arXiv:2602.06038v1 [cs.RO] 5 Feb 2026

CommCP：通过基于大语言模型的通信与共形预测实现高效多智能体协同

Xiaopan Zhang*、Zejin Wang*、Zhixu Li、Jianpeng Yao、Jiachen Li‡

摘要——为了完成由人类以自然语言提供的任务分配，机器人必须理解指令，生成并回答与场景理解相关的问题，以及操控目标物体。在实际部署中，通常需要多台具备不同操控能力的异构机器人协作处理不同任务。除了需要专门的操控技能外，有效的信息收集对于完成这些任务也至关重要。为解决这一问题，我们将完全合作场景中的信息收集过程形式化为一个尚未充分探索的多智能体多任务具身问答（MM-EQA）问题。这是对经典具身问答（EQA）的一种新颖拓展；在这一问题中，有效通信对于协调行动、避免重复工作至关重要。为此，我们提出 CommCP，这是一种专为 MM-EQA 设计的新型基于大语言模型的去中心化通信框架。我们的框架采用保形预测校准生成的消息，从而最大限度地减少接收方干扰并提高通信可靠性。为评估该框架，我们推出了 MM-EQA 基准，其中包含多样化、照片级真实感的家庭场景和具身问题。实验结果表明，与基线方法相比，CommCP 显著提升了任务成功率和探索效率。实验视频、代码和数据集可在我们的项目网站上获取：https://comm-cp.github.io。



图1. 在家庭环境中，机器人通过交换观测和推理，协作完成分配给它们的任务。每个智能体利用大语言模型的校准输出，生成可靠且目标导向的消息。左下图展示了机器人1接收机器人2提供的信息后所采用导航路径的鸟瞰图。右上方的序列呈现了两台机器人的摄像头视图及其不同时间戳。

I. 引言

现代服务机器人旨在理解人类指令，并在真实的家庭环境中完成任务（例如，“如果电视当前处于开机状态，就把它关掉”，“把客厅里的红色枕头拿到卧室床上”）。这一过程包括理解自然语言指令，为场景理解和推理生成并回答相关问题（例如，“电视开着吗？”“枕头是什么颜色？”），并据此操作目标物体。其中一个关键步骤是回答这些问题，这项任务称为具身问答（EQA） [1], ，它要求机器人从随机起始位置出发，高效探索三维环境并主动

收集信息，直到能够给出有把握的答案。先前的研究已在单智能体场景[1]–[5]中对此进行了探讨。相比之下，我们设想未来的家庭中将配备多个异构服务机器人，每个机器人都具备不同的能力，并承担不可转移的任务。虽然它们无法接手彼此的任务，但可以访问所有生成的问题，并共享观测结果和解读，以提升探索效率。我们将这种协作信息收集场景定义为多智能体多任务具身问答（*MM-EQA*）问题，这是一个能够促进现实场景中多机器人协作的新颖挑战。

虽然现有的单智能体EQA解决方案可以通过让每台机器人独立工作来适配多智能体环境，但这种简单的方法效率低下。通信能够实现相互协助，改善探索并提高更快完成任务的可能性。然而，未经校准的通信可能会通过分享无关或误导性信息而降低效率。因此，确保消息准确且与接收者的任务相关至关重要。本研究通过设计一种通信框架，提升多智能体探索效率和任务性能，从而解决MM-EQA问题。

大语言模型（LLM）凭借理解自然语言查询、进行推理并以自然语言作答的卓越能力，在解决 EQA 任务方面展现出巨大潜力 [6]。在 MM-EQA 中，自然语言是理想的通信协议，因为大语言模型本身就是通过对话进行训练的。其他领域已提出多种基于大语言模型的通信方法 [7], [8], , 但这些方法无法直接应用于我们的 MM-EQA 场景。此外，大语言模型经常会产生校准不当且过度自信的输出 [9],, 从而可能提供无关或误导性信息。这会阻碍协作效率，因为智能体可能会共享不准确的数据，

*贡献相同 ‡通讯作者 X. Zhang、Z. Wang、Z. Li、J. Yao 和 J. Li 隶属于美国加利福尼亚州河滨市的加利福尼亚大学河滨分校可信自主系统实验室。 {xzhan006, jiachen.li}@ucr.edu。

进而降低整体探索有效性 [10]。

我们的工作针对这一挑战，为 MM-EQA 开发了一种基于大语言模型的通信框架。我们的核心洞见是智能体只应向伙伴智能体传达它有把握认为与伙伴智能体任务相关的信息（见图1）。我们提出了 CommCP，这是一种新颖的去中心化、基于大语言模型的通信框架，采用保形预测（CP） [11], [12] 来校准大语言模型输出的置信度。我们的框架确保大语言模型生成的输出更加可靠，并减少无关或误导性信息对伙伴智能体造成的负面影响，最终提升多智能体系统的整体任务性能和效率。为评估所提出的框架，我们基于真实场景和 Habitat-Matterport 3D (HM3D) 数据集 [13]构建了一个新颖的 MM-EQA 基准。实验结果表明，与基线方法相比，我们的方法显著提升了任务成功率，并大幅缩短了完成时间。

