arXiv:2301.05223v1 [cs.RO] 12 Jan 2023

NOPA：面向构建具有社交智能的家庭助手的神经引导在线概率辅助

Xavier Puig*、Tianmin Shu*、Joshua B. Tenenbaum、Antonio Torralba

摘要——本文研究如何构建社会智能机器人，帮助人们处理家中事务。具体而言，我们聚焦于在线目标推断辅助，即机器人必须同时推断人类的目标以及如何帮助他们实现这些目标。以往的辅助方法要么缺乏适应性，无法根据目标不确定性调整协助策略（即何时以及如何提供帮助）；要么缺乏可扩展性，无法在庞大的目标空间中进行快速推断。我们的 NOPA（神经引导的在线概率辅助）方法同时应对这两项挑战。NOPA 包含：(1) 一个在线目标推断模块，将神经目标候选与逆向规划和粒子滤波相结合，从而在存在不确定性的情况下实现稳健推断；(2) 一个辅助规划器，用于发现值得协助完成的子目标，并将目标推断中的不确定性纳入考量。我们在一项全新的具身人工智能辅助挑战中，将 NOPA 与多种基线方法进行比较：Online Watch-And-Help。在该挑战中，辅助智能体需要同时观察主智能体的动作、推断其目标，并在逼真的虚拟家庭环境中协助其更快地完成常见家务任务。实验表明，我们的辅助智能体能够稳健地更新目标推断，并根据不确定性程度的变化调整协助计划。[1]

I. 引言

人们对研发社会智能机器人日益感兴趣，希望它们能够在现实世界中安全、高效地与人类协作。以往关于机器人协助的研究在一些场景中取得了成功：机器人预先获知人类的真实目标，或者只需在状态空间较小的简单环境中帮助人类。然而，要打造能够在更自然的环境中帮助人类完成所有日常生活活动的机器人助手，仍然极具挑战性。例如，在家庭环境中，人类目标的范围十分广泛，而且在任意时刻，一个人的目标通常都无法被确切获知。

我们的目标是打造能够在复杂家庭环境中帮助人们完成各种任务的机器人助手。机器人助手必须能够根据过去的观测在线推断人类的真实目标，并规划如何帮助人类，同时不造成干扰
了解他们，并随着任务的推进，同时更新目标推断和协助策略，以适应他们的行为（如图1所示）。迄今为止，机器人一直难以具备这种能力，原因在于存在两个主要的技术障碍。一方面，由于状态空间、动作空间和目标空间庞大，在现实环境中进行在线目标推断极其困难；另一方面，不准确或模糊的目标推断

* 共同贡献。所有作者均来自麻省理工学院。{xpuig, tshu, jbt, torralba}@mit.edu[1]代码可在 https://github.com/xavierpuigf/online_watch_and_help 获取。项目网站：https://www.tshu.io/online_watch_and_help。

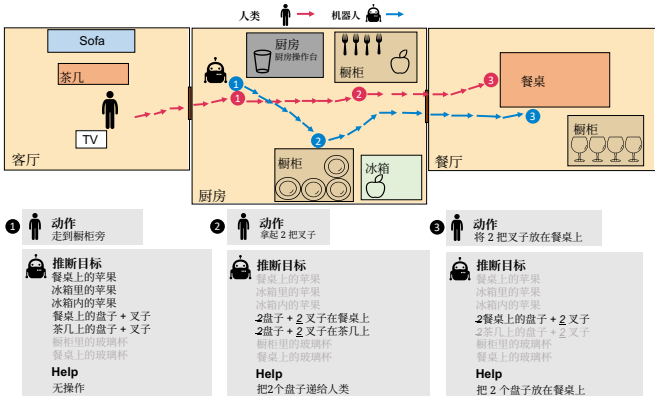


图1：成功在线协助示意图。机器人起初不了解人类的目标，因此会选择观察。随着观察到更多人类动作，它对目标推断越来越有把握，因此会动态调整辅助子目标。例如，在这幅图中，机器人首先看到人类走向橱柜，于是推断目标涉及橱柜内的物品。人类拿起2把叉子后，机器人推断目标是将2套餐具（盘子和叉子）摆放在餐桌或茶几上，但不确定目标位置。因此，它把2个盘子递给人类，而不是随意猜测位置。

常常会导致无效甚至适得其反的帮助尝试，尤其是在系统未意识到自身不确定性的情况下。

为应对这些挑战，我们提出了一种新颖的在线协助方法 NOPA（神经引导在线概率辅助）。如图 2**a**所示，NOPA 由两个主要部分组成：（1）神经网络引导的在线目标推断模块；（2）不确定性感知助人规划器。神经网络引导的在线目标推断模块首先利用神经网络自下而上地产生目标候选，随后通过粒子滤波和逆向规划，维护一组与观测到的动作相一致的目标及未来轨迹预测，从而确保推断既快速又稳健。助人规划器根据最新预测及其确定性，首先确定最值得协助的子目标，然后利用符号规划器规划相应的协助行动。由此生成的辅助计划能够适应预测中的各种不确定性。例如，当某个目标对象存在多个可能的目标位置时，机器人助手会将该对象递给人类代理，而不是冒着放错位置的风险。

