arXiv:2302.09605v1 [cs.LG] 19 Feb 2023

通过自监督信息聚合实现在线与离线多智能体强化学习中的高效通信

Cong Guan、Feng Chen、Lei Yuan、Zongzhang Zhang 和 Yang Yu*

新型软件技术国家重点实验室，南京大学邮箱:

{guanc,chenf,yuanl}@lamda.nju.edu.cn,{zzzhang,yuy}@nju.edu.cn

摘要

利用队友发送的消息能够提升合作式多智能体强化学习（MARL）中的协作能力。以往工作通常将队友的原始消息与局部信息结合，作为策略的输入。然而，忽视消息聚合会显著降低策略学习的效率。受表征学习近期进展的启发，我们认为，高效的消息聚合对于合作式多智能体强化学习中的良好协作至关重要。本文提出通过自监督信息聚合实现多智能体通信（MASIA），使智能体能够将接收到的消息聚合为紧凑且高度相关的表征，以增强局部策略。具体而言，我们设计了一种置换不变消息编码器，从消息中生成通用的信息聚合表征，并通过以自监督方式重构和预测未来信息对其进行优化。因此，每个智能体都可以通过一种新颖的消息提取机制，利用聚合表征中最相关的部分进行决策。此外，考虑到离线学习在现实应用中的潜力，我们构建了多智能体通信离线基准，据我们所知，这是首个此类基准。实验结果表明，我们的方法在在线和离线设置下均表现出优越性。本文还公开了所构建的离线基准，将其作为通信能力验证的测试平台，以促进未来的进一步研究。

1 引言

多智能体强化学习（MARL） [24] 近年来受到广泛关注 [24, 10] ，并在许多复杂领域取得了显著成功 [2], ，例如交通信号控制 [11], 、液滴控制 [36], 、主动电压控制 [65], 和动态算法配置 [72]。为在更多应用中实现更好的协作，非平稳性 [47], 和可扩展性 [4] 等问题仍有待解决。为解决多个策略并发学习所导致的非平稳性，以及智能体数量增加带来的可扩展性问题，近期大多数多智能体强化学习研究采用集中式训练与分散式执行（CTDE） [28, 39] 范式，其中包括基于价值的方法 [57, 52, 64, 76] 、策略梯度方法 [16, 38, 68, 75], 以及 Transformer 等其他技术 [69]。然而，在 CTDE 范式下，由于多智能体环境中的部分可观测性，学习到的策略的协作能力可能较为脆弱；这是许多多智能体任务面临的常见挑战 [41]。原则上，循环神经网络可以通过基于动作-观测历史来确定策略，从而缓解这一问题[23], ，但执行时对其他智能体（例如其状态和动作）的不确定性可能导致灾难性协调失误，甚至造成次优结果 [67, 8]。

通信在解决这些问题方面展现出巨大潜力， [78, 84], 智能体可以通过通信共享观测、意图或经验等信息，从而稳定学习过程，进而更好地理解环境（或其他智能体），并实现更好的协作。此前

*通讯作者

通信方法要么侧重于为消息发送方生成有意义的信息 [67, 27, 82] ，要么设计注意力机制等技术 [6, 44], 消息门控 [40, 8] ，以筛选原始接收消息中最相关的信息。这些方法将接收到的信息视为黑箱，并默认策略网络能够在策略学习过程中自动从多条原始消息中提取最关键的信息。在这种情况下，由于强化学习是唯一提供的信号，这一提取过程可能效率较低，尤其是在复杂场景中。

受近期状态表征学习进展的启发 [54, 33], ，这些进展表明辅助表征目标能够促进策略学习 [14], ，我们旨在通过设计一种新颖的通信框架，高效地聚合信息并将其作为策略的紧凑表征：**Multi-Agent communication via Self-supervised Information Aggregation**（MASIA，多智能体自监督信息聚合通信）。具体而言，我们通过自监督目标优化表征，使其既能抽象真实状态，又能预测未来信息。由于策略学习旨在引导智能体获得更高的累积奖励，使表征与真实状态及未来信息相关联，有助于强化策略学习中的学习信号，从而提升策略学习效率。此外，考虑到表征的置换不变性也能提高效率，我们设计了一种自注意力机制，以保持所得表征的不变性。我们还设计了一个网络，为单个智能体的聚合表征赋予权重，从而得到独特且高度相关的表征，用以增强局部策略并实现高效协作。

另一方面，强化学习（RL）在现实场景中的应用面临重大挑战，因为与环境交互通常代价高昂，甚至不可能 [35]。近期提出的离线强化学习旨在解决这一问题，它只从固定的离线数据集中学习智能体策略，而无需与环境交互。尽管这一主题意义重大且备受关注，当前研究主要聚焦于单智能体设置 [35, 49, 18], ，或不涉及通信的多智能体协作 [74, 46, 58, 5, 79]。尽管通信在多智能体协作中发挥着关键作用，尤其是在部分可观测场景中 [84],，目前既没有针对这一问题的有效方法，也没有用于验证此设置的测试平台。基于此，我们构建了一个离线测试平台，涵盖若干通信确有必要的热门场景，以测试不同通信方法的通信能力。我们希望这些基准测试能够用于评估各种多智能体通信算法的性能，并推动更多有关多智能体通信的离线强化学习研究。

为评估我们的方法，我们在多个合作式多智能体基准环境上进行了广泛实验，包括走廊 [67], 基于等级的觅食 [48], 交通路口 [6], 以及星际争霸多智能体挑战（SMAC）中的两张地图 [67] ，以验证该方法在在线和离线设置下的通信效果。在线实验结果表明，MASIA 优于先前方法、强基线方法以及我们方法的消融版本，证明了 MASIA 在在线学习中的有效性。我们还基于上述四种环境构建了离线数据集。有关离线学习的进一步实验表明，MASIA 在离线设置下也表现良好。

我们的主要贡献如下：

- 我们提出了一种新颖框架，利用消息聚合网络从队友生成的多条消息中提取信息，从而获得置换不变的信息聚合表示。随后，智能体可以利用一种新颖的聚焦网络 提取与决策最相关的信息。

- 我们引入了两个表征目标，使信息表示紧凑 且充分，包括状态重构和多步未来状态预测。

- 我们构建了一个用于多智能体通信的离线数据集，涵盖多种环境和不同的数据集设置。该数据集旨在为离线学习下的不同通信算法提供基准，并推动更多关于离线多智能体通信学习的研究。

- 在各种基准测试和通信条件下取得的充分在线实验结果表明，我们提出的方法显著提升了通信性能，而可视化

结果进一步揭示了其有效的原因。在离线数据集上的补充实验结果证明了我们的方法在离线设置中的有效性，并为该领域的进一步研究提供了启发。

2 问题形式化

本文研究一个完全合作的多智能体强化学习通信问题，该问题可建模为通信条件下的分散式部分可观测马尔可夫决策过程（Dec-POMDP-Com）[45]，并形式化为元组 $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, \mathrm{P}, \Omega, \mathrm{O}, \mathrm{R}, \gamma, \mathcal{M} \rangle$，其中 $\mathcal{N} = \{1, \ldots, \mathrm{n}\}$ 是智能体集合，$\mathcal{S}$ 是全局状态集合，$\mathcal{A}$ 是动作集合，Ω 是观测集合，O 为观测函数，R 表示奖励函数，$\gamma \in [0, 1)$ 表示折扣因子，而 $\mathcal{M}$ 表示消息集合。在每个时间步，由于部分可观测性，每个智能体 $\mathrm{i} \in \mathcal{N}$ 只能获取由观测函数 O(s, i) 生成的观测 $\mathrm{o}_i \in \Omega$，其中 $\mathrm{s} \in \mathcal{S}$；每个智能体持有个体策略 $\pi(\mathrm{a}_i \mid \tau_i, \mathrm{m}_i)$，其中$\tau_i$ 表示智能体 i 在当前时间步 t 的历史记录 $(\mathrm{o}_i^1, \mathrm{a}_i^1, \ldots, \mathrm{o}^{t-1}\mathrm{i}, \mathrm{a}^{t-1}\mathrm{i}, \mathrm{o}_i^t)$，而 $\mathrm{m}_i \in \mathcal{M}$ 是智能体 i 接收到的消息。联合动作 $\mathrm{a} = \langle \mathrm{a}_1, \ldots, \mathrm{a}_n \rangle$ 会导向下一状态 $\mathrm{s}' \sim \mathrm{P}(\mathrm{s}' \mid \mathrm{s}, \mathrm{a})$，并产生全局奖励 R(s, a)。形式化目标是寻找联合策略 $\boldsymbol{\pi}(\boldsymbol{\tau}, \mathrm{a})$，以最大化全局价值函数 $\mathrm{Q}_{\mathrm{tot}}^{\pi}(\boldsymbol{\tau}, \mathrm{a}) = \mathbb{E}_{s,\boldsymbol{a}}\left[\sum_{t=0}^{\infty} \gamma^t \mathrm{R}(\mathrm{s}, \mathrm{a}) \mid \mathrm{s}_0 = \mathrm{s}, \mathrm{a}_0 = \mathrm{a}, \boldsymbol{\pi}\right]$，其中 $\boldsymbol{\tau} = \langle \tau_1, \ldots, \tau_n \rangle$。由于每个智能体既可以充当消息发送方，也可以充当消息接收方，本文研究如何在接收端学习有用的消息表示；智能体仅使用局部信息 o_i 作为消息在团队内共享。

我们通过基于价值的多智能体强化学习优化策略，其中深度 Q 学习 [43]使用深度神经网络实现动作价值函数 $Q(s, \boldsymbol{a})$，即由 $\boldsymbol{\theta}$参数化的 $Q(\boldsymbol{\tau}, \boldsymbol{a}; \boldsymbol{\theta})$。本文采用集中训练、分散执行范式。在集中式训练阶段，深度 Q 学习使用回放记忆库 $\mathcal{D}$ 存储转移元组 $\langle \boldsymbol{\tau}, \boldsymbol{a}, r, \boldsymbol{\tau}' \rangle$。我们使用 $Q(\boldsymbol{\tau}, \boldsymbol{a}; \boldsymbol{\theta})$ 来近似 $Q(s, \boldsymbol{a}; \boldsymbol{\theta})$，以缓解部分可观测性。因此，通过最小化期望时序差分（TD）误差来学习参数 $\boldsymbol{\theta}$:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{\tau}, \boldsymbol{a}, r, \boldsymbol{\tau}') \in \mathcal{D}}\left[\left(r + \gamma V\left(\boldsymbol{\tau}'; \boldsymbol{\theta}^{-}\right) - Q(\boldsymbol{\tau}, \boldsymbol{a}; \boldsymbol{\theta})\right)^2\right],$$

其中，$V(\tau'; \boldsymbol{\theta}^{-}) = \max_{a'} Q(\boldsymbol{\tau}', \boldsymbol{a}'; \boldsymbol{\theta}^{-})$ 是时序差分目标的预期未来回报，$\boldsymbol{\theta}^{-}$ 是目标网络的参数，并通过 $\boldsymbol{\theta}$定期更新。

在离线设置下，我们有一个表示为 $\mathcal{B}$的离线数据集。该数据集由特定行为策略收集，并在整个训练过程中保持固定。对于 Dec-POMDP 问题，离线数据集 $\mathcal{B}$通常由多条轨迹构成，即 $\mathcal{B} := \{(s^t, \{o_i^t\}_{\mathcal{N}}, \{a_i^t\}_{\mathcal{N}}, r^t)_{t=1}^{T}\}$，其中 T 表示轨迹长度，r^t 表示在第 t 个时间步获得的奖励。进一步考虑多智能体通信时，我们还会记录每个智能体发送的消息的接收方 ID。这些信息描述了当前问题所允许的通信信道。

3 方法

本文提出一种通过自监督信息聚合实现的高效多智能体通信方法（MASIA），这是一种旨在提升合作性能的新型多智能体通信机制。冗余通信会增加各智能体进行决策时的信息处理负担；由于原始消息中包含大量无关信息，也会给信息提取带来新的挑战。为设计高效的通信机制，我们认为两项属性至关重要——充分性和紧凑性。充分性意味着信息丰富，而紧凑性则要求更高的信息密度。

为满足充分性标准，我们在智能体之间共享一个全局编码器，并将其称为信息聚合编码器（IAE），用来将智能体广播的信息聚合为共同表示。经过适当训练后，该表示能够反映全局观测，使每个智能体都能从中获取充分的信息以做出决策。至于紧凑性，我们首先在全局表示上设计一种辅助损失，使其与策略学习过程相关联，并让每个智能体仅

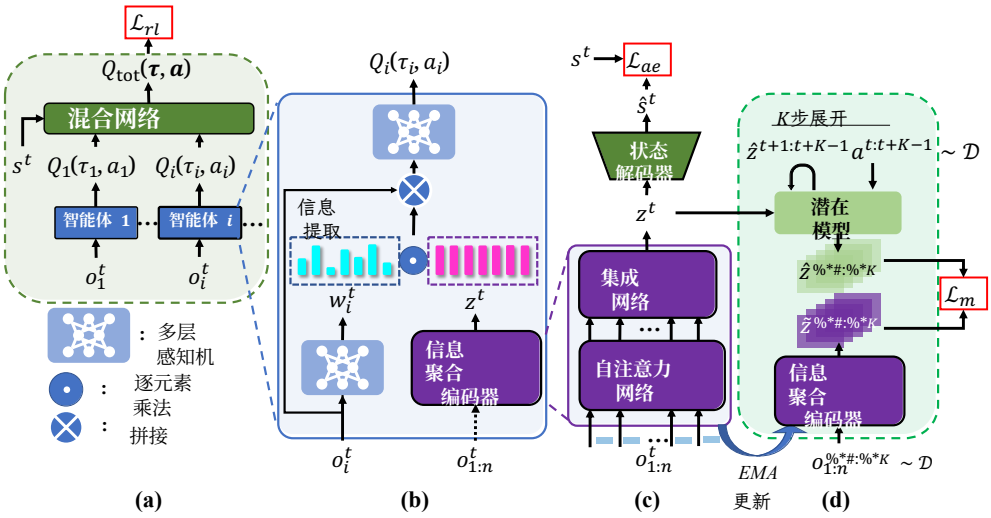


图1：MASIA 的结构。（a）整体架构。（b）信息聚合与提取。（c）信息聚合优化。（d）转移模型学习。

通过所设计的聚焦网络排除无关部分，从而只关注表示中与其性能和协作相关的部分。我们方法的完整框架如图所示1。此外，训练和执行过程的说明见附录B.2。

3.1 信息聚合与提取

信息聚合。 我们认为，真实状态应当通过组合后的消息加以反映，因此设计了聚合编码器，使其能够囊括智能体发送的所有消息。此外，多智能体系统中的通信系统具有灵活性，并且本质上具备置换不变性，这要求聚合编码器采用置换不变结构。基于这些认识，我们采用自注意力机制来聚合来自不同队友的多条消息：

$$\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} = \texttt{MLP}_{Q,K,V}([o_1^t, \ldots, o_i^t, \ldots, o_n^t]), \tag{1}$$

$$\boldsymbol{H} = \texttt{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}})\boldsymbol{V}, \tag{2}$$

其中，可学习矩阵 Q、K和 V 将所有智能体的感知转换为相应的查询 $\boldsymbol{Q}$ 、键 $\boldsymbol{K}$和值 $\boldsymbol{V}$，这些都是注意力机制中定义的概念 [61]。具体而言，$\boldsymbol{H}$ 的每个行向量都可以看作一个智能体针对所有可用信息的查询结果，而隐藏状态 $\boldsymbol{H}$ 将被输入后续的集成网络，最终得到输出的聚合表征 z^t。关于集成网络设计的详细讨论见附录 B.1。在集中式训练阶段，我们将聚合表征 z^t 作为个体观测之外的额外输入，输入价值函数。由于 z^t 包含确定真实状态所需的信息，因此将 z^t 作为额外输入，可以降低对环境状态的不确定性，并使任何基于价值的策略学习算法下的价值函数更准确地估计Q值。

信息提取。 与人类的决策过程类似，在通信系统中，全局消息通常对于单个智能体进行良好协作而言是冗余的。例如，在交通路口任务中 [6]，一个自然的想法是，相较于远处车辆的信息，邻近车辆的信息对于智能体感知而言更为重要；当地图较大时，全局消息中的无关信息有时甚至会使智能体感到困惑，并阻碍学习。一个玩具实验