本文的主要贡献如下：

- 我们将以自然语言给出的任务分配的信息收集过程形式化为一种新颖的多智能体多任务具身问答（MM-EQA）问题：多台机器人组成团队，在共享环境中处理具身问答任务，并通过通信交换信息或答案。
- 我们提出 CommCP，这是一种面向 MM-EQA 的新颖大语言模型去中心化通信框架。该框架采用保形预测对生成的消息进行校准，以减少对其他智能体的干扰，并提高通信的可靠性和效率。
- 我们利用 HM3D 数据集中的照片级真实感场景，创建了一个新颖的 MM-EQA 基准，以验证所提框架的有效性。该基准现已发布，旨在促进未来的研究。

II. 相关工作

A. 基于大语言模型的去中心化多智能体协作

基于大语言模型的多智能体协作近年来受到越来越多的关注 [14], [15]，并涌现出各种用于多智能体任务的系统 [7], [16]–[20]。与单智能体系统或集中式系统不同，去中心化协作系统涉及点对点通信，智能体直接交互；这种架构常见于世界仿真应用 [21], [22]。在这些系统中，通信通常采用由大语言模型生成的自然语言文本，具体内容因应用而异，例如共享环境观测、协调行动或重新分配任务。然而，大语言模型生成的通信究竟有多有效，目前仍缺乏充分研究。正如 [6], 所指出的，去中心化通信往往会产生带宽限制或延迟等成本。因此，智能体必须高效通信，并避免不必要或冗余的消息。当前方法缺乏评估通信质量的机制，只依赖大语言模型的原始输出，导致效率低下，尤其是在智能体依据不完整或不确定的信息采取行动时。

B. 保形预测与校准

近期研究指出，大语言模型存在误校准问题，其文本输出可能表现出过度自信或信心不足。随着基础模型应用于具身任务，这一问题带来了巨大挑战，因为智能体可能对自身决策的置信度校准不当。先前的工作 [12], [23]采用保形预测 [11] ，在机器人规划场景中正式量化大语言模型的不确定性，从而确保机器人以经过校准的置信度执行规划。探索直至置信 [2]将这一方法扩展至具身问答任务，采用多步共形预测来判断视觉语言模型（VLM）何时具有足够的置信度，从而停止探索。据我们所知，我们是首个通过在协作探索过程中校准置信度来运用保形预测增强多智能体通信的工作；此前的研究尚未探讨这一场景。

III. 问题表述

考虑这样一种场景： N_a 台机器人部署在包含多个不同任务分配的三维场景中，每台机器人从一个初始位姿 g_0^i出发，旨在回答$q_{1:N_q}^i$ 与其任务分配相关的问题。目标是在时间范围 T_{max}内回答所有问题，同时最大化成功率并尽量缩短探索时间。每台机器人都知道所有问题，包括分配给其他机器人的问题。它们可以通过自然语言消息进行通信，记作 ζ^i，以交换信息。

每台机器人 $i \in N_a$ 都配备了摄像头，可在每个时间步 t为机器人提供局部场景的RGB图像$I_{c,\ t}^i$ 和深度图像$I_{d,\ t}^i$ 作为观测。每台机器人在每个时间步的位姿（二维位置和朝向）记为 g_t^i，所有机器人的位姿汇总为集合 $G_t = \{g_t^i \mid i = 1, ..., N_a\}$。每台机器人都配备了无碰撞规划器 π 用于导航。给定当前位姿 g_t^i 和目标位置，规划器π 会确定下一个可行位姿 g_{t+1}^i，并由低层控制器在 $t+1$将机器人移动至规划的位姿。

在此情况下，多机器人多任务具身问答（MM-EQA）问题定义为一个元组$\xi := (E,\ G_0,\ T_{\text{max}},\ Q,\ Y)$，其中 E是尺寸为 $L \times W \times H$的三维场景，该场景被离散化为体素地图 M ，由边长为 l的立方体组成。 L、W和 H 分别表示体素地图 M的长度、宽度和高度；$G_0 = \{g_0^i \mid i = 1, ..., N_a\}$ 是机器人的初始位姿集合，T_{max} 是允许机器人探索场景并完成任务的最大时间范围。每个机器人 i 都会被分配 N_q 个问题来回答。集合 $Q = \{q_j^i \mid i = 1, ..., N_{\text{a}}, j = 1, ..., N_{\text{q}}\}$ 汇总了分配给机器人的所有问题，其中 q_j^i 是分配给第 i^{th} 个机器人的第 j^{th} 个问题。每个问题都是一道包含四个选项 {‘A’、‘B’、‘C’、‘D’}的选择题。真实答案用集合 $Y = \{a_j^i \in$‘A’、‘B’、‘C’、‘D’$\} \mid i = 1,\ \text{.....},\ N_a,\ j = 1,\ \text{.....},\ N_q\}$表示。