为进行评估，我们提出了一项新的具身人工智能助

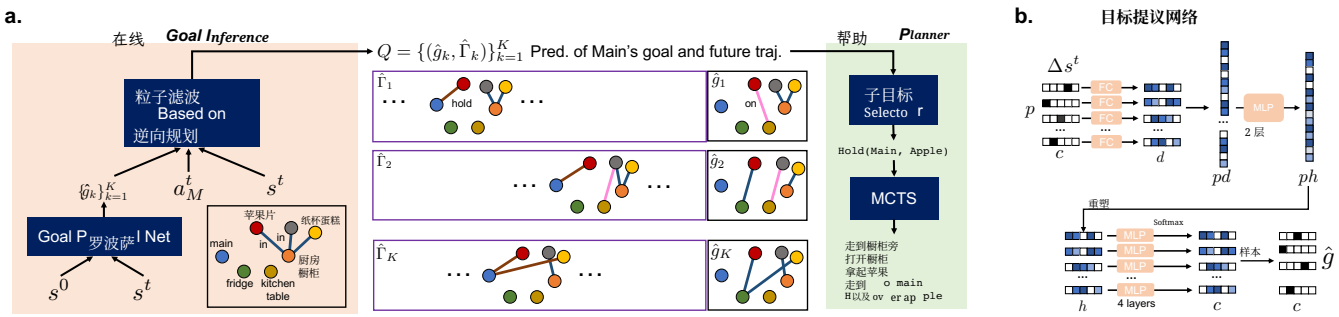


图 2：（**a**）我们方法的概览，该方法由在线目标推断模块和辅助规划器组成。我们使用场景图表示状态和目标。 s^t 是时间 t的状态； a^t_{M} 是主智能体在时间 t执行的动作； $\hat{}g_k$ 是第 k个目标提议； Γ_k 是对应于$\hat{}g_k$的主智能体未来轨迹预测。（**b**）目标提议网络。Δs^t 是一个矩阵，用于编码状态 s^t 和 s^0中的谓词计数差异，其中每一行都是一个独热向量，表示特定谓词计数的变化。 $p = 136$ 是谓词类型的总数， $c = 9$ 是计数的最大值， $d = 100$ 和 $h = 128$ 是中间层的维度。

力挑战：在线观察与协助（O-WAH）。该挑战基于近期推出的多智能体虚拟家庭环境 VirtualHome-Social。与大多数现有挑战（例如 [1], [2]）不同，在 O-WAH 中，辅助智能体需要在线推断主智能体的目标，同时尽可能高效地帮助其实现推断出的目标。我们在一系列家务任务中评估了采用 NOPA 构建的智能体及若干基线方法，为由人类玩家或基于规划器的智能体控制的主智能体提供帮助。实验结果表明，NOPA 的表现显著优于所有基线方法。我们还观察到，NOPA 会根据新的观测进行调整，并由此展现出智能的协助策略。

总之，我们的主要贡献包括：（1）一种新的在线协助方法，用于在复杂环境中构建具有社交智能的家庭助手；（2）一项名为 Online Watch-And-Help 的具身人工智能辅助挑战，可作为训练和测试具身智能体的试验平台，使其能够在逼真的虚拟家庭环境中执行在线目标推断和提供帮助。

II. 相关工作

在线目标推断。 在线目标推断方法通常分为两类：（1）前馈预测，直接将观测到的过去轨迹映射到可能的目标，通常由目标预测网络 [3]–[11], 实现；（2）生成式方法，例如逆向规划 [12]–[19]，通过比较给定目标假设所生成的计划与观测到的动作来进行推断。前馈方法速度快，在简单任务中（如行人的目的地预测 [5]）经过大量数据训练后往往表现良好。然而，在不熟悉的场景中，逆向规划方法通常优于前馈预测，因为它们能够想象各种条件下的理性行为。逆向规划方法的局限之一是，当目标空间较大时，推断速度可能较慢，并且依赖人工设计的启发式方法来加速推断 [17], [18]。我们的工作整合了这两类方法，以兼顾速度和鲁棒性。

具有未知目标的具身人工智能协助。具身人工智能协助领域已有大量研究积累。许多现有研究假设人类与 AI/机器人伙伴共享一个已知的共同目标 [1], [20]–[27]。然而，在现实世界中，机器人往往需要即时推断人类的目标。已有研究探讨了如何帮助实现推断目标 [2], [28]–[32] ，表明准确的目标推断能够提升具身人工智能助手的客观表现和感知表现。然而，当目标推断存在不确定性时，帮助实现推断目标往往会导致适得其反的行为，例如取消已经完成的目标 [2]。为此，一些先前研究在简单环境中设计了应对不确定目标推断的规划器 [33]。近期一项研究提出了一种通过赋能实现的目标无关协助框架 [34], ，旨在改变状态，从而最大限度地提升智能体达成尽可能多目标的能力，而不论其真实意图为何。尽管该方法在某些领域取得了成功，但在不了解人类目标的情况下，于现实环境中协助人类往往会导致适得其反的行为。我们的研究探讨了如何设计一种不确定性感知规划器，在复杂环境中智能地调整帮助行为，使其范围从目标无关策略延伸至目标特定计划。

具身人工智能辅助基准。已有一些基准旨在评估人工智能智能体与人类队友协作的能力 [1], [25]。然而，其中大多数基准关注简单的游戏环境，并假设人工智能智能体和人类代理事先拥有共同目标。基于一个逼真的虚拟家庭环境， [2]提出了 Watch-And-Help 挑战：辅助智能体必须根据预先录制的示范推断主智能体的目标，然后帮助主智能体实现同一目标。我们将这一挑战扩展到在线辅助场景，此时不再提供示范。因此，辅助智能体必须留意主智能体的动作，并在努力实现推断目标的同时不断更新推断及其不确定性，这更接近机器人在真实家庭中应有的表现。这类基准补充了在实验室环境中开展的传统机器人辅助研究 [20], [21],，并提供了一种可复现且

算法1NOPA

输入： $\Gamma_{\mathrm{M}}^{0}=\{(s^{0},\ a_{\mathrm{M}}^{0})\}$, s^{t}, K, $T_{\max}$, T_{prop}, q, w_r, w_c, w_m, $L_{\max}$
$t\leftarrow 1$, $l\leftarrow 0$
$Q\leftarrow\emptyset$
重复
  $Q, l\leftarrow$ **GoalInf**$(t, Q, \Gamma^{t-1}{}_{\mathrm{M}}\ s^{t}, q, K, l, T_{\text{prop}})$
  $\Gamma_{\mathrm{H}}^{t}\leftarrow$ **HelpPlanner**$(Q,\ s^{0},\ s^{t},\ w_r,\ w_c,\ w_m,\ L_{\max})$
  执行辅助计划中的第一个动作 a_{H}^{t}
  从环境中观察 $a_{\mathrm{M}}^{t},\ s^{t+1}$
  $t\leftarrow t+1$
直到 $t=T_{\max}$ 或尚未达到真正的目标