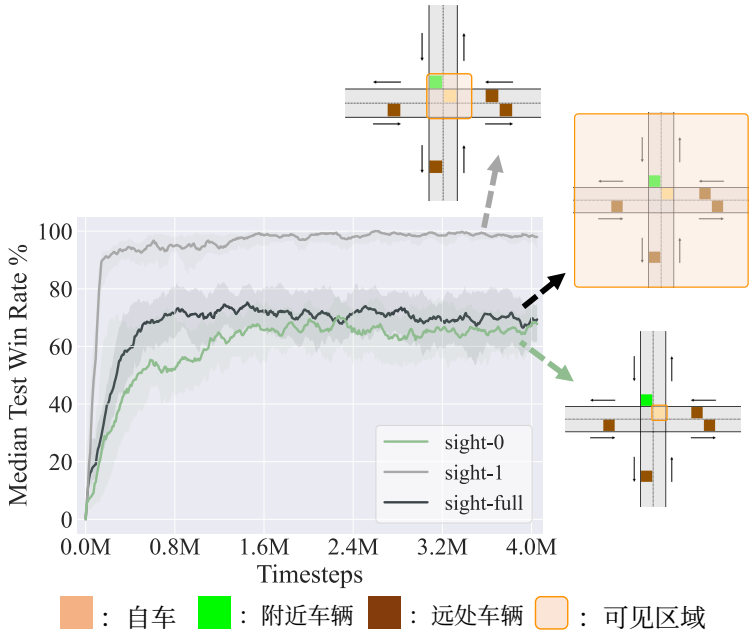


图 2：交通路口任务中用于说明信息冗余的一个简单实验。

如图 2 所示，这一想法得到了支持。我们在具有不同视野设置的交通路口任务中应用QMIX算法。结果表明，与视野极其受限的设置（sight-0）和全视野设置（sight-full）相比，智能体在小视野设置（sight-1）下能够学习到更好的策略，这促使人们寻求高效的消息提取方法。

为了使每个智能体都能够决定自身的感知区域，我们采用**聚焦网络**对每个智能体的聚合表示进行加权。聚焦网络采用多层感知器（MLP）设计，并使用 Sigmoid输出激活函数，以确保w_i^t 的每个维度都介于 0 和 1 之间。通过与z^t进行逐元素乘法，可以为单个智能体提炼出独特的表征。这样，如果聚焦网络为特定维度赋予较高权重，那么这些维度上的聚合表示变化就会更加显著，因此，智能体对这些部分的聚合表征也会更加敏感。相反，如果某些维度的权重接近于零，这些维度上的信息就会被过滤掉。尤其是，尽管信息提取过程在一定程度上能够体现每个智能体的特性，我们仍强调局部信息，并将其与提取出的表征一同输入后续网络。

3.2 信息表示优化

对于聚合表征的学习过程，我们考虑全局编码器训练中的两个典型目标：重构和多步预测，它们分别约束全局编码器生成的表征具备充分性和紧凑性。对于重构目标，我们采用额外的解码器，旨在通过聚合表征重构全局状态，从而对全局编码器进行自监督。具体而言，解码器与聚合编码器一同优化，通过多个接收到的消息 s^t 和$\boldsymbol{o}^t$来重构全局状态：

$$\mathcal{L}_{ae}(\theta,\eta)=\mathbb{E}_{\boldsymbol{o}^t,s^t}\|g_\eta(z^t)-s^t\|_2^2,\quad z^t=f_\theta(\boldsymbol{o}^t), \tag{3}$$

其中，f_θ、g^η 分别表示参数化为 θ 的编码器网络和参数化为η的解码器网络。该损失项类似于经典的自编码器损失，但这里的解码器并非用于

重构输入，而是用于从表征中恢复全局状态。通过利用该损失，我们引导编码器提取有助于推断全局状态的观测特征，并使$z^t = f_\theta(\boldsymbol{o}^t)$ 成为充分表征。

对于多步预测目标，我们约束生成的表征，使其能够预测未来信息。具体而言，我们设计了一个转移模型 h_ψ： $\mathcal{Z} \times \mathcal{A}^n \to \mathcal{Z}$，将其参数化为 ψ，作为辅助模型；该模型通过当前步 z^t 所对应的聚合表征和联合动作 $\boldsymbol{a}^t$，预测下一步 $t+1$ 的聚合表征 z^{t+1}。经过 k 步展开后，我们将预测的聚合表征与未来消息 $\boldsymbol{o}^{t+k}$的实际聚合表征进行回归，并通过多步预测损失更新聚合编码器和辅助模型：

$$
\begin{aligned}
&\mathcal{L}_m(\theta,\psi) = \mathbb{E}_{\boldsymbol{o}^t,s^t,\boldsymbol{a}^t,\ldots,\boldsymbol{a}^{t+K-1},\boldsymbol{o}^{t+K},s^{t+K}}\left[\sum_{k=1}^{K}\|\hat{z}^{t+k}-\bar{z}^{t+k}\|_2^2\right],\\
&\hat{z}^{t+1} = h_\psi(\bar{z}^t,\boldsymbol{a}^t),\\
&\hat{z}^{t+k} = h_\psi(\hat{z}^{t+k-1},\boldsymbol{a}^{t+k-1}),\quad k=2,\ldots,K,\\
&\bar{z}^{t+k} = f_\theta(\boldsymbol{o}^{t+k}),\quad k=0,\ldots,K.
\end{aligned}
\tag{4}
$$

为进一步稳定学习过程，我们在聚合编码器中采用双网络技术：使用两个架构相同但更新频率不同的网络。目标网络采用指数移动平均（EMA）进行更新，与 SPR 中的方法类似 [54]。通过使聚合表征能够预测其未来状态，聚合表征可以与决策所需的信息具有更强的相关性，从而满足紧凑性标准。结合这两个目标，聚合编码器能够提取更有用的信息，帮助智能体更好地协调。为了提升单个智能体的信息提取能力，我们还通过强化学习目标来增强这些组件的学习过程。具体而言，我们考虑最小化时序差分损失：

$$
\begin{aligned}
\mathcal{L}_{rl}(\theta,\phi) = \mathbb{E}_{(\boldsymbol{\tau},\boldsymbol{a},r,\boldsymbol{\tau}')\in\mathcal{D}}\Bigg[\Bigg(&r+\gamma\max_{\boldsymbol{a}'}Q_{\text{tot}}\left(\boldsymbol{\tau}',\boldsymbol{a}';\theta^-,\phi^-\right)\\
&-Q_{\text{tot}}(\boldsymbol{\tau},\boldsymbol{a};\theta,\phi)\Bigg)^2\Bigg],
\end{aligned}
\tag{5}
$$

其中，Q_{tot} 通过各智能体的Q值计算得出。Q 值的计算实际上取决于具体的基于价值的学习算法。我们将其应用于常见的方法，包括 VDN [57]、QMIX [52]和 QPLEX [64]。此外，聚焦网络的更新与强化学习目标相耦合，使聚焦网络生成的权重能够对任务敏感，从而也有助于策略学习。

3.3 离线学习的表征预训练

鉴于离线学习对现实应用的重要性，本文也涵盖了离线设置。将我们的方法用于离线通信学习时，我们只能访问固定的离线数据集，无法与环境交互。因此，难点在于如何充分利用有限的离线数据，高效地学习多智能体通信策略。另一方面，由于聚合编码器与 Q 网络一同更新，它所导出的表征空间 $\mathcal{Z}$ 在一定程度上是动态的。通常，编码器需要一定量的数据进行训练，才能获得相对稳定的表征空间。

受上述分析启发，我们提出在使用离线数据集优化智能体策略之前进行表征预训练。核心思路是首先获得一个相对良好的聚合编码器，使其生成的表征空间在一定程度上保持稳定。随后，我们利用离线数据集进一步更新 Q网络，并联合微调聚合编码器。整体流程如图3所示。事实上，这种做法有一些优点：（1）表征预训练过程通常优化无监督学习目标，且只更新聚合编码器网络，因此

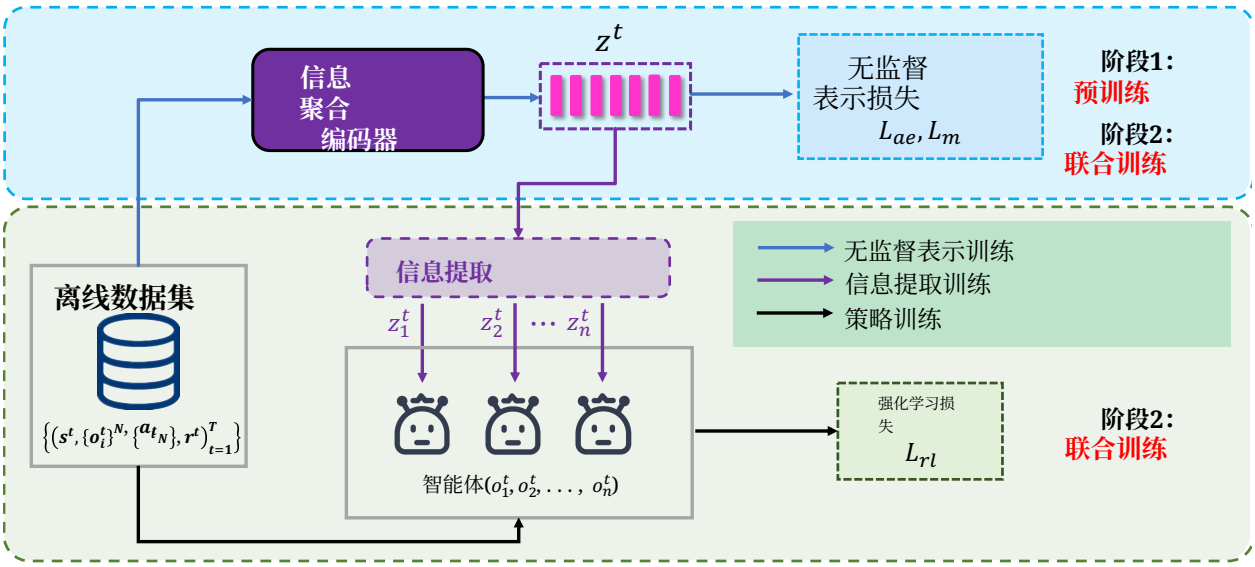


图3：MASIA用于离线学习的整体训练流程，包括两个训练阶段，分别是预训练和联合训练。在预训练阶段，我们仅利用离线数据优化无监督表示损失；而在联合训练阶段，我们同时优化表示损失和强化学习损失。此外，我们使用 agent$(o_1^t, ..., o_n^t)$ 表示智能体策略的网络推理过程，该过程用于输出动作或Q值。

无需考虑 Q 发散问题；（2）预训练聚合编码器网络能够带来更稳定的聚合表示空间，从而有助于后续强化学习目标的优化。通过这种做法，我们旨在结合本方法的属性，更好地利用离线数据集。从某种意义上说，这也可以视为我们方法机制带来的一项优势，因此我们在所有离线实验中均采用这种做法。第4.2.2 节中的进一步消融研究证明了这种做法的有效性。

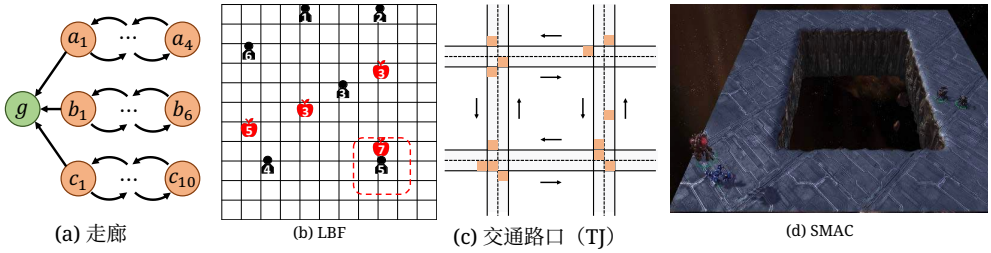
4 实验

为了评估我们的方法在在线和离线问题设置中的有效性，我们首先在四个基准测试上开展在线通信学习，以验证MASIA的有效性。此外，我们构建了一个离线数据集，以支持离线通信学习。其目的是检验各种多智能体通信学习算法在离线设置中的收敛性能，并验证MASIA是否仅凭固定的离线数据集仍能取得良好表现。

4.1 在线多智能体通信学习

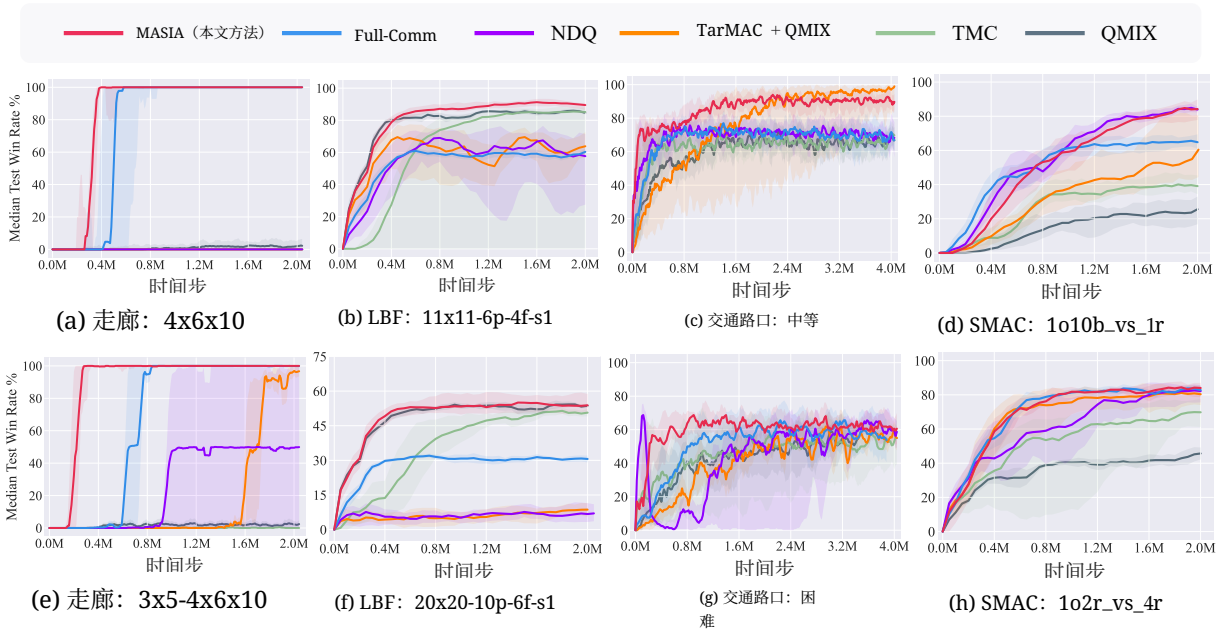
我们在具有不同通信请求级别的各种基准测试上开展实验[1]。具体而言，本节旨在回答以下问题：1）与多种基线方法相比，我们的方法在不同场景中的表现如何（第4.1.1节）？ 2）信息聚合编码器学到了什么样的知识（第4.1.2节）？ 3）信息提取模块如何从已学习的嵌入空间中提取对单个智能体最相关的信息（第4.1.3节）？ 4）MASIA能否应用于不同的价值分解基线方法，以提升其在各种通信条件下的协作能力和鲁棒性（第4.1.4节）？ 我们将MASIA与多种

[1]代码可在https://github.com/chenf-ai/MASIA



(a) 走廊 (b) LBF (c) 交通路口（TJ） (d) SMAC

图4：我们实验中使用的多个基准测试。



(a) 走廊：4x6x10 (b) LBF：11x11-6p-4f-s1 (c) 交通路口：中等 (d) SMAC：1o10b_vs_1r

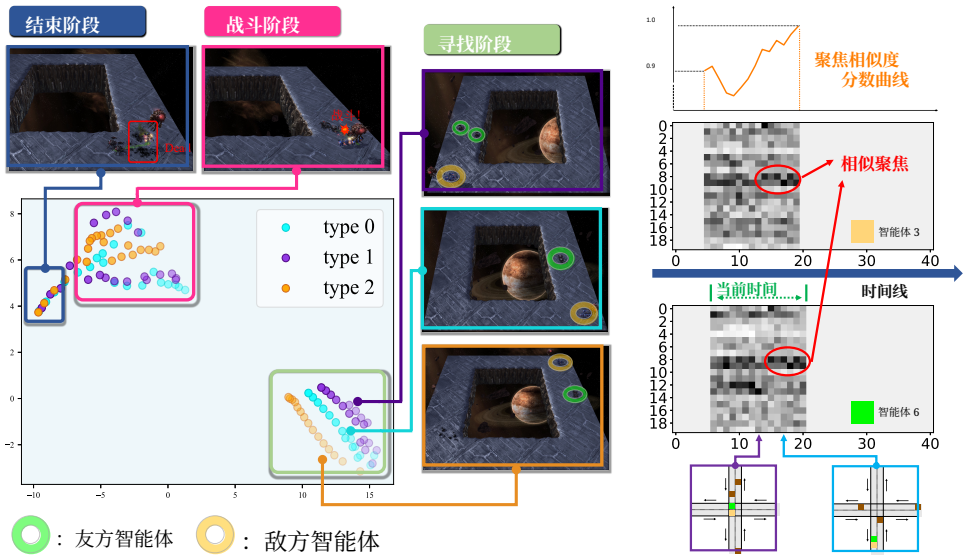
(e) 走廊：3x5-4x6x10 (f) LBF：20x20-10p-6f-s1 (g) 交通路口：困难 (h) SMAC：1o2r_vs_4r

