakob Foerster 和 Shimon Whiteson。QMIX：面向深度多智能体强化学习的单调价值函数分解。收录于 Jennifer Dy 和 Andreas Krause 编，《第*35*届国际机器学习会议论文集》，机器学习研究论文集第80卷，第4295–4304页。PMLR，2018年7月10日至15日。[Tan，1993] Ming Tan。《多智能体强化学习：独立智能体与合作智能体》。收录于第十届国际机器学习会议论文集，第330–337页，1993年。[Tang等，2025] Chen Tang、Ben Abbatematteo、Jiaheng Hu、Rohan Chandra、Roberto Martín-Martín 和 Peter Stone。《面向机器人学的深度强化学习：现实世界中的成功应用综述》。*AAAI*人工智能会议论文集，39(27)：28694–28698，2025年4月。[Wang 和 Zhang，2024] Haoyu Wang 和 Wei-Qiang Zhang。《自监督预训练语音模型的非结构化剪枝与低秩分解》。*IEEE*信号处理精选主题期刊，2024年。[Yang等，2020] Huanrui Yang、Minxue Tang、Wei Wen、Feng Yan、Daniel Hu、Ang Li、Hai Li 和 Yiran Chen。《通过奇异向量正交性正则化和奇异值稀疏化学习低秩深度神经网络》。收录于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）研讨会论文集，2020年6月。