算法 2 GoalInf

输入： t, Q, $\Gamma^{t-1}\mathrm{M}=\{(s^{\tau},a_{\mathrm{M}}^{\tau})\}_{\tau=0}^{t-1}$, s^{t}, q, K, l, T_{prop}
输出： 已更新的提议 Q' 以及距上次提议的步数 l'
$Q'\leftarrow\emptyset$
如果 $Q\neq\emptyset$ 且 $l<T_{\text{prop}}$ **则**
  对于 $k=1,\ \cdots,\ |Q|$ **执行**
    如果 a^{t-1} M 是计划 $\hat{}\,\Gamma_k$ 的一部分**则**
      $Q'\leftarrow Q'\cup\{(\hat{g}_k,\ \hat{}\,\Gamma_k)\}$
    结束如果
  结束循环
结束 if
如果 $Q'=\emptyset$ **则**
  对于 $k=1,\cdots,K$ **do**
    $\hat{g}_k\sim q(g|s^{0},s^{t})$ // 对目标提案进行采样
    $\hat{}\,\Gamma_k\leftarrow$ **蒙特卡洛树搜索**$(s^{t},\hat{g}_k,T_{\text{prop}})$ // 对计划进行采样
  $Q'\leftarrow Q'\cup\{(\hat{g}_k,\ \hat{\Gamma}_k)\}$
  结束循环
  $l'\leftarrow 0$
否则
  $l'\leftarrow l+1$
结束条件判断
返回 Q', l'

可扩展的方法，用于比较不同的方法。

III. 方法

A. 问题设定

我们将在线辅助问题定义为混合可观测性马尔可夫决策过程（MOMDP） [35], ，其中辅助智能体需要推断主智能体的目标，并帮助主智能体更快地实现其目标。该问题可形式化为 $\langle\mathcal{S},\mathcal{G},\mathcal{A}_{\mathrm{H}},\mathcal{O},\mathcal{T}_{\mathcal{S}},\mathcal{T}_{\mathcal{G}},Z,R_{\mathrm{H}},\gamma\rangle$。整体状态包含两个部分：世界状态 $s\in\mathcal{S}$，两个智能体都可以完全观测；以及主智能体的目标 $g\in\mathcal{G}$，辅助智能体只能部分观测。 $\mathcal{A}_{\mathrm{H}}$ 是辅助智能体的动作空间。辅助智能体的观测由世界状态和主智能体的动作构成，即 $\mathcal{O}=\mathcal{S}\times\mathcal{A}_{\mathrm{M}}$。 $\mathcal{T}_{\mathcal{S}}(s,g,a_{\mathrm{H}},s')=p(s'|s,g,a_{\mathrm{H}})$ 是世界状态的转移函数，而 $\mathcal{T}_{\mathcal{G}}(s,g,a_{\mathrm{H}},s',g')=p(g'|s,g,a_{\mathrm{H}},s')$ 是目标的转移函数。 $Z(s',g',a_{\mathrm{H}},o)=p(o=(s,a_{\mathrm{M}})|s',g',a_{\mathrm{H}})$ 是观测结果的条件概率函数。在步骤 $t+1$，辅助智能体根据主智能体在 t 时刻之前的历史轨迹推断其目标，即 $\Gamma^{t}=\{(s^{\tau},a_{\mathrm{M}}^{\tau})\}_{\tau=1}^{t}$。辅助智能体的期望奖励函数定义为 $R_{\mathrm{H}}(s,a|\Gamma^{t})=E_{p(g|\Gamma^{t})}[\mathbb{1}(s=g)]-c_{\mathrm{H}}(a)$，其中 $c_{\mathrm{H}}(a)$ 是动作 a 的成本，而 $\mathbf{1}(\cdot)$ 用于判断当前世界状态 s 是否满足目标。 γ 是折扣因子。请注意，在目标未知的辅助问题中，假设两个智能体都能完全观测世界状态是先前研究中的常见做法 [29], [31], [32]。

B. 方法概述

为了解决上述形式化的在线辅助问题，我们提出了 NOPA（神经引导在线概率辅助）。图 **2a** 展示了 NOPA 的概览，包括两个主要组成部分：i）神经引导的在线目标推断；ii）不确定性感知助人规划器。如算法1所示，NOPA 根据观测到的状态和主智能体的动作更新一组粒子。每个粒子对应一个可能的最终目标。常见的辅助框架 [2], [29], [31], [32] 通常只考虑用于辅助的最终目标。然而，当目标推断存在不确定性时，智能助手应寻找有较高把握能够提供帮助的中间子目标。为此，我们还会针对每个粒子预测主智能体的未来轨迹。我们将中间状态和最终目标都表示为场景图中的一组边 [36],[37], $\langle O,E\rangle$，如图 **2a**所示。每个节点 $o\in O$ 代表一个实体（智能体/对象）；每条边 $e\in E$ 对应一个谓词（例如 IN(苹果, 厨房橱柜)），表示两个实体之间的空间关系。这类表示已广泛应用于机器人学和具身人工智能领域 [38]–[41]。给定这些粒子，助人规划器会评估中间状态和最终目标中各边的价值，并选择价值最高的边作为辅助子目标。下面我们将介绍具体细节。

C. 神经引导的在线目标推断

与先前关于在线目标推断的研究不同，本研究中的在线目标推断旨在帮助下游任务，即提供协助。这带来了额外的挑战：(1) 辅助智能体必须估计推断中的不确定性，而不只是预测最可能的目标；(2) 必须确保推断在动态环境中具有鲁棒性；(3) 推断必须高效，使助手能够迅速响应并提供协助。为此，我们提出了一种神经网络引导的在线目标推断算法，如算法 2 所示，该算法结合了逆向规划和神经网络。

如图 **2b** 所示，我们使用目标提议网络（GPN）来学习提议分布 $q(g|s^{0},\ s^{t})$，并从中采样 K 目标候选 $\{\hat{}\,g_k\}_{k=1}^{K}$，这些候选由初始状态 s^{0} 和当前状态 s^{t}确定。每个目标候选都是一组目标谓词。在此，我们仅将初始状态和当前状态作为 GPN 的输入，而不使用一系列过去的状态，这样一来，即便状态变化序列与训练序列可能大不相同，在仅由主智能体执行任务的回合上训练出的 GPN 也能稳健地应用于辅助情境。