图5：在多个基准测试上与基线方法的性能比较。

基线方法进行比较，包括无通信方法和一些最先进的通信方法。QMIX [52]是一种强大的无通信基线方法，我们采用 PyMARL 提供的实现[2] 进行比较，该实现已在多种多智能体基准测试中展现出优异性能 [53]。TarMAC 利用注意力机制，根据消息的相对重要性选择消息。我们使用的实现由 [67], 提供，记为 TarMAC + QMIX。NDQ [67] 旨在通过生成有意义的消息并最小化通信，学习近似可分解的 Q函数。TMC [82] 在消息发送方一端采用时间平滑技术，大幅减少智能体之间交换的信息量。在消融研究中，我们设计了一个仅在通信协议上有所不同的基线方法，该方法采用完全通信范式，即每个智能体在每个时间步都会接收其他所有队友发送的消息，记为 FullComm。

我们在图 4所示的多个基准测试上评估了所提出的方法。走廊 [67] 是一个具有部分可观测性的合作环境，其中 m 个智能体随机初始化在不同位置，并需要同时到达目标 g 。我们考察了两种场景，其中智能体数量各异且分属不同群组，不同群组必须在不同时间到达。基于等级的觅食（LBF） [48] 是另一种具有部分可观测性的合作网格世界游戏，智能体需要协同合作，同时收集食物。交通路口（TJ） [6] 是一种常用的基准测试，用于检验通信能力，其中

[2]我们的实验均基于使用 SC2.4.6.2.6923 的 PyMARL框架。



(a) (b)

图 6: （a）信息聚合可视化。图中的每个点都代表一个聚合表征。三种不同的颜色分别代表敌方实体的三种不同初始化情形。例如，类型 0 表示敌人初始化于右下角的情形。为区分不同时间步上的聚合表示，我们用颜色较深的点标记较大的时间步。（b）单个回合中所选智能体关注权重 w_i^t 变化的可视化。横轴表示单个回合中的时间步，纵轴表示聚合表征的维度。权重通过亮度体现，单元格颜色越深，权重越大。

许多汽车沿着双向道路行驶，遵循预定义路线经过一个或多个道路交叉口；我们在中等和困难地图上进行测试。SMAC 中的两张地图 1o2r_对_4r 和 1o10b_对_1r [67]要求智能体相互协作并通信，以获取敌人的位置。评估时，所有结果均以 5 个随机种子上的中位数性能和 95% 置信区间报告。有关基准测试、网络架构以及本方法超参数选择的详细信息分别见附录 A和 B.1。

4.1.1 通信性能

我们首先将 MASIA 与多种基线方法进行比较，以考察其在各种基准测试中的通信效率。如图 5所示，MASIA 在所有基准测试中都取得了最佳表现，且方差较低，表明 MASIA 在各种难度的场景中都具有很强的适用性。在走廊环境中（图 5a & 图 5e），良好表现需要频繁通信（QMIX 等不使用通信的方法会失败），而 TarMAC、NDQ 和 TMC 等其他通信方法在该环境中的表现不佳，甚至会失败。这表明，不恰当的消息生成或消息选择会损害学习过程。我们认为，FullComm 能够成功的原因在于走廊环境中智能体观测之间几乎没有冗余。我们的 MASIA 在该环境中也取得了成功，表现优于其他方法。在扩展走廊环境中（图 5e），MASIA 的优势表现更加显著，因为智能体被划分为不同的群组。在此环境中，MASIA 可以帮助智能体提取有关队友的信息，并更高效地学习协作模式。在 LBF 环境中（图 5b& 图 5f），现有基于通信的多智能体强化学习方法（如 NDQ、Fullcomm 和 TarMAC）因奖励稀疏而难以取得良好表现，尤其是在食物分布更加稀疏时（图 5f）。与 QMIX 在走廊环境中的表现相反，QMIX 在

LBF 中表现良好，这归因于智能体能够观测到附近的网格，而且 QMIX 的混合网络可以在训练阶段帮助提升固定群组中智能体的协作能力。我们的方法取得了与 QMIX 和 TarMAC 相当的表现，展现出其在稀疏奖励场景中的强大协作能力。在交通路口环境中（图 5c& 图5g），TarMAC 和 NDQ 因消息不稳定而方差较大，而 MASIA 的样本效率较高，并且由于其目标是重构状态，能够生成稳定的消息。在 SMAC基准测试中（图 5d & 图5h），我们在两个需要通信才能成功的复杂场景中进行测试，其中一名监督者负责获取敌人的信息。由于智能体的视野有限，因此需要传递消息，其他队友需要监督者发送的消息来识别敌人的位置。我们的 MASIA 方法能够保持较高的学习效率，并且收敛后始终具有竞争力，优于其他基线方法。

4.1.2 信息聚合编码器的洞察

为了确定编码器通过训练学习了何种知识，我们在 SMAC 的地图 1o2r_与_4r 上进行可视化分析，以展示聚合表示中所包含的信息z^t。我们通过 t-SNE 将聚合表示向量投影到二维平面上 [60]，如图6a 所示。在一项任务中，智能体必须在开始时寻找敌人、发现敌人，并最终与其战斗，以获得更好的性能；我们从 3 个具有不同初始化类型的场景中获取轨迹。可以观察到：(1) 聚合表示能够按照阶段得到良好区分。寻找阶段中的投影表示与战斗阶段和结束阶段中的表示相距较远。这表明我们学习到的表示能够很好地反映真实状态。(2) 在敌人完成初始化后的寻找阶段，聚合表示最初彼此分离，但随后逐渐交织在一起，直至结束阶段；此时，敌人在一场激烈的战斗后已被消灭。由于敌人与决策密切相关，这一结果验证了聚合表示也利用了奖励信息。总而言之，可视化结果表明，MASIA 能够通过这些表示提取有价值的全局信息。

4.1.3 单个智能体的信息提取研究

为展示信息提取模块的有效性，我们分析了聚焦网络计算出的权重。具体而言，我们选取 TJ（中等）任务进行评估，并比较两个不同智能体生成的权重。在此环境中，智能体可以动态进入或离开平面，因此环境中的智能体会随时间灵活变化。可以观察到，智能体 3 和智能体 6 将注意力集中在聚合表征中相似的区域。特别是在时间步 15 之后，当智能体 3 和智能体 6 处于相似情境且远离交叉路口时，它们关注的区域几乎相同。这验证了全局状态信息已成功提取并传递给单个智能体。此外，在图的上方，我们绘制了一幅图，用于衡量不同智能体的权重向量余弦相似度与时间步之间的关系。图中显示，当这两个智能体开始沿同一条路线前进后（如图6b 下方的渲染图像所示），它们关注点的相似度随之上升。这也符合我们的直觉：相似的观测应提取出相似的消息。

4.1.4 我们方法的普适性

我们旨在验证所提出的方法不受不同视野范围和所采用的基于价值的多智能体强化学习方法的影响。我们首先在地图 1o10b_vs_1r 上进行实验，以展示 MASIA 也能很好地泛化到观测受限的智能体。第 4.1.1 节的结果展示了智能体视野范围为 9 时 MASIA 的性能。如图 7a 所示，当我们缩小智能体的视野范围时，通过接收并聚合队友发送的消息，MASIA 的性能并未显著下降。我们的信息聚合与提取模块防止智能体在视野范围进一步受限时丧失有关状态的知识。

表 1：我们的多智能体通信离线数据集的格式。

步骤 0	···	步 t		步 $t+1$		···	步 T
s^0	···	s^t		s^{t+1}		···	s^T
r^0	···	r^t		r^{t+1}		···	r^T
智能体$^0_{i=1}$	···	智能体$^t_{i=1}$	o^t_1 a^t_1 完成$^t_1$ 接收方 ID^t_1	智能体$^{t+1}{}_{i=1}$	o^{t+1}_1 a^{t+1}_1 完成$^{t+1}_1$ 接收方 ID^{t+1}_1	···	智能体$^T_{i=1}$
···	···	···		···		···	···
智能体$^0_{i=n}$	···	智能体$^t_{i=n}$	o^t_n a^t_n 完成^t_n 接收方 $ID^t{}_n$	智能体$^{t+1}{}_{i=n}$	o^{t+1}_n a^{t+1}_n 完成$^{t+1}_n$ 接收方 ID^{t+1}_n	···	智能体$^T_{i=n}$

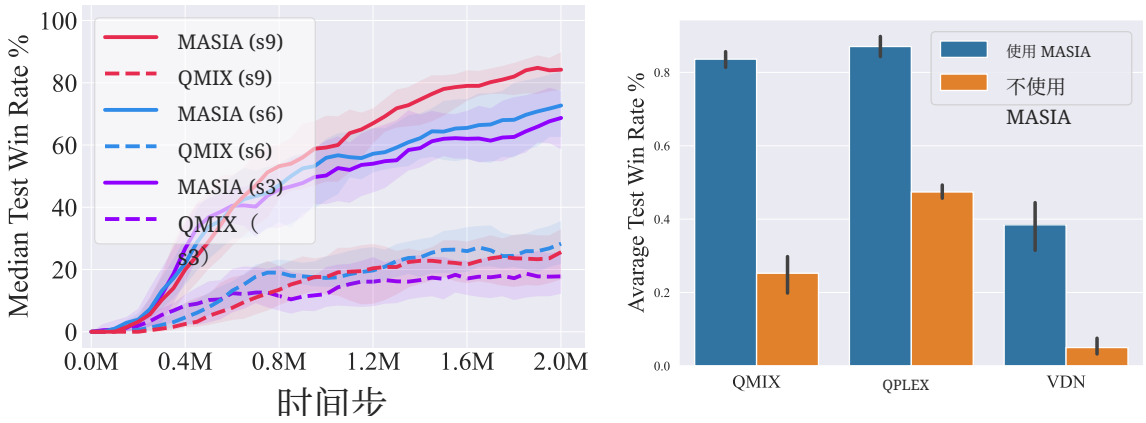
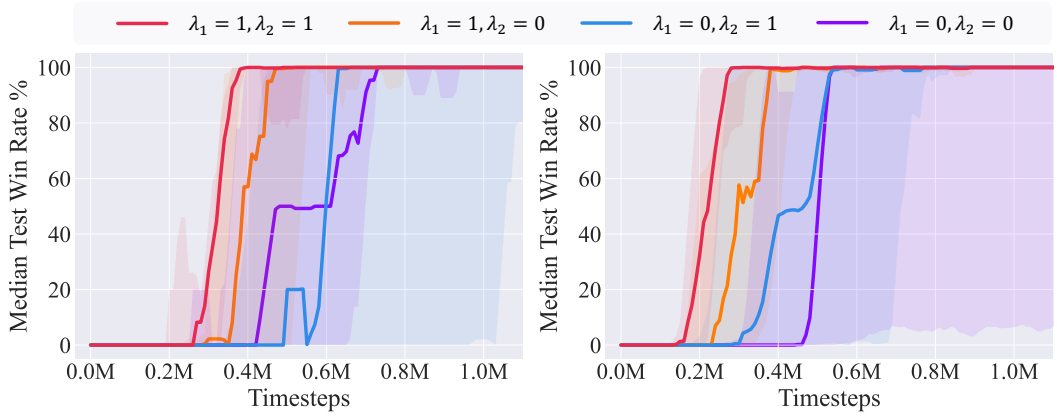


图7：(a) 不同视野范围下的性能比较，其中 sn 表示视野范围为n。(b) MASIA 在地图 1o10b_vs_1r 上带来的胜率提升。

为展示 MASIA框架 的通用性，我们还开展了将 MASIA 与当前基线方法相结合的实验，包括 VDN、QMIX 和 QPLEX。如图 7b 所示，与 MASIA 集成后，这些基线方法在来自 SMAC 的地图 1o10b_vs_1r 上的性能均得到大幅提升。在该场景中，一名监督者负责监控敌人。在没有通信的情况下，其他智能体必须在地图上全面搜索敌人。而借助可靠的通信，它们可以彼此以及与监督者进行交流，从而实现更好的协作。结果表明，MASIA 能够高效地聚合消息，并提升这些基于价值的多智能体强化学习方法中智能体的协作能力。

4.1.5 在线学习的消融实验

在我们的工作中，我们提出了两个表征目标，使聚合信息表征更加紧凑且充分。为了进一步验证这两个目标的有效性，我们针对在线实验开展了消融研究。具体而言，我们设计了三组消融实验：(1)$\lambda_1 = 1$， $\lambda_2 = 0$；(2)$\lambda_1 = 0$， $\lambda_2 = 1$；(3)$\lambda_1 = 0$， $\lambda_2 = 0$，它们分别对应于：(1) 仅使用编码器-解码器学习损失；(2) 仅使用潜在模型学习损失；(3) 不使用任何损失。而我们的方法（MASIA）对应于 $\lambda_0 = 1$， $\lambda_1 = 1$。Hallway 任务的实验结果如图 8所示。



(a) 走廊：4x6x10　　(b) Hallway：3x5-4x6x10

图 8：走廊环境在线学习的消融实验。

从实验结果中可以看出，采用两个表征目标的 MASIA 优于各组消融实验。所提出的两个表征目标有助于 MASIA 在 4x6x10 和 3x5-4x6x10 两种设置下更快地学会解决任务。特别是，第三组消融实验中的一些随机种子甚至无法在 2M 个样本内解决 3x5-4x6x10 任务，这表明这两个表征目标发挥着不可或缺的作用。它们提供了良好的指导，并加速了任务学习。

4.2 离线多智能体通信学习

目前，在多智能体通信强化学习领域，离线学习仍缺乏适当的评估标准。因此，为了探索利用离线数据集学习多智能体通信策略的可能性，并测试各种多智能体通信算法在离线设置下的有效性，我们构建了一组离线数据集并在其上开展实验。具体而言，为了使我们构建的数据集尽可能贴近真实世界的数据，我们借鉴了 D4RL [17] 的做法。D4RL 为单智能体离线学习提供了基准，我们以类似的方式收集数据来构建数据集。与 D4RL 不同，我们的数据集考虑了分散式部分可观测马尔可夫决策过程的属性，这意味着其中也包含每个智能体的观测。此外，由于我们关注的是多智能体通信设置，数据集还包含了每个时间步的消息接收方信息。我们在表 1 中介绍了整个数据集的结构。从表中可以看出，数据集中的每条轨迹都由每个智能体的个体观测和个体动作组成。此外，变量 接收方 ID_i^t 表示在时间步 i 时，时间步 t的智能体可以接收谁发送的消息。具体而言，数据通常以向量形式存储。例如，接收方 ID_i^t 可以存储为 $[0, 1]$，这表示智能体 0 和智能体 1 可以接收智能体 i 在时间步 t发送的消息。

考虑到现实世界的数据质量通常差异很大，我们设计了三种不同的数据集设置：1）**专家**、2）**噪声**和 3）**回放**。这些数据收集方案旨在系统地涵盖不同的现实场景。详情如下：

- **专家数据集**。我们训练在线策略直至收敛，并使用最终的专家策略贪心地采样数据。Fu 等人也采用了这种做法。 [17], Gulcehre 等人 [22] 和 Kumar 等人 [31]。
- **噪声数据集**。噪声数据集由专家策略生成，该策略通过 ε-贪心方法选择动作，并使用 $\varepsilon = 0.2$。通过固定的噪声策略创建数据集，与 Fujimoto 等人 [18], 、Kumar 等人 [31] 和 Gulcehre 等人 [22]的数据集收集过程类似。此方案旨在模拟现实场景，即专家可能偶尔会犯错。
- **回放数据集**　t。该数据集包含代表在线策略不同时期的样本　y

表 2： 三种不同等级行为策略的评估回报值范围 .