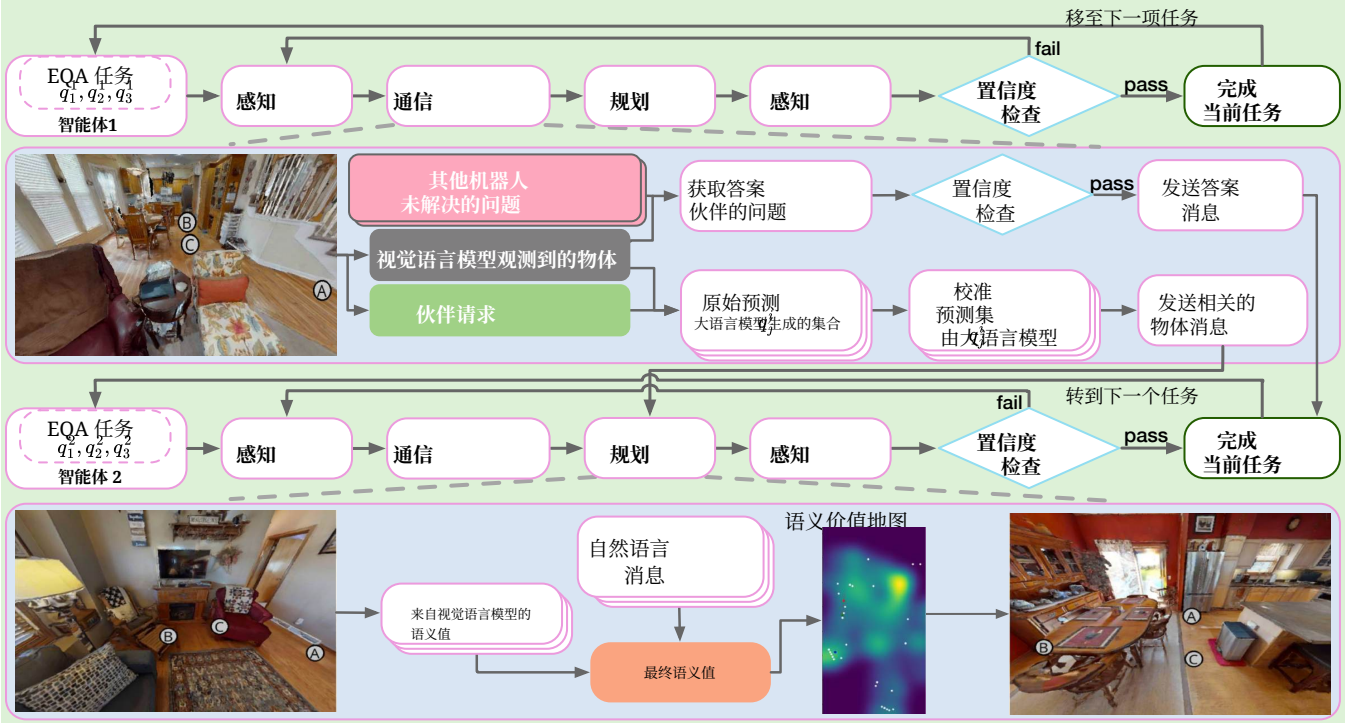

图2. 我们框架的概览展示了每个机器人所包含的模块：感知模块、通信模块、规划模块和置信度检查模块。在每个时间步，机器人根据当前视野和来自其他机器人的通信消息生成局部和全局语义值（SV）。机器人利用二维加权语义值地图进行导航，并处理来自其他智能体的相关物体检查请求。消息根据机器人的当前视野生成，并通过保形预测进行校准，以提高相关性。

IV. 方法

本节介绍 CommCP 框架，该框架利用通信增强针对 MM-EQA 问题的多智能体探索。在 [2] 提出的单智能体单任务具身问答方法基础上，我们的框架进一步支持通信，以提高任务完成效率和成功率。此外，我们的框架还可轻松扩展，以应对多智能体环境中更复杂的人类任务分配：机器人根据其获取的信息执行下游任务，而分配给机器人的问题则基于其执行下游任务的能力。

整体框架架构如图2所示，由四个核心模块组成：感知模块、通信模块、规划模块和置信度检查模块。规划模块和置信度检查模块由 [2]修改而来。在通信模块中，消息会被生成并投射到规划模块中的语义价值地图上，以指导机器人的探索策略。该模块还使机器人能够在置信度足够时回答其他机器人的问题。本节中的所有记号均对应于时间 t。对于每个机器人 i，感知模块会使用视觉语言模型，根据以下提示从RGB图像 I_c^i 中检测一组观测物体 O^i_{observe}：