我们使用逆向规划来评估 GPN 提出的目标，并拒绝与观测到的主智能体动作不一致的目标。为了在有限理性条件下，根据每个目标 $\hat{}\,g_k$ 对智能体的行为进行建模，我们使用内置规划器来预测主智能体接下来 T_{prop} 步 $\hat{}\,\Gamma_k=\{(\hat{}\,s^{t+\tau_k},\ \hat{a}^{t+\tau_k})\}_{\tau=1}^{T_{\text{prop}}}$ 的未来轨迹。具体而言，理性程度可以通过模拟次数和

算法 3 HelpPlanner

输入： Q, s^0, s^t, w_r, w_c, w_m, $L_{\max}$
输出： 辅助计划 Γ_{H}^t
对于 $e \in \mathcal{E}$，**执行**
  $L_{\mathrm{M}}(e) \leftarrow \infty$
  $p(e) \leftarrow 0$
  $V(e) \leftarrow -\infty$
  $\Gamma_{\mathrm{H}}(e) \leftarrow \emptyset$
  如果 e 出现在初始状态**那么**
    $L_{\mathrm{M}}(e) \leftarrow 0$
    $p(e) \leftarrow 1$
  else
    对于 $k = 1, \cdots, |Q|$ **执行**
      如果 e 出现在未来轨迹中。$\hat{}\ \Gamma_k$ **则**
        令 $\tau_k(e)$ 为 e 出现时的第一步 $\hat{\Gamma}_k$
        $L_{\mathrm{M}}(e) \leftarrow \min(L_{\mathrm{M}}(e)\ \tau_k(e))$
        $p(e) \leftarrow p(e) + 1/|Q|$
      否则，如果 e 出现在预测目标中$\hat{g}_k$ **则**
        $L_{\mathrm{M}}(e) \leftarrow$ 分钟$(L_{\mathrm{M}}(e)\ L_{\max})$
        $p(e) \leftarrow p(e) + 1/|Q|$
      结束如果
    结束循环
  结束如果
  如果 $p(e) > 0$ **那么**
    $\Gamma_{\mathrm{H}}(e) \leftarrow$**蒙特卡洛树搜索**$(s^t, \{e\}, \infty)$ // 助手针对子目标 e的计划
    $L_{\mathrm{H}}(e) \leftarrow |\Gamma_{\mathrm{H}}(e)|$
    按公式 (1) 计算 $V(s^t,\ e)$
  结束 if
结束 for
$e^* \leftarrow \arg\max V(s^t,\ e)$
返回 $\Gamma_{\mathrm{H}}(e^*)$

滚动执行的长度进行调整。随后，我们创建 $K\hat{}$ 个粒子 $Q = \{\hat{}g_k,\ \Gamma_k\}$。每当主智能体执行一个新动作 a_{M}^t 时，我们都会检查该动作是否属于每个粒子的预测计划。如果对于某个粒子 k 而言，该动作未包含在预测计划中，那么这表明相应目标下的理性行为与观测到的动作不一致。因此，该目标很可能是错误的，需要拒绝该粒子。当不再有粒子，或达到预测时域 T_{prop} 时，我们会根据最新状态从 GPN 中重新采样 K 个目标，并创建新的粒子。由于某些粒子可能共享同一个目标，但具有不同的预测计划，因此我们的方法可以考虑达成目标的不同方式。

D. 不确定性感知辅助规划器

在每一步寻找最优辅助计划的代价很高。为了平衡速度和最优性，我们的辅助规划器（算法 3）专注于寻找有价值的子目标，而不是在每一步更新整个计划。它将主智能体的预测计划中最终目标和中间状态里出现的所有边都视为候选辅助子目标。此外，辅助智能体在更新目标推断时，或在主智能体完成相应子目标后，可能会发现自己拿取的对象已不再需要。为了让辅助智能体能够将这些对象放回其初始位置，辅助子目标空间还包括初始状态中的边。对于每条边 e，我们估计主智能体到达该子目标需要多长时间，即 $L_{\mathrm{M}}(e)$。如果这条边出现在粒子中的某条预测轨迹里，我们就可以根据它出现在轨迹中的位置，方便地估计 $L_{\mathrm{M}}(e)$。如果它只出现在最终目标中，我们就使用固定长度 $L_{\max}$，预测主智能体到达该子目标需要多少步。我们还可以使用蒙特卡洛树搜索，为辅助智能体搜索一个计划 $\Gamma_{\mathrm{H}}(e)$，以到达同一子目标。设辅助智能体计划的长度为 $L_{\mathrm{H}}(e)$，则我们将辅助子目标 e 的辅助收益定义为辅助智能体通过到达子目标 e 所能带来的加速，即 $\max(L_{\mathrm{M}}(e) - L_{\mathrm{H}}(e),\ 0)$。为考虑推断中的不确定性，我们通过统计有多少粒子在中间状态或最终目标中包含 e，来估计 e 是否可能是必要的，即 $p(e)$。最后，我们定义一个价值函数，用于为辅助智能体选择最佳子目标：

$$V(s^t, e) = \begin{array}{l} w_r p(e) |L_{\mathrm{M}}(e) - L_{\mathrm{H}}(e)|_+ - w_c L_{\mathrm{H}}(e) \\ -w_m(\mathcal{D}(s^0, \hat{s}(e)) - \mathcal{D}(s^0, s^t)), \end{array} \quad (1)$$

其中，w_r、w_c和 w_m 是常数权重；$\mathcal{D}$ 衡量两个状态之间的差异；而$\hat{s}(e)$是从当前状态 s^t到达子目标 e 后的状态。公式（1）中的三个项分别评估：i）帮助实现子目标的预期辅助收益；ii）辅助智能体的成本；以及 iii）子目标引入的额外状态变化（相对于初始状态）。这三个项确保辅助智能体选择的子目标满足以下条件：i）在高度确定的情况下有望加速任务；ii）对辅助智能体而言成本不过高；以及 iii）能够分别恢复任务不需要的对象的初始状态。给定子目标e^*，我们执行辅助计划 $\Gamma_{\mathrm{H}}(e^*)$中的第一个动作。