环境 \ 等级	Good	中等	Poor
走廊：4x6x10	0.75 - 1.0	0.5 - 0.75	0.0 - 0.5
Hallway：3x5-4x6x10	1.5 - 2.0	1.0 - 1.5	0.0 - 1.0
LBF：11x11-6p-4f-s1	0.75 - 1.0	0.5 - 0.75	0.0 - 0.5
LBF：20x20-10p-6f-s1	0.75 - 1.0	0.5 - 0.75	0.0 - 0.5
SMAC：1o2r _vs_4r	15 - 20	10 - 15	0 - 10
SMAC：1o10b _vs_1r	15 - 20	10 - 15	0 - 10
交通路口：容易	0 - (-7)	(-7) - (-17)	(-17) - (-600)
TJ：中等	0 - (-7)	(-7) - (-17)	(-17) - (-600)

学习，这种类似方案可见于 Agarwal 等人 [1] 和 Fujimoto 等人 [18]的研究。为考察不同质量的离线数据，我们纳入了三种形式的**回放**数据集。具体而言，我们在在线训练期间保存若干中间策略，并根据表2所示的评估回报值，将这些保存的策略人为划分为三个等级（较差、中等和良好）。随后，我们收集了三种由不同行为策略混合构成的数据集：1）**回放（较差）**完全由较差等级行为策略收集的数据（差等级数据）构成；2）**回放（中等）**由 80% 的中等级数据和 20% 的差等级数据构成；（3）**回放（良好）**由 60% 的好等级数据、20% 的中等级数据和 20% 的差等级数据构成。

4.2.1 通信性能

考虑到离线学习中已得到深入研究的 Q 发散问题， [35] 我们提出将这些通信算法与离线多智能体强化学习算法 ICQ 结合， [74], 该算法采用保守学习范式来缓解 Q 发散问题。具体而言，我们将不同通信算法的设计整合到行动者网络结构中，从而使智能体能够学习基于消息的策略。也就是说，智能体根据通信消息做出决策。我们还补充了 ICQ 算法本身的实验结果作为消融实验，以验证智能体通信的有效性。整体结果列于表3中。尽管离策略评估（OPE） [62] 通常用于离线强化学习中测试最终策略的性能，但为了提供更准确的评估，我们直接在环境中评估策略并报告评估结果。事实上，令人欣喜的是，在不同的离线数据集设置下，MASIA 在大多数任务中都取得了最佳通信性能。例如，使用**专家**离线数据集进行学习时，除 SMAC: 1o2r_vs_4r 外，MASIA 在所有场景中都取得了最佳性能。而在任务 SMAC: 1o2r_vs_4r 上，尽管方差略大，MASIA 的平均性能仍与 Full-Comm 一样良好。再以 Hallway：3x5-4x6x10 任务为例，MASIA 在所有离线数据集模式下都取得了最高成功率，尤其是在**回放**数据集上，性能远超所有其他基线方法。MASIA 在**回放（较差）**和**回放（中等）**数据集上的良好表现证明了 MASIA 的鲁棒性，因为即使使用低质量数据，它仍能学习良好的通信策略。

为了进一步比较 MASIA 与其他基线方法在不同任务上的学习趋势，我们还选择性地展示了 Hallway：3x5-4x6x10 和 SMAC：1o2r 的部分数据集上的离线实验学习曲线_与_4r，见图9。如图9a 和9b 所示，MASIA 在 Hallway：3x5-4x6x10 的学习曲线中展现出更快的学习速度和更好的收敛性能。在**回放（较差）**数据集上，MASIA 是唯一一种能够在 4M 个学习样本内达到超过 60% 测试成功率中位数的方法。同样，在 Hallway：3x5-4x6x10 的**回放（中等）**数据集上，只有 MASIA 能够稳定收敛到超过 80% 的成功率，而其他所有方法均未能做到。在 SMAC：1o2r_与_4r 上，MASIA 同样表现出更快的收敛速度和更好的收敛性能，与 Full-Comm 算法不相上下。

表 3：不同基准测试下不同方法的实验结果。

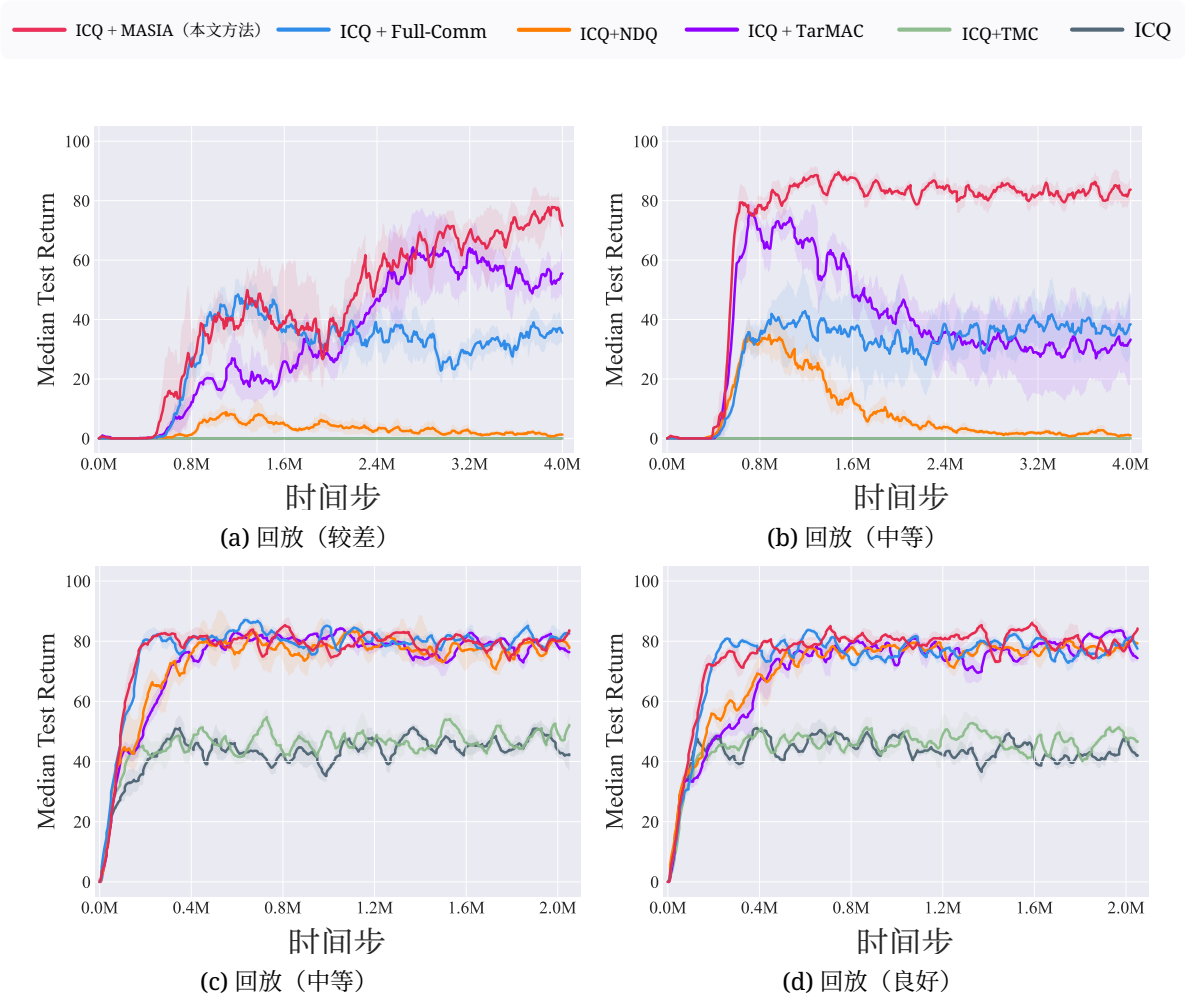
环境/数据集 \ 方法		ICQ+MASIA	ICQ+Fullcomm	ICQ+NDQ	ICQ+TarMAC	ICQ+TMC	ICQ
走廊：4x6x10	专家	**100.00 ± 0.00**	**100.00 ± 0.00**	85.10 ± 5.55	97.80 ± 4.28	0.04 ± 0.24	0.25 ± 0.64
	噪声	**99.90 ± 0.07**	99.60 ± 0.95	99.50 ± 0.74	99.80 ± 0.51	14.10 ± 11.62	1.55 ± 2.82
	回放（较差）	**94.50 ± 3.59**	87.90 ± 8.95	62.20 ± 9.21	83.60 ± 7.92	0.52 ± 0.94	0.11 ± 0.35
	回放（中等）	**97.40 ± 3.60**	94.40 ± 5.07	57.40 ± 8.24	92.80 ± 5.51	1.39 ± 1.69	0.12 ± 0.40
	回放（良好）	**99.70 ± 0.72**	91.10 ± 8.46	11.30 ± 13.95	97.70 ± 2.59	0.33 ± 1.01	0.02 ± 0.17
Hallway：3x5-4x6x10	专家	**99.70 ± 0.57**	98.90 ± 0.48	86.50 ± 5.17	98.50 ± 2.26	1.18 ± 1.48	0.36 ± 0.73
	噪声	**99.90 ± 0.10**	99.90 ± 0.20	99.10 ± 1.17	99.80 ± 1.84	7.67 ± 4.43	5.97 ± 3.71
	回放（较差）	**73.40 ± 10.63**	40.80 ± 20.46	10.20 ± 16.61	53.40 ± 26.28	0.00 ± 0.00	0.00 ± 0.00
	回放（中等）	**81.90 ± 11.54**	38.60 ± 15.92	2.78 ± 4.02	35.60 ± 24.91	0.00 ± 0.00	0.00 ± 0.00
	回放（良好）	**95.20 ± 4.02**	77.50 ± 13.98	76.50 ± 12.65	88.30 ± 8.35	0.01 ± 0.10	0.06 ± 0.46
LBF：11x11-6p-4f-s1	专家	**88.70 ± 13.09**	76.50 ± 15.61	82.60 ± 4.33	88.30 ± 11.37	80.50 ± 14.13	82.90 ± 12.53
	噪声	**89.60 ± 9.42**	78.40 ± 14.59	80.60 ± 4.93	88.80 ± 10.09	82.10 ± 10.94	80.10 ± 11.59
	回放（较差）	77.20 ± 15.14	66.40 ± 15.53	70.10 ± 5.38	**78.10 ± 13.81**	70.90 ± 15.14	71.20 ± 12.55
	回放（中等）	**82.70 ± 10.92**	70.70 ± 15.93	74.80 ± 5.74	80.50 ± 11.75	73.60 ± 13.45	75.70 ± 13.13
	回放（良好）	**89.10 ± 11.41**	76.90 ± 13.43	79.30 ± 4.60	87.40 ± 8.96	78.20 ± 12.09	78.30 ± 11.92
LBF：20x20-10p-6f-s1	专家	**54.70 ± 10.39**	42.90 ± 10.53	45.80 ± 4.27	52.50 ± 10.83	46.40 ± 10.26	44.90 ± 10.53
	噪声	54.10 ± 10.66	43.30 ± 11.89	47.30 ± 5.30	**54.60 ± 12.71**	48.70 ± 11.72	46.60 ± 11.92
	回放（较差）	**48.90 ± 12.75**	36.80 ± 11.49	44.50 ± 4.85	44.78 ± 10.23	45.10 ± 12.29	45.10 ± 11.72
	回放（中等）	**50.20 ± 10.92**	39.90 ± 11.04	46.40 ± 5.13	48.10 ± 11.85	45.70 ± 10.45	44.70 ± 10.56
	回放（良好）	**51.50 ± 12.78**	41.70 ± 10.38	45.60 ± 5.50	48.90 ± 11.81	46.30 ± 10.54	45.20 ± 10.19
SMAC：1o2r vs 4r_ _	专家	83.60 ± 17.68	**83.60 ± 15.04**	79.30 ± 12.71	80.70 ± 19.69	45.60 ± 25.81	43.70 ± 25.58
	噪声	**81.10 ± 15.08**	76.70 ± 21.21	77.40 ± 10.59	79.90 ± 18.75	57.90 ± 26.88	41.50 ± 25.26
	回放（较差）	**85.70 ± 15.61**	83.60 ± 19.86	82.10 ± 10.76	84.70 ± 15.56	40.30 ± 25.85	45.40 ± 25.34
	回放（中等）	79.60 ± 18.17	**80.50 ± 20.29**	77.70 ± 10.86	79.20 ± 18.55	50.20 ± 26.39	45.20 ± 25.71
	回放（良好）	**80.10 ± 15.26**	75.90 ± 23.41	77.90 ± 10.78	19.60 ± 21.23	48.80 ± 26.75	43.80 ± 26.68
SMAC：1o10b 对 1r_ _	专家	**84.10 ± 21.36**	77.30 ± 21.48	82.80 ± 12.41	74.30 ± 23.08	11.90 ± 16.75	9.34 ± 18.37
	噪声	**84.60 ± 17.77**	82.40 ± 18.89	78.70 ± 10.19	58.20 ± 26.63	13.30 ± 19.76	15.60 ± 18.11
	回放（较差）	**82.90 ± 17.19**	78.70 ± 21.93	77.70 ± 10.05	80.90 ± 20.91	7.70 ± 11.57	9.79 ± 18.18
	回放（中等）	82.90 ± 21.26	82.80 ± 18.54	81.70 ± 8.90	**85.40 ± 16.14**	9.19 ± 14.65	7.77 ± 13.97
	回放（良好）	**85.60 ± 17.79**	80.90 ± 20.71	81.10 ± 10.19	80.30 ± 23.97	10.30 ± 17.21	6.44 ± 12.16
交通路口：容易	专家	**96.70 ± 10.04**	68.60 ± 29.36	89.70 ± 7.51	96.70 ± 12.93	92.00 ± 16.31	91.20 ± 18.64
	噪声	**97.90 ± 9.93**	93.60 ± 12.65	96.20 ± 4.39	96.40 ± 11.47	95.80 ± 10.24	95.10 ± 16.29
	回放（较差）	91.10 ± 19.53	75.60 ± 27.95	95.30 ± 5.22	92.20 ± 16.53	90.20 ± 20.76	**95.50 ± 13.71**
	回放（中等）	87.30 ± 21.93	71.90 ± 28.52	93.60 ± 7.89	89.80 ± 18.95	**98.10 ± 9.69**	96.80 ± 12.41
	回放（良好）	90.80 ± 20.92	75.60 ± 27.01	89.20 ± 12.01	**96.70 ± 10.23**	94.40 ± 17.71	96.30 ± 13.93
TJ：中等	专家	**88.60 ± 19.98**	15.80 ± 24.56	78.70 ± 14.59	75.50 ± 31.17	81.50 ± 28.94	87.90 ± 22.49
	噪声	72.70 ± 29.63	48.20 ± 32.49	71.30 ± 17.58	56.40 ± 34.31	76.10 ± 28.35	**78.10 ± 26.14**
	回放（较差）	**99.60 ± 3.04**	69.90 ± 30.91	96.30 ± 5.30	98.10 ± 6.89	97.50 ± 9.80	98.10 ± 9.24
	回放（中等）	89.90 ± 19.21	49.20 ± 33.36	83.70 ± 11.69	87.80 ± 21.61	**91.20 ± 17.81**	80.10 ± 27.11
	回放（良好）	92.60 ± 14.81	72.20 ± 30.12	90.30 ± 6.89	85.90 ± 23.69	**96.10 ± 12.88**	91.10 ± 20.63

4.2.2 离线学习的消融研究

在前面的实验中，我们证明了在线设置下所提出的两个无监督表示学习目标的重要性。为了研究这两个目标在离线设置下的作用，我们进一步针对离线实验开展了消融研究。在 Hallway：3x5-4x6x10 任务上，使用数据集**回放（中等）**和**回放（较差）**的消融结果如图10所示。

与在线实验的消融研究类似，这里我们分别对两个无监督目标进行消融。从结果可以看出，消融任一目标都会导致最终通信性能下降，这表明所设计的两个无监督目标在离线设置中发挥着不可或缺的作用。事实上，可以看出，在离线设置中消融无监督学习目标的影响相对大于在线实验。造成这一现象的主要原因是，在离线设置中，由于不允许在环境中进行探索，因此学习良好的通信策略更具挑战性。此外，数据集质量问题，尤其是数据集**回放（较差）**，加剧了这一问题。

此外，我们还在离线实验中加入了针对表征预训练做法的消融实验。从结果可以看出，去除表征预训练也会严重损害性能，这证明了该做法的有效性。更多消融实验与分析请参见附录C.2。



(a) 回放（较差）　(b) 回放（中等）

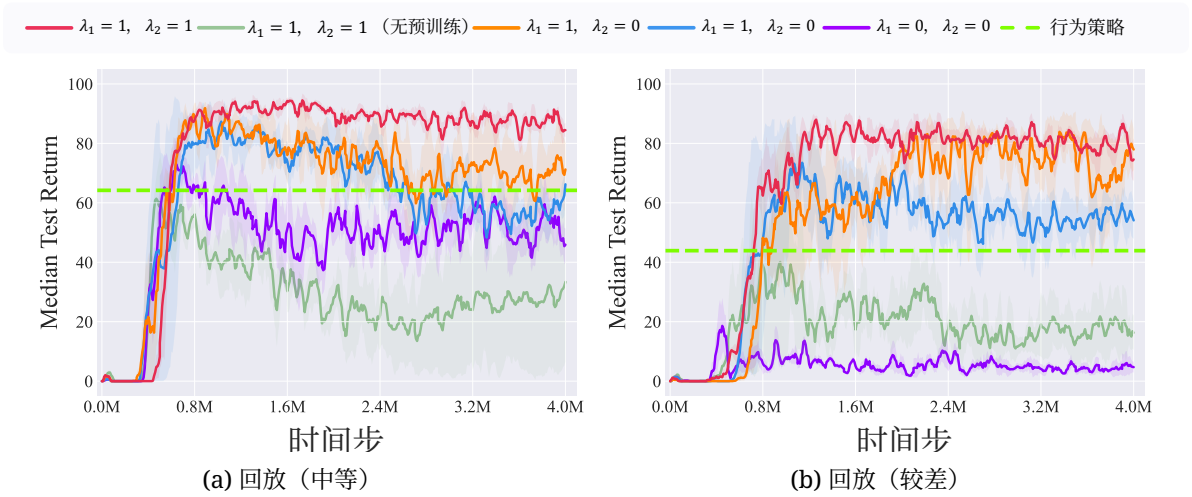
(c) 回放（中等）　(d) 回放（良好）

图 9：离线实验的学习曲线。(a) 和 (b) 在 Hallway：3x5-4x6x10 上进行，(c) 和 (d) 在 SMAC：1o2r_vs_4r 上进行。

5 相关工作

多智能体强化学习（MARL）。近年来，多智能体强化学习取得了显著进展。许多方法诞生于集中训练、分散执行范式之下，旨在缓解非平稳性问题，并在近年来取得了显著进展。它们大致可分为基于策略的方法和基于价值的方法。典型的策略梯度方法包括 MADDPG [38], COMA [16], MAAC [25],SQDDPG [66], FOP [83], 和 HAPPO [29]，这些方法探索了多智能体策略梯度方法的优化；而基于价值的方法则主要关注全局价值函数的分解 [63, 9]。VDN [57] 采用简单的加性分解，将联合价值函数分解为智能体级价值函数。QMIX [52] 在结构上强制学习到的联合价值函数相对于智能体的效用保持单调，从而能够表示更丰富的一类价值函数。QPLEX [64] 进一步采用双重竞争网络架构对联合价值函数进行分解，实现了个体-全局最大化（IGM）的完全表达能力 [55]。

多智能体通信。在部分可观测条件下，通信在多智能体协作中发挥着重要作用 [13, 84]。人们已广泛研究如何学习通信协议，以



(a) 回放（中等）　　(b) 回放（较差）

Fi图 10：Hallway：3x5-4x6x10 任务的离线学习消融实验。

提升协作任务的表现 [20, 15, 34, 73, 12, 37, 77, 71, 21]。以往工作可分为两类。一类侧重于为消息发送方生成有意义的消息。最简单的方法是将原始局部观测或局部信息历史记录视为消息 [15, 56]。VBC [81]和 TMC [82] 在发送方采用基于方差的控制和时间平滑等技术，使生成的消息对策略学习更有意义、更有价值。NDQ [67] 为不同队友生成精简消息，以学习近似可分解价值函数，并基于两种不同的基于信息论的正则项优化消息生成器，从而实现富有表现力的通信。相反，另一类工作试图在接收方高效地提取最有用的消息，并设计机制来区分消息的重要性。I2C [8] 和 ACML [40] 采用门控机制，有选择地接收消息。还有一些工作受到注意力机制广泛应用的启发 [3, 7]。TarMAC [6] 通过一种简单的基于签名的软注意力机制实现定向通信：发送方广播一个编码智能体属性的键，随后接收方对所有收到的消息进行注意力加权求和，以作出决策。SARNet [51] 和 MAGIC [44] 进一步移除了 TarMAC 中的签名，并利用基于注意力的网络学习实体之间高效且可解释的关系，从而决定何时与谁通信。

离线多智能体强化学习。离线强化学习 [35] 因其无需与环境交互的数据驱动训练范式而受到广泛关注 [50]。先前的工作 [19]讨论了离线学习中的分布偏移问题，并考虑学习行为约束策略，以缓解基于未见数据估计所导致的外推误差 [70, 30, 32]。离线多智能体强化学习是一个富有前景的研究方向 [80]，其从静态数据集训练策略。继承在线多智能体强化学习方法——即将策略梯度算法扩展到多智能体场景 [38, 16, 68]，或采用结合价值分解的Q学习范式 [57, 52, 55, 64],——现有离线多智能体强化学习方法尝试利用带有策略约束的离线数据。ICQ [74] 仅信任离线数据，从而有效缓解了外推误差。MABCQ [26] 引入了完全去中心化的离线多智能体强化学习设定，并利用价值偏差和转移归一化技术实现高效学习。OMAR [46] 结合一阶策略梯度与零阶优化方法，以避免不协调的局部最优。MADT [42] 利用Transformer的序列建模能力，并将其无缝整合到离线和在线多智能体强化学习任务中。 [58] 研究了明确考虑智能体轨迹多样性的离线多智能体强化学习，并提出名为共享个体轨迹（SIT）的新框架来解决这一问题。 [59] 提出首先训练一个能够访问每个智能体的观测、动作和奖励的教师策略。教师策略识别并重新组合数据集中的“良好”行为后，他们创建独立的学生策略并进行蒸馏，

不仅将教师策略的特征，还将其不同智能体特征之间的结构关系一并传递给学生策略。ODIS [79] 提出了一种新颖的离线多智能体强化学习算法，用于从多任务数据中发现协作技能（ODIS）。 [5] 近期发布了一个名为Off-the-Grid多智能体强化学习（OG-MARL）的框架，用于在无通信条件下生成离线多智能体强化学习数据集和算法；该框架发布了首批数据集和基线方法，用于协作式离线多智能体强化学习，并配套提供了标准化的评估协议。

据我们所知，现有的多智能体强化学习通信方法均未明确考虑如何优化接收到的多条消息，以提高策略学习效率。队友提供的冗余信息可能会使智能体感到困惑，而仅将原始消息并入局部策略，也可能增加学习负担。同时，目前还没有针对离线多智能体通信设置的测试平台。我们提出的方法采用消息聚合模块来学习紧凑的信息表示，并在在线和离线设置中提取与决策最相关的部分。

6 结论与未来工作

本文研究了多智能体通信中的信息表示。以往工作要么着重于生成有意义的消息，要么设计机制以原始方式选择最相关的消息，却忽略了消息聚合，导致在复杂场景中的样本效率较低。我们的方法通过学习紧凑的信息表示来锚定真实状态，并以自监督方式对其进行优化，从而提高通信效率。此外，我们还采用聚焦网络提取与决策最相关的部分。我们在多种基准测试中进行了充分实验，以验证所提方法在在线和离线设置中的有效性；更多可视化结果也进一步揭示了我们的方法为何有效。我们还发布了新构建的多智能体通信离线基准，希望能推动多智能体强化学习通信在现实中的应用。未来，值得进一步研究图像输入相关结果，以及如何通过智能体分组等技术解决面对数百或数千个智能体时的可扩展性问题。此外，在在线策略部署中遭遇分布偏移时，如何获得鲁棒的通信策略也是亟待研究的课题。

致谢

本研究得到中国国家重点研发计划（2020AAA0107200）、中国国家自然科学基金（61921006、61876119、62276126）、江苏省自然科学基金（BK20221442）以及南京大学优秀博士生候选人B计划的支持。感谢 Lichao Zhang 和 Chuneng Fan 提供的宝贵支持、建议与讨论。

参考文献

[1] Rishabh Agarwal、Dale Schuurmans 和 Mohammad Norouzi。对离线强化学习的乐观视角。载于国际机器学习会议，第 104–114 页。PMLR，2020年。[2] Lorenzo Canese、Gian Carlo Cardarilli、Luca Di Nunzio、Rocco Fazzolari、Daniele Giardino、Marco Re 和 Sergio Spanò。多智能体强化学习：挑战与应用综述。*AppliedSciences*，11(11):4948，2021年。[3] Sneha Chaudhari、Varun Mithal、Gungor Polatkan 和 Rohan Ramanath。注意力模型综述。*arXiv* 预印本 *arXiv:1904.02874*，2019年。[4] 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。载于国际机器学习会议，第 1989–1998 页，2021年。

[5] Formanek Claude、Jeewa Asad、Shock Jonathan 和 Arnu Pretorius。Off-the-Grid多智能体强化学习：用于生成数据集的框架，以及协作式离线多智能体强化学习的基线方法。*arXiv* 预印本 *arXiv:2302.00521*，2023年。[6] Abhishek Das、Théophile Gervet、Joshua Romoff、Dhruv Batra、Devi Parikh、Mike Rabbat 和 Joelle Pineau。TarMAC：目标导向的多智能体通信。载于国际机器学习会议，第 1538–1546 页，2019年。[7] Alana de Santana Correia 和 Esther Luna Colombini。请注意！深度学习中神经注意力模型综述。*arXiv* 预印本 *arXiv:2103.16775*，2021年。[8] Ziluo Ding、Tiejun Huang 和 Zongqing Lu。为多智能体协作学习个体推断通信。载于神经信息处理系统进展，第 22069–22079 页，2020年。[9] Zehao Dou、Jakub Grudzien Kuba 和 Yaodong Yang。理解深度协作式多智能体强化学习中的价值分解算法。*arXiv* 预印本 *arXiv:2202.04868*，2022年。[10] Wei Du 和 Shifei Ding。多智能体深度强化学习综述：从挑战与应用的角度出发。《人工智能评论》，54(5):3215–3238，2021年。[11] Xin Du、Jiahai Wang、Siyuan Chen 和 Zhiyue Liu。通过时空特征融合进行交通信号控制的多智能体深度强化学习。载于欧洲机器学习与数据库知识发现联合会议，第 470–485 页，2021年。[12] Yali Du、Bo Liu、Vincent Moens、Ziqi Liu、Zhicheng Ren、Jun Wang、Xu Chen 和 Haifeng Zhang。在多智能体强化学习中学习相关通信拓扑。载于国际自主智能体与多智能体系统会议，第 456–464 页，2021年。[13] Tom Eccles、Yoram Bachrach、Guy Lever、Angeliki Lazaridou 和 Thore Graepel。多智能体强化学习中涌现通信的偏置。载于神经信息处理系统进展，第 13111–13121 页，2019年。[14] Linus Ericsson、Henry Gouk、Chen Change Loy 和 Timothy M Hospedales。自监督表征学习：简介、进展与挑战。*arXiv* 预印本 *arXiv:2110.09327*，2021年。[15] Jakob Foerster、Ioannis Alexandros Assael、Nando De Freitas 和 Shimon Whiteson。通过深度多智能体强化学习学习通信。载于神经信息处理系统进展，第 2137–2145 页，2016年。[16] Jakob N. Foerster、Gregory Farquhar、Triantafyllos Afouras、Nantas Nardelli 和 Shimon Whiteson。反事实多智能体策略梯度。载于*AAAI*人工智能会议，第 2974–2982 页，2018年。[17] Justin Fu、Aviral Kumar、Ofir Nachum、George Tucker 和 Sergey Levine。D4RL：深度数据驱动强化学习数据集。*arXiv* 预印本 *arXiv:2004.07219*，2020年。[18] Scott Fujimoto、Edoardo Conti、Mohammad Ghavamzadeh 和 Joelle Pineau。批量深度强化学习算法的基准测试。*arXiv* 预印本 *arXiv:1910.01708*，2019年。[19] Scott Fujimoto、David Meger 和 Doina Precup。无需探索的离策略深度强化学习。载于国际机器学习会议，第 2052–2062 页，2019年。[20] C Lee Giles 和 Kam-Chuen Jim。为多智能体系统学习通信。载于激进智能体概念研讨会，第 377–390 页。施普林格，2002年。[21] Cong Guan、Feng Chen、Lei Yuan、Chenghe Wang、Hao Yin、Zongzhang Zhang 和 Yang Yu。基于自监督信息聚合的高效多智能体通信。载于神经信息处理系统进展，2022年。

[22] Caglar Gulcehre, Sergio Gómez Colmenarejo, Ziyu Wang, Jakub Sygnowski, Thomas Paine, Konrad Zolna, Yutian Chen, Matthew Hoffman, Razvan Pascanu 和 Nando de Freitas。正则化行为价值估计。*arXiv*预印本 *arXiv:2103.09575*，2021年。[23] Matthew J. Hausknecht 和 Peter Stone。面向部分可观测MDP的深度循环Q学习。载于*AAAI*人工智能会议，第29–37页，2015年。[24] Pablo Hernandez-Leal、Bilal Kartal 和 Matthew E Taylor。多智能体深度强化学习综述与评析。《自主智能体与多智能体系统》，33(6):750–797，2019年。[25] Shariq Iqbal 和 Fei Sha。用于多智能体强化学习的演员-注意力-评论家。载于国际机器学习会议，第2961–2970页，2019年。[26] Jiechuan Jiang 和 Zongqing Lu。离线去中心化多智能体强化学习。*arXiv*预印本 *arXiv:2108.01832*，2021年。[27] Woojun Kim、Jongeui Park 和 Youngchul Sung。多智能体强化学习中的通信：意图共享。载于国际学习表征会议，2020年。[28] Landon Kraemer 和 Bikramjit Banerjee。多智能体强化学习作为去中心化规划的演练。《神经计算》，190:82–94，2016年。[29] Jakub Grudzien Kuba、Ruiqing Chen、Munning Wen、Ying Wen、Fanglei Sun、Jun Wang 和 Yaodong Yang。多智能体强化学习中的信赖域策略优化。*arXiv*预印本 *arXiv:2109.11251*，2021年。[30] Aviral Kumar、Justin Fu、Matthew Soh、George Tucker 和 Sergey Levine。通过自举误差缩减稳定离策略Q学习。载于神经信息处理系统进展，第32卷，2019年。[31] Aviral Kumar、Aurick Zhou、George Tucker 和 Sergey Levine。面向离线强化学习的保守Q学习。神经信息处理系统进展，33:1179–1191，2020年。[32] Aviral Kumar、Aurick Zhou、George Tucker 和 Sergey Levine。面向离线强化学习的保守Q学习。载于神经信息处理系统进展，第33卷，第1179–1191页，2020年。[33] Michael Laskin、Aravind Srinivas 和 Pieter Abbeel。CURL：用于强化学习的对比无监督表征。载于国际机器学习会议，第5639–5650页，2020年。[34] Angeliki Lazaridou 和 Marco Baroni。深度学习时代涌现式多智能体通信。*arXiv*预印本 *arXiv:2006.02419*，2020年。[35] Sergey Levine、Aviral Kumar、George Tucker 和 Justin Fu。离线强化学习：教程、综述与开放问题展望。*arXiv*预印本 *arXiv:2005.01643*，2020年。[36] Tung-Che Liang、Jin Zhou、Yun-Sheng Chan、Tsung-Yi Ho、Krishnendu Chakrabarty 和 Cy Lee。使用多智能体强化学习在MEDA生物芯片上进行液滴并行控制。载于国际机器学习会议，第6588–6599页，2021年。[37] Toru Lin、Jacob Huh、Christopher Stauffer、Ser Nam Lim 和 Phillip Isola。利用自编码器学习多智能体通信的落地。载于神经信息处理系统进展，第15230–15242页，2021年。[38] Ryan Lowe、Yi Wu、Aviv Tamar、Jean Harb、Pieter Abbeel 和 Igor Mordatch。面向混合合作竞争环境的多智能体演员-评论家。载于神经信息处理系统进展，第6379–6390页，2017年。