IV. 在线 Watch-And-Help

为了评估不同的辅助方法，我们提出了 Online Watch-And-Help（O-WAH）这一具身人工智能辅助挑战。在该挑战中，辅助智能体必须推断主智能体的目标，并尽可能快速地帮助其实现目标。该挑战将现有的 Watch-And-Help（WAH）扩展为在线辅助问题。O-WAH 构建于逼真的多智能体虚拟平台 VirtualHome-Social [2], 之上，用于模拟日常家务任务（如图 3 所示）。每项任务的目标由一组谓词及其数量定义，用于表示环境中不同对象的目标位置。我们从五种常见的家务任务类型中采样挑战中的每个目标：摆餐桌、放入洗碗机、补充冰箱库存、准备餐食和拿零食。需要注意的是，我们定义任务类型只是为了确保目标模拟现实生活中的家务任务，但不会将这些信息提供给辅助智能体。如表 I 所总结，这些任务可能产生不同类型的不确定性：i）对象数量的不确定性；ii）所需对象的不确定性；以及 iii）目标位置的不确定性。与先前研究相比，O-WAH 中的目标空间扩大了 1 到 2 个数量级。我们在 [2]中采用相同的动作空间。

为了创建训练回合，我们首先从五套训练公寓中的一套采样一个目标和初始环境，然后使用内置规划器控制主智能体独自执行任务。内置规划器与 [2] 中的分层规划器相同。我们创建了一个包含 6,000 个回合的大型训练集，以及一个包含 300 个



图 3：O-WAH 在其中一套模拟公寓中的示例设置。

表 I：每种任务类型的目标定义及独特目标的数量。

任务名称	目标定义	#目标
摆餐桌	将 N 个盘子、N 把叉子、N 个 OBJ 放在 LOC 上，其中 N ∼ U(1,3), OBJ ∼choice(⌈水杯、红酒杯⌋), LOC ∼choice(⌈厨房桌，茶几⌋)	12
放入洗碗机	将 OBJ_POOL 中的 N 个对象放入洗碗机，其中 N ∼ U(3,7), OBJ_POOL = ⌈叉子、盘子、水杯、红酒杯⌋	315
补充冰箱库存	将 OBJ_POOL 中的 N 个对象放入冰箱，其中 N ∼U(3,7), 对象_POOL = ⌈鲑鱼、苹果、纸杯蛋糕、布丁⌋	315
准备餐食	将 N 条鲑鱼、N 个苹果、N 个对象放在位置上，其中 N ∼ U(1,3), 对象 ∼choice(⌈纸杯蛋糕、布丁⌋), 位置 ∼choice(⌈厨房桌、茶几、炉灶⌋)	18
拿零食	将1个遥控器、1份调味品和1包薯片放在茶几上	1

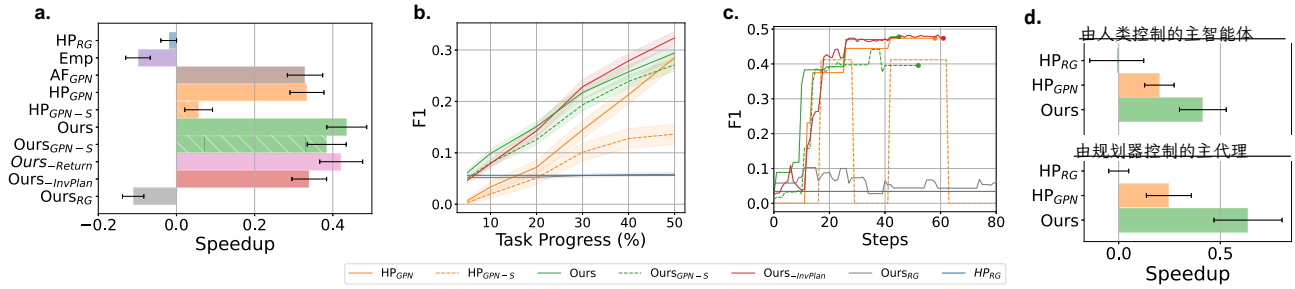


图4：**(a)** 不同方法的加速比（条纹柱表示使用小型训练集）。误差线表示标准误差。**(b)** 任务执行过程中预测目标的F1分数。x轴根据主智能体单独完成各项任务所需的步数进行归一化。曲线表示均值，阴影区域表示标准误差。**c)** 单个测试回合中不同方法随时间变化的F1分数，圆点表示某个基线方法完成任务所用的步数。**(b)**和**(c)**中的虚线表示使用小型训练集。**(d)** 人体实验的结果。此处展示主智能体由内置规划器或人类玩家控制时不同方法的加速比（请注意，这两种条件下的结果均基于相同的10个测试回合）。

个回合的小型训练集。测试集包含 100 个回合，来自训练期间未见过的两套测试公寓。

我们使用目标谓词的 F1分数来衡量目标推断准确率。为评估协助效果，我们使用加速比，即比较辅助智能体与主智能体协作时的回合长度（L_{H}）和主智能体单独工作时的回合长度（L_{M}），也就是$L_{\mathrm{M}}/L_{\mathrm{H}}-1$。每个回合的时间限制设为 250 步，并报告 3 次运行的平均表现。

V. 实验

A. 基线方法

我们将 NOPA 与若干基线方法进行比较。

$\mathbf{HP}_{\mathbf{GPN}}$：我们采用原始 Watch-And-Help 挑战赛中表现最佳的方法作为此基线方法。该方法是一种分层规划器（HP），基于 GPN 认为概率最高的目标。具体而言，在每一步，**$\mathbf{HP}_{\mathbf{GPN}}$**使用目标 $\hat{g}=\arg\max_g q(g|s^0,\ s^t)$。

$\mathbf{AF}_{\mathbf{GPN}}$：我们通过采用 NOPA 的在线目标推断来扩展**$\mathbf{HP}_{\mathbf{GPN}}$**。我们使用 HP 为每个预测目标生成计划，并执行所有计划中出现频率最高的第一个动作。**赋能**：借鉴赋能的理念 [34]，此基线方法在每一步均匀采样 K 个目标，为这些目标预测计划和中间状态，并选择中间状态中出现频率最高的边作为辅助子目标（即对任意 目标而言最常见的子目标）。

$\mathbf{HP}_{\mathbf{RG}}$：一种基于回合开始时随机采样目标的分层规划器。

我们考虑以下消融方法，以评估 NOPA 不同组件的效果。

我们的方法$_{\mathbf{RG}}$：我们将算法 2 中的提议分布 q 替换为均匀分布。

我们的方法$_{\mathbf{-InvPlan}}$：不使用逆向规划的我们的方法。

我们的方法$_{\mathbf{-Return}}$：我们的方法不将无关物体送回其初始位置（$w_m=0$ 见公式 (1)）。

默认情况下，GPN 在较大的训练集上训练。为了评估 NOPA 的样本效率，我们还报告了 GPN 在小型训练集上训练时**我们的方法** 和**$\mathbf{HP}_{\mathbf{GPN}}$**的表现，其中小型训练集以下标**GPN-S**表示。为了衡量协助效果的上界，我们还实现了一个预言机辅助者**$\mathbf{HP}_{\mathbf{GT}}$**，它知道真实目标，并由一个 HP 控制。

我们为 NOPA 设置 $T_{\max}=250$、$T_{\text{prop}}=15$、$w_r=1$、$w_c=1$、$w_d=5$ 和 $L_{\max}=100$。对于所有提出多个目标的方法，我们使用 $K=20$ 个提议。我们使用 Adam [42] 训练 GPN，学习率设为。0009，批量大小为 256。

B. 结果

1) 由规划器控制的主代理：我们使用由内置规划器控制的主智能体评估所有方法，并在图**4a**中报告辅助加速比（跨回合的平均值和标准误差）。对于具有不同目标推断模块的方法，我们还在图**4b**中报告其目标推断结果的F1分数。预言机智能体在掌握目标真实信息的情况下运行，其加速比为 **$\mathbf{HP}_{\mathbf{GT}}$** 1.29。

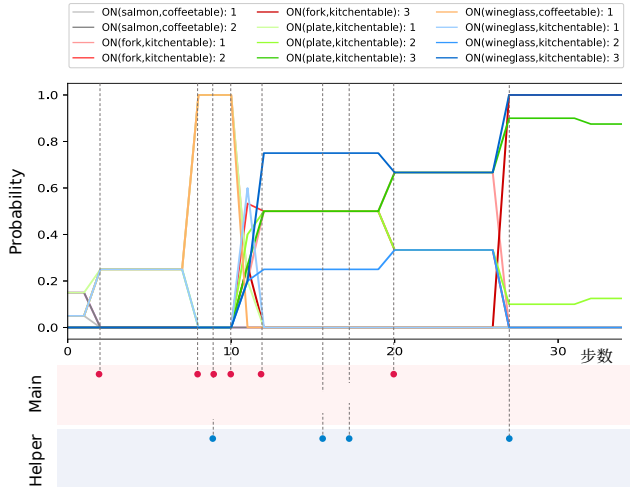


图5：NOPA 对图4c所示同一任务的目标推断和计划，该任务是为3人布置厨房餐桌。我们展示了每一步中基于粒子得出的主要谓词及其数量的后验概率、主智能体的关键动作（以红点标示）以及关键协助行动（以蓝点标示）。在第2步，观察到主智能体朝洗碗机走去后，NOPA 排除了涉及附近物品（例如苹果、鲑鱼）但不在洗碗机内的提议。第8步主智能体拿起一把叉子后，NOPA 高度确信目标是为至少一人布置桌子。因此，在下一步，辅助智能体采取了它的第一个动作——走过去拿起一个盘子。第10步看到主代理朝厨房餐桌走去后，目标位置变得确定。观察到更多动作后，推断逐渐收敛到为3人布置厨房餐桌。

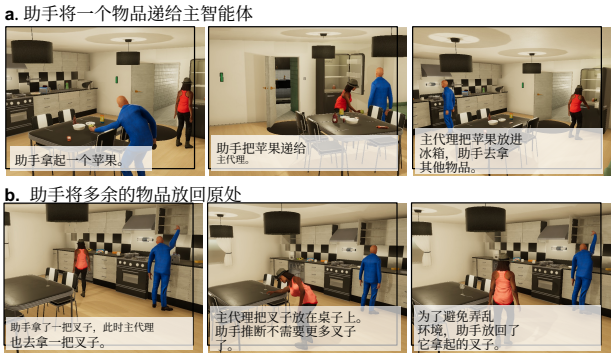


图6：超越直接实现最终目标的协助计划示例。主代理为红色，助手为蓝色。

NOPA（**我们的方法**）优于所有基线方法，取得了最高的加速比。在任务早期阶段，它的目标推断准确率也最高，这为其成功协助奠定了基础。从图**4c** 中可以更清楚地看到这一优势（由于每个回合的时间归一化方式不同，提升幅度看起来较小）。**赋能** 较低的加速比表明，在线目标推断对于有效协助是必要的，尽管先前研究[34]展示了它在某些领域中的成功。鉴于预测目标可能存在不确定性，使用多个目标候选能够带来更好的协助效果，这一点可通过比较 $\textbf{Ours}_{\textbf{GPN}}$ 和 $\textbf{HP}_{\textbf{GPN}}$看出。当 GPN 使用更少的数据进行训练、因而准确度较低时（**GPN-S**），这一效果会更加明显。我们还发现，与均匀目标候选相比，神经网络引导的目标候选可以大幅提升目标推断效果（$\textbf{Ours}_{\textbf{RG}}$）。此外，结果表明，逆向规划对于过滤来自 **GPN** 的虚假目标候选至关重要；由于它能让目标推断比 $\textbf{Ours}_{\textbf{-InvPlan}}$ 及其他基线方法更早达到相对较高的准确度（见图**4c**），因此显著提升了相较于 $\textbf{Ours}_{\textbf{-InvPlan}}$ 的加速比。最后，通过比较 NOPA 与 $\textbf{Ours}_{\textbf{-Return}}$，我们可以看到，避免不必要地扭曲环境可略微提升加速比；$\textbf{Ours}_{\textbf{-Return}}$ 还会造成多出 11.2% 的不必要状态变化。

图5展示了 NOPA 的一个典型成功案例，其中任务与图**4c**中的任务相同。该案例表明，NOPA 能够：i）通过过滤掉与主智能体动作不一致的目标候选，尽早准确地进行目标推断；ii）根据更多观察，正确更新目标推断及其不确定性估计；iii）基于筛选后的目标候选以及推断中的不确定性，规划有效的协助行动。请注意，辅助智能体会保持静止，但一旦目标推断变得足够确定，就会立即采取有用的协助行动；此外，得益于它逐步更新所需对象数量，也能避免拿取多余对象。