[39] Xueguang Lyu、Yuchen Xiao、Brett Daley 和 Christopher Amato。多智能体强化学习中集中式与去中心化评论家的比较。*arXiv*预印本 *arXiv:2102.04402*，2021年。[40] Hangyu Mao、Zhengchao Zhang、Zhen Xiao、Zhibo Gong 和 Yan Ni。通过消息剪枝，在有限带宽下学习智能体通信。载于*AAAI*人工智能会议，第5142–5149页，2020年。[41] Weichao Mao、Kaiqing Zhang、Erik Miehling 和 Tamer Ba¸sar。部分可观测协作式多智能体强化学习中的信息状态嵌入。载于*IEEE*决策与控制会议，第6124–6131页，2020年。[42] Linghui Meng、Muning Wen、Yaodong Yang、Chenyang Le、Xiyun Li、Weinan Zhang、Ying Wen、Haifeng Zhang、Jun Wang 和 Bo Xu。离线预训练多智能体决策变换器：一个大型序列模型征服所有《星际争霸II》任务。*arXiv*预印本 *arXiv:2112.02845*，2021年。[43] Volodymyr Mnih、Koray Kavukcuoglu、David Silver、Andrei A. Rusu、Joel Veness、Marc G. Bellemare、Alex Graves、Martin A. Riedmiller、Andreas Fidjeland、Georg Ostrovski、Stig Petersen、Charles Beattie、Amir Sadik、Ioannis Antonoglou、Helen King、Dharshan Kumaran、Daan Wierstra、Shane Legg 和 Demis Hassabis。通过深度强化学习实现人类水平的控制。《自然》，518(7540):529–533，2015年。[44] Yaru Niu、Rohan Paleja 和 Matthew Gombolay。多智能体图注意力通信与协同组队。载于国际自主智能体与多智能体系统会议，第964–973页，2021年。[45] Frans A Oliehoek 和 Christopher Amato。《去中心化部分可观测马尔可夫决策过程简明导论》。施普林格，2016年。[46] Ling Pan、Longbo Huang、Tengyu Ma 和 Huazhe Xu。秉持保守原则，实现更优规划：通过演员网络校正进行离线多智能体强化学习。载于国际机器学习会议，第17221–17237页，2022年。[47] Georgios Papoudakis、Filippos Christianos、Arrasy Rahman 和 Stefano V Albrecht。应对多智能体深度强化学习中的非平稳性。*arXiv*预印本 *arXiv:1906.04737*，2019年。[48] Georgios Papoudakis、Filippos Christianos、Lukas Schäfer 和 Stefano V Albrecht。协作任务中的多智能体深度强化学习算法基准测试。载于神经信息处理系统数据集与基准测试专题论文集，2021年。[49] Xue Bin Peng、Aviral Kumar、Grace Zhang 和 Sergey Levine。优势加权回归：简单且可扩展的离策略强化学习。*arXiv*预印本 *arXiv:1910.00177*，2019年。[50] Rafael Figueiredo Prudencio、Marcos ROA Maximo 和 Esther Luna Colombini。离线强化学习综述：分类体系、回顾与开放问题。*arXiv*预印本 *arXiv:2203.01387*，2022年。[51] Murtaza Rangwala 和 Ryan Williams。通过结构化注意力推理学习多智能体通信。载于神经信息处理系统进展，第10088–10098页，2020年。[52] Tabish Rashid、Mikayel Samvelyan、Christian Schroeder、Gregory Farquhar、Jakob Foerster 和 Shimon Whiteson。QMIX：用于深度多智能体强化学习的单调价值函数分解。载于国际机器学习会议，第4295–4304页，2018年。[53] Mikayel Samvelyan、Tabish Rashid、Christian Schröder de Witt、Gregory Farquhar、Nantas Nardelli、Tim GJ Rudner、Chia-Man Hung、Philip HS Torr、Jakob N Foerster 和 Shimon Whiteson。星际争霸多智能体挑战。载于国际自主智能体与多智能体系统会议，第2186–2188页，2019年。

[54] Max Schwarzer、Ankesh Anand、Rishab Goel、R Devon Hjelm、Aaron Courville 和 Philip Bachman。基于自预测表征的数据高效强化学习。发表于国际学习表征会议，2020年。[55] Kyunghwan Son、Daewoo Kim、Wan Ju Kang、David Earl Hostallero 和 Yung Yi。QTRAN：通过变换学习分解协作式多智能体强化学习。发表于国际机器学习会议，第5887–5896页，2019年。[56] Sainbayar Sukhbaatar、Arthur Szlam 和 Rob Fergus。通过反向传播学习多智能体通信。发表于神经信息处理系统进展，第2244–2252页，2016年。[57] Peter Sunehag、Guy Lever、Audrunas Gruslys、Wojciech Marian Czarnecki、Vinícius Flores Zambaldi、Max Jaderberg、Marc Lanctot、Nicolas Sonnerat、Joel Z Leibo、Karl Tuyls 和 Thore Graepel。基于团队奖励的协作式多智能体学习价值分解网络。发表于国际自主智能体与多智能体系统会议，第2085–2087页，2018年。[58] Qi Tian、Kun Kuang、Furui Liu 和 Baoxiang Wang。在离线多智能体强化学习中从优质轨迹中学习。*arXiv*预印本 *arXiv:2211.15612*，2022年。[59] Wei-Cheng Tseng、Tsun-Hsuan Wang、Yen-Chen Lin 和 Phillip Isola。通过知识蒸馏进行离线多智能体强化学习。发表于神经信息处理系统进展，2022年。[60] Laurens Van der Maaten 和 Geoffrey Hinton。使用t-SNE可视化数据。机器学习研究杂志，9(11):2579–2605，2008年。[61] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力机制即所需的一切。发表于神经信息处理系统进展，第5998–6008页，2017年。[62] Cameron Voloshin、Hoang Minh Le、Nan Jiang 和 Yisong Yue。强化学习离策略策略评估的实证研究。发表于神经信息处理系统进展，2021年。[63] Jianhao Wang、Zhizhou Ren、Beining Han、Jianing Ye 和 Chongjie Zhang。迈向理解基于价值分解的协作式多智能体Q学习。发表于神经信息处理系统进展，第29142–29155页，2021年。[64] Jianhao Wang、Zhizhou Ren、Terry Liu、Yang Yu 和 Chongjie Zhang。QPLEX：双重决斗多智能体Q学习。发表于国际学习表征会议，2021年。[65] Jianhong Wang、Wangkun Xu、Yunjie Gu、Wenbin Song 和 Tim C. Green。配电网络主动电压控制的多智能体强化学习。发表于神经信息处理系统进展，第 3271–3284页，2021年。[66] Jianhong Wang、Yuan Zhang、Tae-Kyun Kim 和 Yunjie Gu。夏普利Q值：解决全局奖励博弈的一种局部奖励方法。发表于*AAAI*人工智能会议，第7285–7292页，2020年。[67] Tonghan Wang、Jianhao Wang、Chongyi Zheng 和 Chongjie Zhang。通过通信最小化学习近似可分解价值函数。发表于国际学习表征会议，2020年。[68] Yihan Wang、Beining Han、Tonghan Wang、Heng Dong 和 Chongjie Zhang。DOP：离策略多智能体分解策略梯度。发表于国际学习表征会议，2021年。[69] Muning Wen、Jakub Grudzien Kuba、Runji Lin、Weinan Zhang、Ying Wen、Jun Wang 和 Yaodong Yang。多智能体强化学习是一个序列建模问题。发表于神经信息处理系统进展，2022年。

[70] Yifan Wu、George Tucker 和 Ofir Nachum。行为正则化离线强化学习。*arXiv*预印本 *arXiv:1911.11361*，2019年。[71] Di Xue、Lei Yuan、Zongzhang Zhang 和 Yang Yu。通过夏普利消息价值实现高效多智能体通信。发表于国际人工智能联合会议，第578–584页，2022年。[72] Ke Xue、Jiacheng Xu、Lei Yuan、Miqing Li、Chao Qian、Zongzhang Zhang 和 Yang Yu。多智能体动态算法配置。发表于神经信息处理系统进展，2022年。[73] Wanqi Xue、Wei Qiu、Bo An、Zinovi Rabinovich、Svetlana Obraztsova 和 Chai Kiat Yeo。说错还是误导：实现多智能体通信强化学习的鲁棒性。发表于国际自主智能体与多智能体系统会议，第1418–1426页，2022年。[74] Yiqin Yang、Xiaoteng Ma、Chenghao Li、Zewu Zheng、Qiyuan Zhang、Gao Huang、Jun Yang 和 Qianchuan Zhao。眼见为实：离线多智能体强化学习的隐式约束方法。发表于神经信息处理系统进展，2021年。[75] Jianing Ye、Chenghao Li、Jianhao Wang 和 Chongjie Zhang。通过序贯变换实现协作式多智能体强化学习的全局最优性。*arXiv*预印本 *arXiv:2207.11143*，2022年。[76] Lei Yuan、Chenghe Wang、Jianhao Wang、Fuxiang Zhang、Feng Chen、Cong Guan、Zongzhang Zhang、Chongjie Zhang 和 Yang Yu。通过去中心化任务表征实现多智能体集中式协调。发表于国际人工智能联合会议，第599–605页，2022年。[77] Lei Yuan、Jianhao Wang、Fuxiang Zhang、Chenghe Wang、Zongzhang Zhang、Yang Yu 和 Chongjie Zhang。通过去中心化队友建模实现多智能体激励通信。发表于*AAAI*人工智能会议，第9466–9474页，2022年。[78] Mohamed Salah Zaïem 和 Etienne Bennequin。多智能体强化学习中的通信学习：综述。*arXiv*预印本 *arXiv:1911.05438*，2019年。[79] Fuxiang Zhang、Chengxing Jia、Yi-Chen Li、Lei Yuan、Yang Yu 和 Zongzhang Zhang。从多任务离线数据中发现可泛化的多智能体协调技能。发表于国际学习表征会议，2023年。[80] Kaiqing Zhang、Zhuoran Yang、Han Liu、Tong Zhang 和 Tamer Basar。网络化智能体去中心化批量多智能体强化学习的有限样本分析。*IEEE*自动控制汇刊，66(12):5925–5940，2021年。[81] Sai Qian Zhang、Qi Zhang 和 Jieyu Lin。通过基于方差的控制实现多智能体强化学习中的高效通信。发表于神经信息处理系统进展，第3230–3239页，2019年。[82] Sai Qian Zhang、Qi Zhang 和 Jieyu Lin。通过时间消息控制实现简洁且鲁棒的多智能体通信。发表于神经信息处理系统进展，第17271–17282页，2020年。[83] Tianhao Zhang、Yueheng Li、Chen Wang、Guangming Xie 和 Zongqing Lu。FOP：最大熵多智能体强化学习最优联合策略的因子分解。发表于国际机器学习会议，第12491–12500页，2021年。[84] Changxi Zhu、Mehdi Dastani 和 Shihan Wang。带通信的多智能体强化学习综述。*arXiv*预印本 *arXiv:2203.08975*，2022年。

所涉及的基准测试与算法详述

如图所示，我们介绍四类测试环境 4 和表 4，包括走廊 [67],基于等级的觅食（LBF） [48], 交通路口（TJ） [6], 以及两个名为 1o2r_对_4随机和 1o10b_对_1随机的地图，这些地图来自需要通信的星际争霸多智能体挑战（SMAC） [67]。

表4：各环境的属性。

环境 \ 属性	狭窄	无向	多任务	稀疏奖励	次优	真实	部分可观测性
走廊：4x6x10	✓	✓		✓			✓
走廊：3x5-4x6x10	✓	✓		✓			✓
LBF：11x11-6p-4f-s1	✓		✓		✓		✓
LBF：20x20-10p-6f-s1	✓		✓		✓		✓
SMAC：1o2r _vs_4r	✓				✓		✓
SMAC：1o10b _vs_1r	✓				✓		✓
TJ：容易	✓				✓		✓
TJ：中等	✓				✓		✓

走廊：我们设计了两种走廊环境实例。在第一个实例中，我们设置了长度分别为 4、6 和 10 的三条走廊。这意味着我们让三个智能体a、b、c 分别随机初始化于状态a_1 至a_4, 、b_1 至b_6, 以及c_1 至c_{10}, ，并要求它们同时到达状态g 。在第二个实例中，我们将 5 个智能体分成两个群组。第一组的走廊长度为 3 和 5，第二组的走廊长度为 4、6 和 10。如果其中一个群组同时到达目标g ，则给予 1 的奖励。但是，如果两个群组同时到达目标，则给予−0.5 的惩罚。

基于等级的觅食（LBF）：我们使用原始环境的一个变体版本 [48], ，其中我们将状态定义为一种能够表示真实全局状态的数据结构，而不是直接拼接所有智能体的观测。在此基础上，我们使用了配置不同的两个环境实例，其中一个是包含 6 个智能体和 4 个食物的 11×11 网格世界，另一个是包含 10 个智能体和 6 个食物的 20×20 网格世界。在这两个实例中，智能体的观测都是其周围的 3×3 视野。

交通路口（TJ）：我们使用交通路口环境的中等 和困难 版本。中等 版本的智能体数量上限为 10，道路尺寸为 14；而困难 版本的智能体数量上限为 20，道路尺寸为 18。在这两个实例中，智能体的视野均被限制为 0，这意味着每个智能体只能观察其周围的 1×1 视野。

星际争霸多智能体挑战（SMAC）：我们在 SMAC 中使用了两张名为 1o2r_vs_4r 和 1o10b_vs_1r 的地图，这两张地图在 NDQ [67] 中有所介绍。在 1o2r_vs_4r 中，监督者发现了 4 个死神，而友军单位 2 个蟑螂需要接近敌人并将其消灭。类似地，1o10b_vs_1r 是一张布满悬崖的地图，监督者会发现一个蟑螂，而随机生成的友军单位 10 个爆虫需要接近并消灭敌人。

表 5：本文使用的各种算法对比。

Name	通信类型	信息处理位置（发送方/接收方）	适用场景
MASIA（我们的方法）	Full	接收方	无限制
Full-Comm	Full	No	没有冗余信息
NDQ	Full	发送方	价值函数几乎可以分解
TarMAC+QMIX	Full	接收方	带有相对重要性的消息
TMC	时间部分	发送方和接收方	带有传输损失的消息
QMIX	No	No	完全观测或容易协作

我们的工作涉及多种算法，包括 Full-Comm、NDQ [67], TarMAC [6], TMC [82],和 QMIX [52]。在我们的实验中，所有这些方法都遵循分散式部分可观测马尔可夫决策过程的设定，这意味着每个智能体在每个时间步只能获取自身的部分观测。一些算法

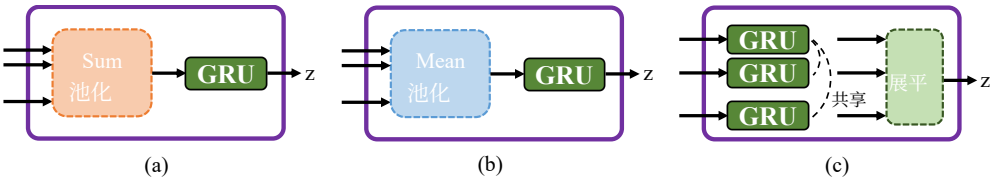


图 11：集成网络的三种实现版本。(c) 中的虚线表示这些门控循环单元网络共享相同的参数。

（包括 Full-Comm、NDQ、TarMAC 和 TMC）通过通信，让每个智能体能够基于更丰富的信息做出决策；而 QMIX 不进行通信，始终让智能体根据本地观测（或观测历史）做出决策。在表5中，我们从不同维度对这些基线方法与我们的方法进行了比较。

B 实现细节

B.1 网络架构和超参数

集成网络。 我们方法中的信息聚合编码器（IAE）由自注意力网络和集成网络组成。下面我们将详细介绍集成网络。如图 11 所示，我们介绍了 3 种不同的集成网络。

在这三种集成网络中，（a）和（b）分别对自注意力网络的输出向量执行均值 和求和 池化操作；而在第三种版本（c）中，我们让自注意力网络的输出向量分别通过一个共享的 GRU 网络，最后将其展平，得到聚合表示。在（a）和（b）中，无论如何对智能体进行排列，我们始终会得到相同的聚合表示。在（c）中，对智能体进行排列只会影响聚合表示中某些维度的顺序，而不会像某些传统设计（例如采用多层感知器的网络）那样得到完全不同的表示。在实践中，我们发现设计（c）在这三种集成网络中取得了最佳性能，并且本文展示的所有实验结果均基于这一实现。