我们还观察到，NOPA 支持多种不同的协助行为，并非仅限于直接实现最终目标，如图6所示。首先，辅助智能体有时会选择一个子目标，即将对象递交给主智能体。例如，如图**6a**所示，辅助智能体将苹果递交给紧挨着冰箱的主智能体，从而加快任务执行。其次，辅助智能体意识到多余对象并非实现目标所需后，可以将它们放回初始位置（图**6b**）。补充视频[2] 展示了更多示例。

2）由人类控制的主智能体：为了评估辅助智能体协助真实人类的效果，我们开展了一项人体实验，其中主智能体由人类玩家控制。我们使用了10个测试回合，共进行了40次试验。在每次试验中，我们会要求一名人类参与者独自完成任务（以估算独自完成每项任务所需的步数），或与由三种方法之一控制的辅助智能体合作：NOPA、$\textbf{HP}_{\textbf{GPN}}$ 和 $\textbf{HP}_{\textbf{RG}}$。参与者并不知道自己与哪种辅助智能体合作。我们招募了10名此前未接触过本系统的参与者（平均年龄为= 32.3岁；其中4名女性）。如图4（**d**）所示，主智能体由人类玩家控制时，各方法的排名仍然一致。在两种条件下，NOPA的表现没有显著差异（$t(9) = 0.87, \rho = 0.40$）。

[2]补充视频可在 https://youtu.be/Oawo9pynPL0 查看。

VI. 结论

在这项工作中，我们提出了一种用于构建具有社交智能的家庭助手的新方法——神经引导在线概率辅助（NOPA）。该方法融合了：（1）一种结合目标提议网络与逆向规划的混合在线目标推断算法；（2）一种不确定性感知助人规划器，可根据最终目标和中间状态确定有价值的助人子目标。为实现系统化且可扩展的评估，我们基于逼真的虚拟家庭平台推出了一项新的具身人工智能辅助挑战——Online Watch-And-Help。在这项挑战中，我们使用多种基线方法评估了NOPA，主智能体由内置规划器或人类控制。实验结果表明，NOPA显著优于基线方法，且训练样本效率很高。未来，我们计划将NOPA扩展到部分可观测设定，并将其应用于真实家庭中的机器人。