超参数。 我们的 MASIA 实现基于 EPyMARL[3] [53]，使用星际争霸 2.4.6.2.69232，并采用其默认超参数设置。例如，我们对每种方法都应用默认的 ϵ-贪心动作选择算法，这意味着 ϵ 会在 50K 个时间步内从 1 衰减至 0.05。我们方法中引入的额外参数的选择列于表 6 中。除消融实验外，本文展示的所有实验均使用这组参数。

B.2 训练和测试的总体流程

首先，为了直观展示我们的工作，我们先介绍在我们的方法（MASIA）中，智能体在分散式执行与集中式训练期间的行为差异。主要区别在于，在训练阶段，我们会使用状态信息来帮助计算自编码器损失，并估计 Q_{tot}（如果Q-网络中存在混合网络，例如 MASIA+QMIX）。此外，多步预测损失也仅在训练阶段进行计算和优化。在分散式执行期间，状态解码器、潜在模型以及可能存在的混合网络都会被舍弃。我们保留 IAE、聚焦网络和个体 Q 网络，以确保分散式执行过程。我们设计的表示损失项旨在充分训练编码器网络。

为说明训练过程，MASIA 的整体训练流程如算法 1所示。各行5∼16 表示完整的训练过程，其中我们采用离策略学习算法并迭代地

[3]https://github.com/uoe-agents/epymarl

算法1 训练框架

1：初始化回放缓冲区 $\mathcal{D}$。2：使用随机参数 θ 初始化信息聚合编码器，使用随机参数 η 初始化状态预测解码器。3：使用随机参数 ϕ 初始化Q 网络，使用随机参数 ψ 初始化潜在模型。4：初始化目标编码器的参数 $\theta^- = \theta$，以及目标Q 网络的参数 $\phi^- = \phi$。5：**对于**回合 $= 1$ 到M，**执行**6：在环境中使用 ϵ-贪心策略展开一条轨迹 $\boldsymbol{\tau}$。7：将轨迹 $\boldsymbol{\tau}$ 存储到 $\mathcal{D}$ 中。8：**如果** $|\mathcal{D}|$ 大于批次大小m，**则**9：从 $\mathcal{D}$ 中采样一个包含m 条轨迹的小批量 $\mathcal{B}$。10：计算编码器-解码器损失：

$$\mathcal{L}_{ae}(\theta, \eta) = \sum_{\text{traj} \in \mathcal{B}} \sum_{t=1}^{T} \| g_\eta(z^t) - s^t \|_2^2,\ z^t = f_\theta(\boldsymbol{o}^t).$$

11：计算潜在模型损失：

$$\begin{aligned}
&\mathcal{L}_m(\theta, \psi) = \sum_{\text{traj} \in \mathcal{B}} \sum_{t=1}^{T-K} \sum_{k=1}^{K} \| \hat{z}^{t+k} - z^{t+k} \|_2^2, \\
&\hat{z}^{k+1} = h_\psi(z^t, \boldsymbol{a}^t), \\
&\hat{z}^{t+k} = h_\psi(\hat{z}^{t+k-1}, \boldsymbol{a}^{t+k-1}), k = 2, \ldots, K, \\
&z^{t+k} = f_{\theta^-}(\boldsymbol{o}^{t+k}), k = 0, \ldots, K,
\end{aligned}$$

12：计算强化学习损失：

$$\begin{aligned}
\mathcal{L}_{rl}(\theta, \phi) = \sum_{\text{traj} \in \mathcal{B}} \sum_{t=1}^{T-1} \Big(& r + \\
\gamma \max_{\boldsymbol{a}'} Q_{\text{tot}}(\boldsymbol{\tau}^{t+1}, \boldsymbol{a}'; \theta^-, \phi^-) - Q_{\text{tot}}(\boldsymbol{\tau}^t, \boldsymbol{a}^t; \theta, \phi) \Big)^2 .
\end{aligned}$$

13: 通过最小化 θ，η，ϕ，ψ 来更新 $\mathcal{L}_{rl}(\theta,\ \phi) + \lambda_1 \mathcal{L}_{ae}(\theta,\ \eta) + \lambda_2 \mathcal{L}_m(\theta,\ \psi)$。
14: **结束条件**15: 使用指数移动平均方法更新目标网络参数 θ^-，ϕ^-。16: **结束循环**

表 6：实验中的超参数。

name	值
自注意力模块中查询和键的隐藏维度	16
自注意力模块的输出维度	32
λ_3 （是否预热表征学习）	1
λ_2 （潜在模型学习损失的系数）	1
λ_1 （编码器-解码器学习损失的系数）	1
潜在模型中的动作嵌入维度	8
聚合表示中每个智能体的维度	8
与z拼接之前的观测嵌入维度	32
是否预测下一状态的残差	True
潜在模型中的预测长度 K	2
潜在模型中隐藏状态的隐藏维度	64

更新模型参数。具体而言，我们分别在第 10、11 和 12 行计算编码器-解码器损失、潜在模型损失和时序差分损失，并在第 13 行统一进行参数更新。此外，MASIA 的执行流程如算法 2所示。在执行阶段，智能体首先相互广播各自的观测，然后每个智能体通过应用训练期间学习到的聚焦网络和个体 Q 网络来计算自身的动作，该过程在第 4 和第 5 行中描述。

算法2 执行流程

要求： 参数为 θ 的信息聚合编码器、参数为 ϕ_i 的独立 Q 网络
每个 Q_i、每个智能体对应的参数为 ω_i 的聚焦网络，以及智能体数量 n。

1: **对于** 步骤 $= 0$ 到回合_上限**执行**
2: 每个智能体在时间步 t广播其观测信息 o_i^t，然后每个智能体输入收集到的将观测$\boldsymbol{o}^t = \{o_i^t\}_n$ 输入信息聚合网络，得到$z^t = f_\theta(\boldsymbol{o}^t)$。
3: **对于**$i = 1$至 n，**执行**
4: 智能体i使用聚焦网络计算$w_i^t = F^{w_i}(o_i^t)$，并获得提取的信息$\bar{z}_i^t = w_i^t \cdot z^t$.
5: 将o_i^t和 $\bar{z}_i^t$输入个体Q 网络以计算$Q_i(\tau_i^t, \cdot)$，智能体i 获得动作$a_i^t =$最大值参数$_a\, Q_i(\tau_i^t, a)$.
6: **结束循环**
7: 智能体系统通过执行动作 $\{a_i^t\}_n$ 与环境交互。
8: **结束循环**

B.3 实验细节

我们的实验在一台配备 4 块 NVIDIA GTX 3090 GPU 的台式机上进行。对于论文中的所有性能曲线，我们每隔 M 个时间步暂停训练，并使用去中心化贪心动作选择进行 N 个回合的评估。走廊中的 (M,N) 、基于等级的觅食 [67], 交通路口 [48], 和 SMAC [6], 分别对应(10K, 100)、(50K, 100)、(10K, 40)和(50K, 100)。对于所有任务，我们通过在N个测试回合中智能体在时间限制内赢得游戏的回合占比，评估测试胜率。

C 更多关于离线学习的结果

C.1 关于离线数据集的详细信息

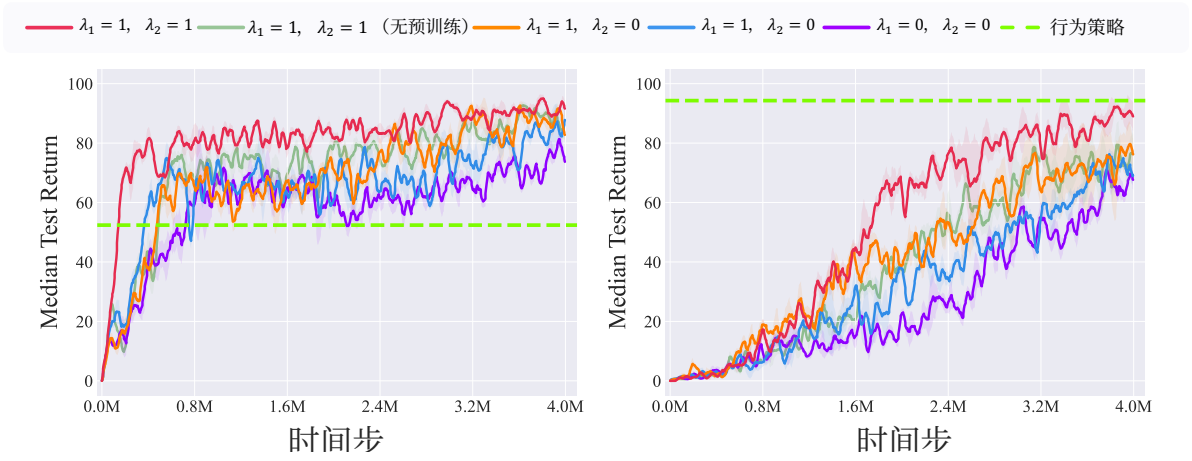
由于在现实世界中与环境交互的成本高昂且受到诸多限制，离线学习被视为推动强化学习发展的一个有前景的方向。此外，许多现实世界的问题都可以建模为多智能体系统，尤其是多智能体通信系统，因此亟需推进离线通信学习的相关研究。然而，在多智能体通信强化学习领域，目前仍缺乏良好的离线学习评估标准，这在很大程度上阻碍了该领域的发展。鉴于此，我们构建了一个离线数据集，其中包含多个多智能体通信环境和不同的数据集设置，以支持离线多智能体通信学习研究。

具体而言，我们的数据集包含四个环境，每个环境包括两个场景，共计八个场景。其中包含的环境与在线实验中的任务环境一致，如图 4所示。正如我们在实验部分（第 4节）中所分析的，这些环境代表了具有不同特征的通信场景。为了进一步了解离线设置下的这四个环境，我们参考 D4RL [17]提出的若干数据集属性，在表 4中额外提供了这些环境的描述。在这些属性中，所有环境都具有**狭窄**、非**真实**以及**部分可观测性**。后两点是因为数据均在模拟器中采集，不包含真实的人类数据，并且由于通信设置的限制，智能体的视野范围受到限制。Hallway环境的数据是**无向**且具有**稀疏奖励**的，因为智能体只有在共同到达目标时才能获得奖励，而且该环境中的数据可以进行拼接。此外，这是唯一一个专家策略能够生成最优轨迹的环境。除此之外，LBF环境的数据具有**多任务**特性，因为目标水果的位置会在每个回合开始时随机生成，因此可以将其视为包含多个导航任务的数据。

此外，考虑到真实数据可能具有不同的质量，我们设置了多个数据集设置，以涵盖不同的数据分布。在第 4.2 节中，我们介绍了这三种不同的数据集设置以及它们的构建方式。事实上，由于数据收集方法不同，这些数据集的数据质量也有所差异。例如，**专家数据集** 中的数据通常是高回报轨迹。表 7 列出了这些数据集的更多详细信息，其中展示了不同数据集的回报分布。此外，为了更直观地呈现数据分布，我们在图 13-20中绘制了更多信息。以 LBF 任务场景 11x11-6p-4f-s1 为例，**专家**、**噪声**和**回放（良好）数据集** 中大多数轨迹的回报值都集中在 1.0，这也是该任务的最优回报值。而**回放（较差）**和**回放（中等）数据集**中没有轨迹达到 1.0 的回报值，且大多数轨迹的回报值都集中在低得多的区间。

C.2 更多离线学习消融实验结果

在第4.2.2节中，我们在 Hallway：3x5-4x6x10 的**回放（中等）**和**回放（较差）**数据集上进行了离线消融研究，以验证两种无监督表示损失以及离线预训练做法的有效性。这里，我们进一步提供 TJ：中等任务上的消融结果。如图12所示，TJ：中等任务上的结论与正文中的结论相似：同时采用两种无监督表示损失和离线预训练的 MASIA，相比其他消融设置，学习速度更快，收敛性能更好。