致谢

本研究得到美国国防高级研究计划局机器常识计划、ONR MURI N00014-13-1-0333 以及洛克希德·马丁的资助。

参考文献

[1] M. Carroll、R. Shah、M. K. Ho、T. Griffiths、S. Seshia、P. Abbeel 和 A. Dragan，《学习人类知识对人机协调的效用》，发表于神经信息处理系统进展，第 32 卷，2019年。[2] X. Puig、T. Shu、S. Li、Z. Wang、Y.-H. Liao、J. B. Tenenbaum、S. Fidler 和 A. Torralba，《Watch-And-Help：社会感知与人机协作的一项挑战》，发表于国际学习表征会议，2021年。[3] N. Rabinowitz、F. Perbet、F. Song、C. Zhang、S. A. Eslami 和 M. Botvinick，《机器心智理论》，发表于国际机器学习会议。PMLR，2018年，第 4218–4227 页。[4] Z. Cao、H. Gao、K. Mangalam、Q.-Z. Cai、M. Vo 和 J. Malik，《结合场景上下文的长期人体运动预测》，发表于欧洲计算机视觉会议。Springer，2020年，第 387–404 页。[5] B. Liu、E. Adeli、Z. Cao、K.-H. Lee、A. Shenoi、A. Gaidon 和 J. C. Niebles，《行人意图预测的时空关系推理》，发表于*IEEE*机器人与自动化快报，第 5 卷，第 2 期，第 3485–3492 页，2020年。[6] Z. Nan、T. Shu、R. Gong、S. Wang、P. Wei、S.-C. Zhu 和 N. Zheng，《学习推断日常活动中的人类注意力》，发表于*Pattern Recognition*，第 103 卷，第 107314 页，2020年。[7] P. Dendorfer、A. Osep 和 L. Leal-Taixé，《Goal-GAN：基于目标位置估计的多模态轨迹预测》，发表于亚洲计算机视觉会议论文集，2020年。[8] H. Zhao、J. Gao、T. Lan、C. Sun、B. Sapp、B. Varadarajan、Y. Shen、Y. Shen、Y. Chai、C. Schmid、等，《TNT：目标驱动的轨迹预测》，发表于*arXiv*预印本 *arXiv:2008.08294*，2020年。[9] Y. Yao、E. Atkins、M. Johnson-Roberson、R. Vasudevan 和 X. Du，《BITRAP：结合多模态目标估计的双向行人轨迹预测》，发表于*IEEE*机器人与自动化快报，第 6 卷，第 2 期，第 1463–1470 页，2021年。[10] H. Tran、V. Le 和 T. Tran，《目标驱动的长期轨迹预测》，发表于*IEEE/CVF*冬季计算机视觉应用会议论文集，2021年，第 796–805 页。[11] K. Mangalam、Y. An、H. Girase 和 J. Malik，《从目标、路径点和路径到长期人体轨迹预测》，发表于*IEEE/CVF*国际计算机视觉会议论文集，2021年，第 15 233–15 242 页。[12] S. Carberry，《计划识别技术》，发表于用户建模与用户适应交互，第 11 卷，第 1 期，第 31–48 页，2001年。[13] M. Ramírez 和 H. Geffner，《作为规划的计划识别》，发表于第二十一届国际人工智能联合会议，2009年。[14] S. Sohrabi、A. V. Riabov 和 O. Udrea，《重新审视作为规划的计划识别》，发表于*IJCAI*，2016年，第 3258–3264 页。[15] C. L. Baker、J. Jara-Ettinger、R. Saxe 和 J. B. Tenenbaum，“人类心智化中信念、欲望和感知的理性定量归因”，《自然·人类行为》，第 1 卷，第 4 期，第 1–10 页，2017年。[16] T. Ullman、C. Baker、O. Macindoe、O. Evans、N. Goodman 和 J. Tenenbaum，“帮助还是阻碍：社会目标推断的贝叶斯模型”，神经信息处理系统进展，第 22 卷，2009年。[17] T. Zhi-Xuan、J. Mann、T. Silver、J. Tenenbaum 和 V. Mansinghka，“有界理性规划智能体的在线贝叶斯目标推断”，神经信息处理系统进展，第 33 卷，第 19 238–19 250 页，2020年。[18] A. Netanyahu、T. Shu、B. Katz、A. Barbu 和 J. B. Tenenbaum，“PHASE：用于机器社会感知的物理基础抽象社会事件”，载于第 35 届 AAAI人工智能会议（AAAI），2021年。[19] R. Tejwani、Y.-L. Kuo、T. Shu、B. Katz 和 A. Barbu，“递归马尔可夫决策过程中的社会互动”，载于机器人学习会议。PMLR，2022年，第 949–958 页。[20] M. A. Goodrich 和 A. C. Schultz，“人机交互：综述”，《人机交互基础与趋势》，第 1 卷，第 3 期，第 203–275 页，2007年。[21] P. A. Lasota、T. Fong、J. A. Shah 等，《安全人机交互方法综述》。Now Publishers，代尔夫特，荷兰，2017年，第 104 卷。[22] S. Nikolaidis、R. Ramakrishnan、K. Gu 和 J. Shah，“从联合行动示范中高效学习模型，以完成 人机协作任务”，载于2015年 ACM/IEEE 人机交互国际会议（HRI）第 10 届会议。IEEE，2015年，第 189–196 页。[23] L. Rozo、S. Calinon、D. G. Caldwell、P. Jimenez 和 C. Torras，“从人类示范中学习物理协作机器人行为”，《IEEE机器人学汇刊》，第 32 卷，第 3 期，第 513–527 页，2016年。[24] M. Jaderberg、W. M. Czarnecki、I. Dunning、L. Marris、G. Lever、A. G. Castaneda、C. Beattie、N. C. Rabinowitz、A. S. Morcos、A. Ruderman 等，“基于群体的强化学习在 3D多人游戏中达到人类水平的表现”，《科学》，第 364 卷，第 6443 期，第 859–865 页，2019年。[25] N. Bard、J. N. Foerster、S. Chandar、N. Burch、M. Lanctot、H. F. Song、E. Parisotto、V. Dumoulin、S. Moitra、E. Hughes 等，“Hanabi挑战：人工智能研究的新前沿”，《人工智能》，第 280 卷，文章编号 103216，2020年。[26] H. Hu、A. Lerer、A. Peysakhovich 和 J. Foerster，“用于零样本协调的‘他者玩法’”，载于国际机器学习会议。PMLR，2020年，第 4399–4410 页。[27] S. A. Wu、R. E. Wang、J. A. Evans、J. B. Tenenbaum、D. C. Parkes 和 M. Kleiman-Weiner，“人多手杂：用于协调多智能体协作的贝叶斯推断”，《认知科学专题》，第 13 卷，第 2 期，第 414–432 页，2021年。[28] A. Fern、S. Natarajan、K. Judah 和 P. Tadepalli，“基于决策理论的辅助模型”，《人工智能研究杂志》，第 50 卷，第 71–104 页，2014年。[29] C. Liu、J. B. Hamrick、J. F. Fisac、A. D. Dragan、J. K. Hedrick、S. S. Sastry 和 T. L. Griffiths，“目标推断提升了人机协作中的客观表现和主观感知表现”，载于第 15 届自主智能体与多智能体系统国际会议（AAMAS），2016年。[30] S. V. Albrecht 和 P. Stone，“对其他智能体进行建模的自主智能体：全面综述与待解决问题”，《人工智能》，第 258 卷，第 66–95 页，2018年。[31] D. Hadfield-Menell、S. J. Russell、P. Abbeel 和 A. Dragan，“合作式逆向强化学习”，神经信息处理系统进展，第 29 卷，2016年。[32] S. Javdani、H. Admoni、S. Pellegrinelli、S. S. Srinivasa 和 J. A. Bagnell，“通过事后优化实现遥操作与团队协作中的共享自主性”，《国际机器人研究杂志》，第 37 卷，第 7 期，第 717–742 页，2018年。[33] A. Shah、S. Li 和 J. Shah，“在不确定规格下进行规划（PUnS）”，《IEEE机器人与自动化快报》，第 5 卷，第 2 期，第 3414–3421 页，2020年。[34] Y. Du、S. Tiomkin、E. Kiciman、D. Polani、P. Abbeel 和 A. Dragan，“AVE：通过赋能实现辅助”，神经信息处理系统进展，第 33 卷，第 4560–4571 页，2020年。[35] S. C. Ong、S. W. Png、D. Hsu 和 W. S. Lee，“针对混合可观测性机器人任务的不确定性规划”，《国际机器人研究杂志》，第 29 卷，第 8 期，第 1053–1068 页，2010年。

[36] J. Johnson、R. Krishna、M. Stark、L.-J. Li、D. Shamma、M. Bernstein 和 L. Fei-Fei，“利用场景图进行图像检索”，载于*IEEE*计算机视觉与模式识别会议论文集，2015年，第 3668–3678 页。[37] Y.-H. Liao、X. Puig、M. Boben、A. Torralba 和 S. Fidler，“通过活动草图合成环境感知活动”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2019年，第 6291–6299 页。[38] F. Yan、D. Wang 和 H. He，“利用神经逻辑学习实现机器人对空间关系的理解”，载于*2020*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*）。IEEE，2020年，第 8358–8365 页。[39] S. Nguyen、O. S. Oguz、V. N. Hartmann 和 M. Toussaint，“用于机器人序贯操作规划的场景图表示自监督学习”，载于机器人学习会议，2020年，第 2104–2119 页。[40] Y. Zhu、J. Tremblay、S. Birchfield 和 Y. Zhu，“利用几何与符号场景图进行长时程操作的分层规划”，载于*2021*年 *IEEE*机器人与自动化国际会议（*ICRA*）。IEEE，2021年，第 6541–6548 页。[41] S. Srivastava、C. Li、M. Lingelbach、R. Martín-Martín、F. Xia、K. E. Vainio、Z. Lian、C. Gokmen、S. Buch、K. Liu、等，“BEHAVIOR：虚拟、交互式和生态环境中的日常家庭活动基准”，载于机器人学习会议。PMLR，2022年，第 477–490 页。[42] D. P. Kingma 和 J. Ba，“Adam：一种随机优化方法”，*arXiv*预印本 *arXiv:1412.6980*，2014年。