（a）回放（中等）　　（b）专家

图12：TJ：中等任务上的离线学习消融实验。

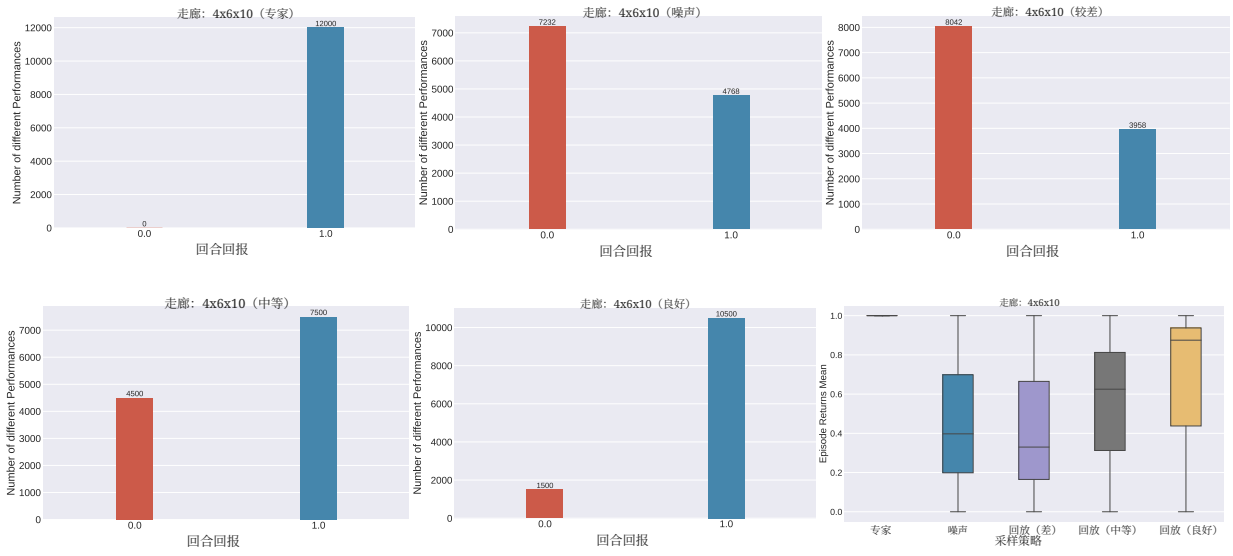


图13：走廊：4x6x10 离线数据集分布。

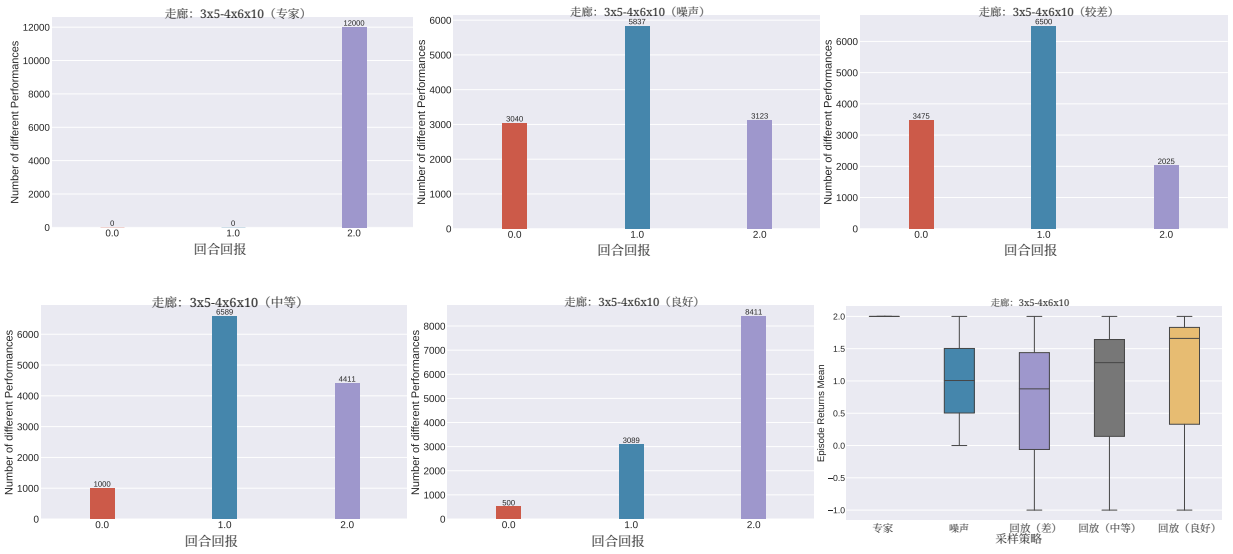


图14：Hallway：3x5-4x6x10 离线数据集分布。

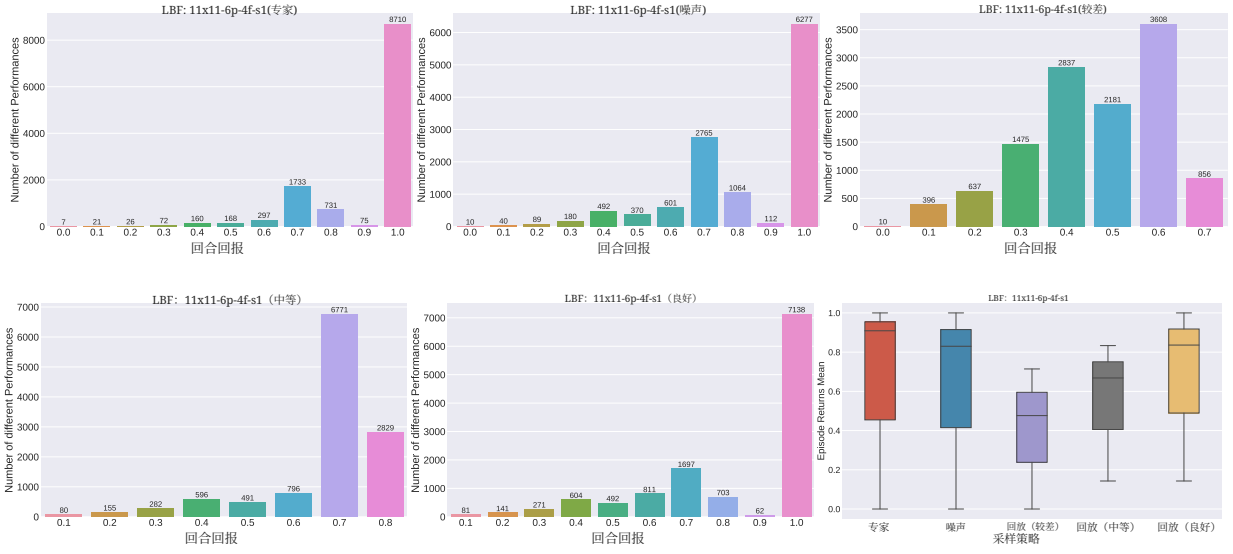


图15：LBF：11x11-6p-4f-s1 离线数据集分布。

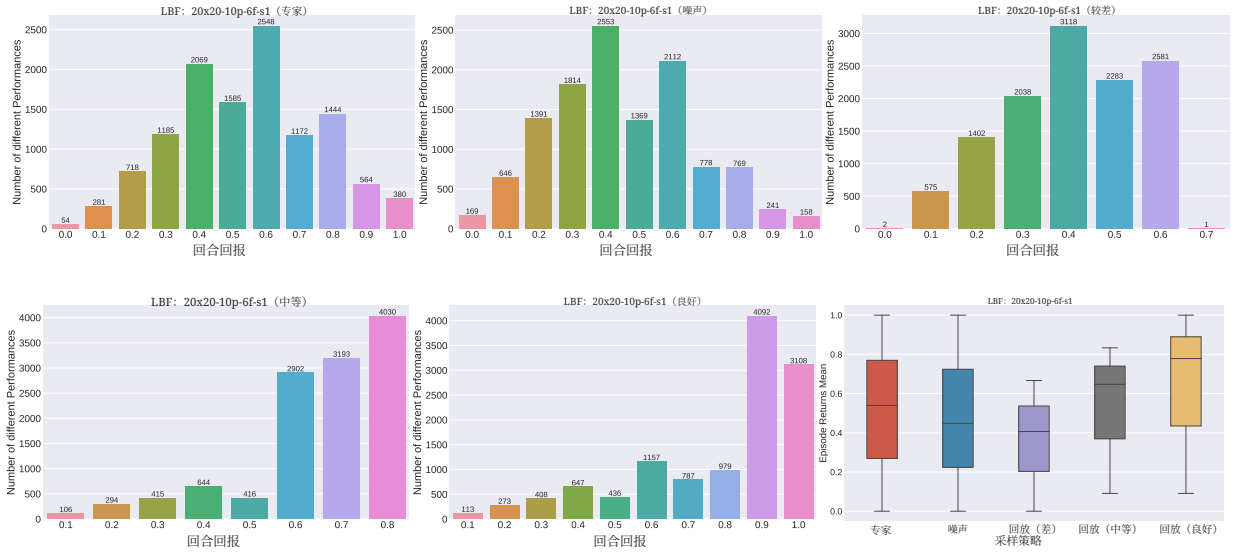


图16：LBF：20x20-10p-6f-s1 离线数据集分布。

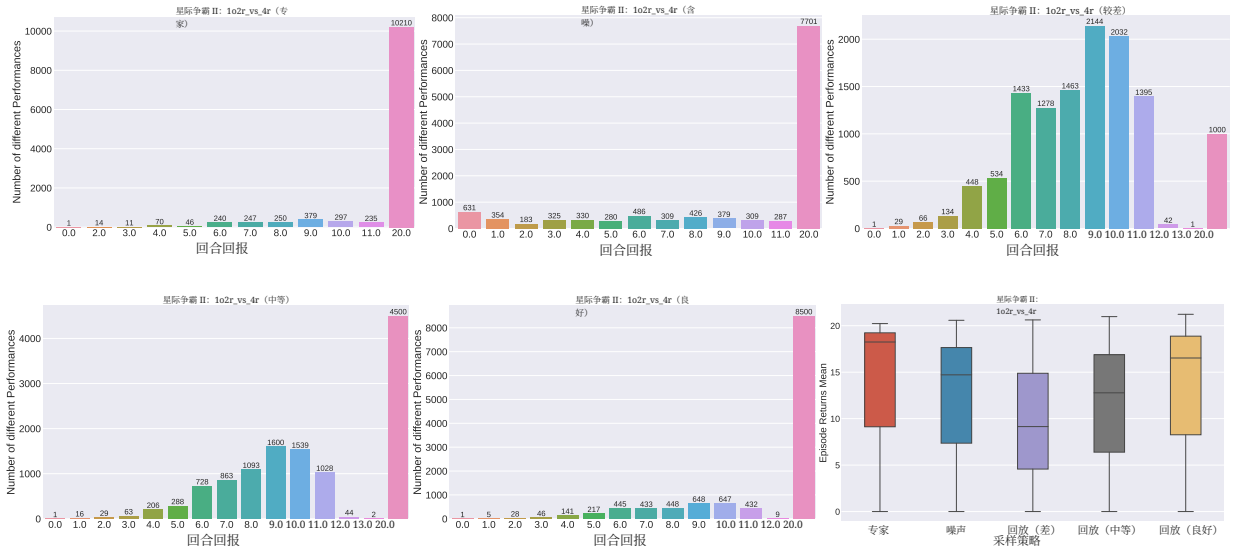


图17：SMAC：1o2r_vs_4r 离线数据集分布。

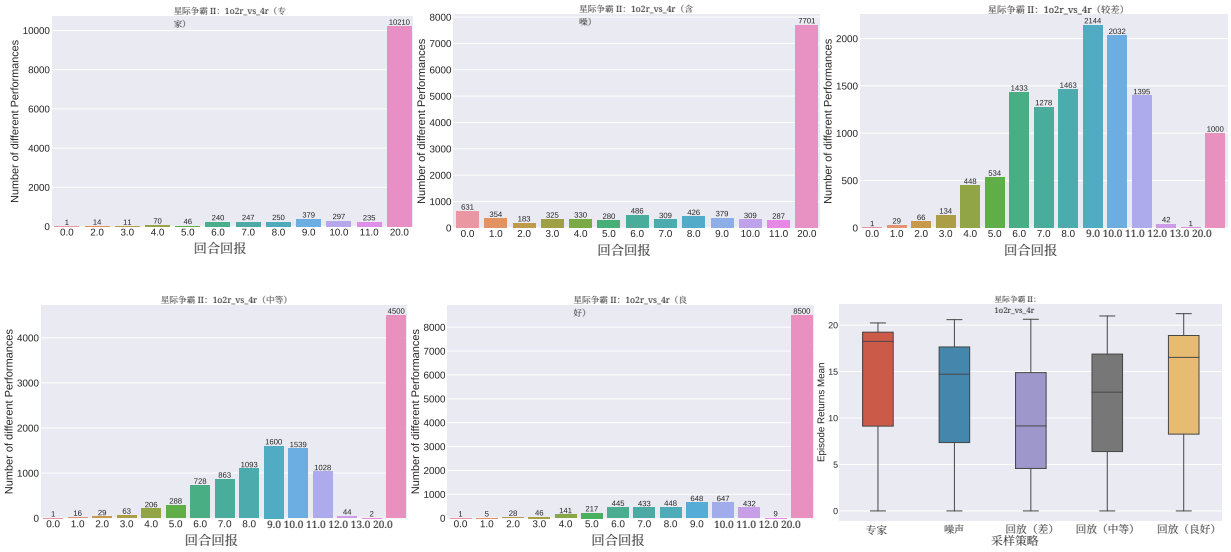


图18：SMAC：1o10b_vs_1r 离线数据集分布。

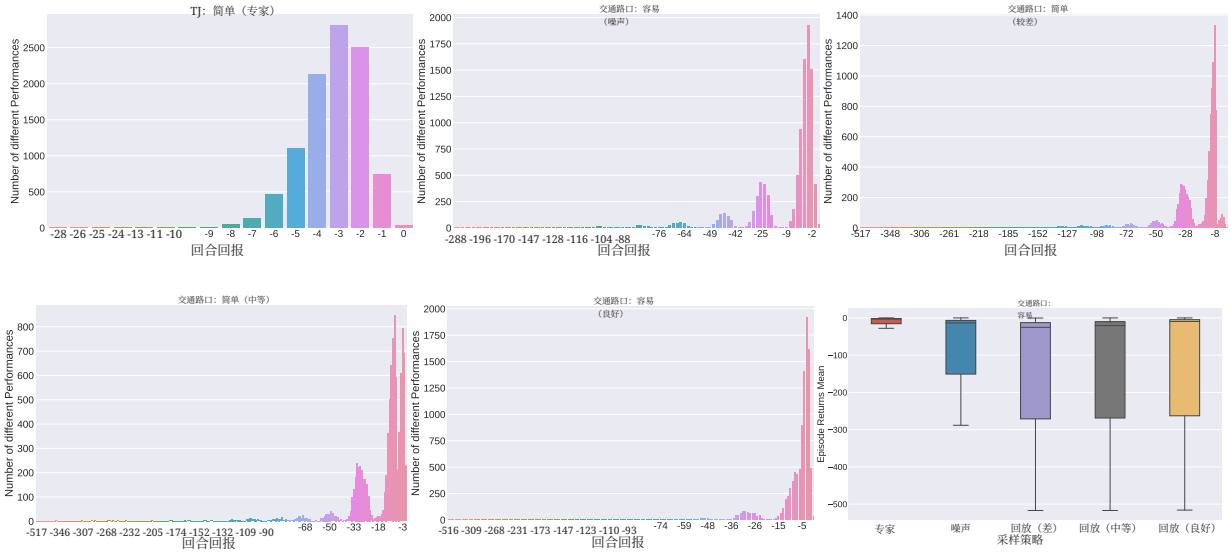


图19：交通路口：容易的离线数据集分布。

表7：不同离线数据集的属性。

环境场景	数据集	行为策略 测试成功率	样本数	回报分布
走廊：4x6x10	专家	100.00 ± 0.00	12000	1.00
	噪声	100.00 ± 0.00		0.40
	回放（较差）	32.90 ± 47.01		0.33
	回放（中等）	62.50 ± 48.41		0.62
	回放（良好）	87.50 ± 33.07		0.88
走廊：3x5-4x6x10	专家	99.80 ± 0.25	12000	2.00
	噪声	99.80 ± 0.25		1.01
	回放（较差）	43.90 ± 33.31		0.88
	回放（中等）	64.20 ± 30.41		1.28
	回放（良好）	82.90 ± 27.74		1.66
LBF：11x11-6p-4f-s1	专家	90.20 ± 2.59	12000	0.91
	噪声	90.20 ± 2.59		0.83
	回放（较差）	46.60 ± 14.93		0.48
	回放（中等）	67.40 ± 13.36		0.67
	回放（良好）	83.60 ± 22.71		0.84
20x20-10p-6f-s1	专家	53.80 ± 2.34	12000	0.54
	噪声	53.80 ± 2.34		0.45
	回放（较差）	38.70 ± 14.55		0.41
	回放（中等）	45.50 ± 15.75		0.65
	回放（良好）	49.10 ± 22.67		0.78
SMAC：1o2r 对 4r_ _	专家	81.50 ± 5.72	12000	18.25
	噪声	81.50 ± 5.72		14.71
	回放（较差）	45.70 ± 19.25		9.14
	回放（中等）	63.80 ± 29.03		12.78
	回放（良好）	76.60 ± 27.69		16.53
SMAC：1o10b vs 1r_ _	专家	82.80 ± 2.66	12000	18.52
	噪声	82.80 ± 2.66		12.27
	回放（较差）	48.20 ± 31.48		9.62
	回放（中等）	67.30 ± 34.65		14.63
	回放（良好）	78.70 ± 30.43		16.80
交通路口：容易	专家	98.70 ± 1.25	10000	-3.28
	噪声	98.70 ± 1.25		-13.48
	回放（较差）	45.60 ± 17.28		-25.27
	回放（中等）	57.80 ± 16.57		-20.61
	回放（良好）	73.90 ± 16.98		-9.58
TJ：中等	专家	94.30 ± 1.08	36000	-7.08
	噪声	94.30 ± 1.08		-15.53
	回放（较差）	42.70 ± 14.86		-27.77
	回放（中等）	52.40 ± 21.79		-23.32
	回放（良好）	67.10 ± 18.16		-17.17

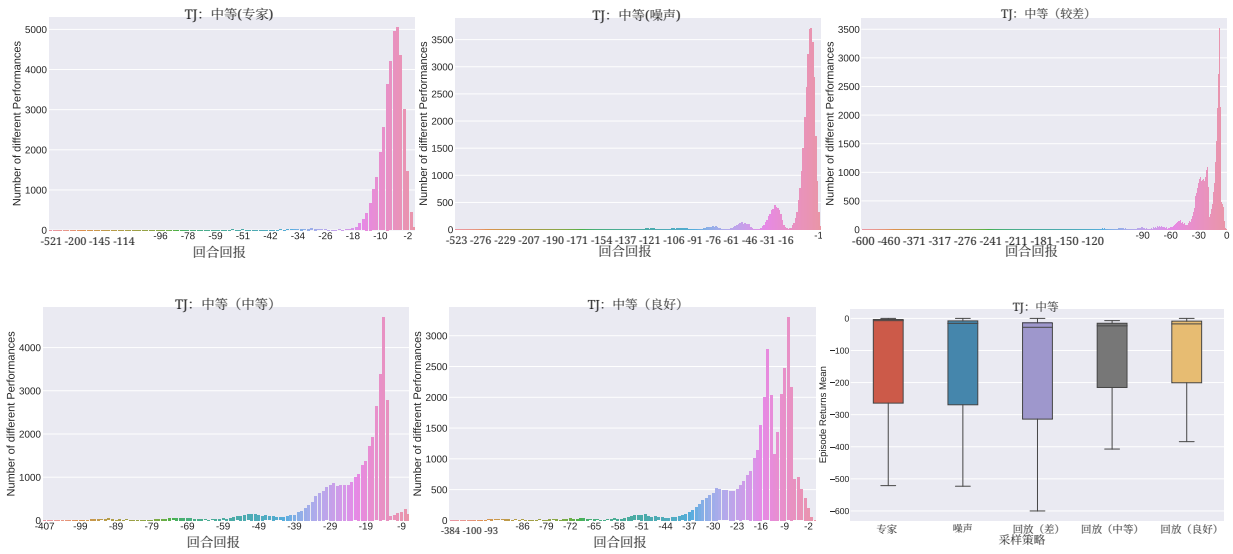


图 20：TJ：中等离线数据集分布。