> 请考虑室内场景，分析所提供的图像并列出你观察到的所有物体。提供每个物体的名称及其颜色，并以逗号分隔。

感知模块检测到的 O^i_{observe} 会被输入推理流程，以确定哪些对象应纳入通信，并采用保形预测。系统依次将每个未解决的问题连同 O^i_{observe} 一起提示大语言模型，以生成答案。如果答案通过了第 IV.D 节所述的置信度检查，且该问题被分配给作答的机器人，则继续处理。否则，系统会将答案发送给负责的机器人。

A. 基于大语言模型的对象相关性推理

在室内场景中，物体通常会按照人类使用习惯形成一定的组织模式，而大语言模型可以利用这些模式中的常识知识。因此，如果大语言模型判断某个区域内观察到的物体与目标物体高度相关，那么目标物体很可能就在附近。基于这一认识，我们设计了一种通信框架，使机器人能够共享相关物体信息，或回答其他机器人提出的问题。

在探索过程中，每个机器人 i 发送一个请求 $\hat{r}^i_j$ 以寻求帮助回答其问题 q^i_j，并提供其目标物体 $\hat{O}^i_{\text{request}}$。大语言模型会评估机器人 i观察到的物体，这些物体记为 O^i_{observe}；观察到的物体和目标物体分别标记为**观察到的物体** 和**请求**。我们采用零样本思维链 [24] 提示大语言模型，在生成最终输出之前进行详细推理。以下是用于大语言模型的提示词。其中，*系统提示* 是给大语言模型的指令，而*用户提示* 是与大语言模型对话的内容。

> **系统提示：**
> 作为家中的机器人，你的伙伴正在寻找 {请求}，你可以告知他们你观察到的情况。

用户提示：

你观察到 {观察到的内容}。你的伙伴想要寻找 {请求}。由于你和伙伴不在不同的位置，请评估你的伙伴是否值得前往你所在的位置寻找 {请求}。

你有四个选项：***A*** （是。因为 {请求} 和{观察到的内容}相同，你可以直接找到 {请求}。）***B*** （是。它们高度相关。这意味着 {请求} 应该与 {观察到的内容}相近。）***C*** （否。它们没有很强的关联。）***D*** （否。{观察到的内容}是一个常见特征。）

对于选项 *A*，请考虑 {观察到的内容} 和 {请求} 是否是同一事物。对于选项 *B*，请考虑 {请求}最可能出现在哪里。然后判断 {观察到的内容}是否也可能出现在那个位置。对于选项 *C* 和 *D*，由于你和你的搭档不在同一位置，请评估是否值得顺路查看 {观察到的内容}。选项 *D* 意味着 {观察到的内容} 只是房子里的一个常见特征。现在请给出你的分析。

大语言模型的输出分析用于计算每个选项的概率。我们使用以下提示，引导大语言模型根据分析只选择一个选项。

系统提示：
你只能输出一个字母。
用户提示：
你现在观察到 $O^{observe,\,t\,k}$ ^{i} ，。你的搭档想要找到 {请求}。由于你们身处不同位置，你需要判断你的搭档是否值得前来寻找 {请求}。你有四个选项：***A***、***B***、***C***、***D***。以下是你之前的分析：{分析}。你只能输出一个字母。

为机器人 i 观测到的每个物体都会分配一个选项-概率对$O^{i}_{\text{observe},\,k}$:= {选项$_k$，$p_k \mid k=1,\ldots,N_k$}，其中 N_k 是观测到的物体数量，p_k 是大语言模型生成与选项 $_k$ 对应词元的概率。如果返回**选项 A**，则该物体很可能对应请求r^i_j的目标物体$O^{i}_{\text{target},\,j}$。如果选择**选项 B**，则它很可能是与请求 r^i_j相关的物体$O^{i}_{\text{relevant},\,j}$ ^。选项为 C 和 D 的物体将被忽略。

B. 基于保形预测的消息生成

为缓解大语言模型生成的目标列表和相关物体列表可能存在的误校准问题（这会导致无关信息），我们采用保形预测（CP） [11] 来校准用于消息生成的物体列表。保形预测（CP）能够提供具有统计保证的预测集，并以用户指定的概率包含真实值 [12]。具体而言，我们采用分割保形预测 [25],，其步骤包括设计一致性分数以衡量预测的可靠性，收集校准集以建立测试期间比较一致性分数的标准，以及确定是否将某个预测结果纳入最终预测集。在深度学习中，一致性分数定义为模型输出的概率，因为这反映了模型对其所生成答案的置信度。

在我们的框架中，我们采用 p_k 作为一致性分数，仅将 p_k 高于计算所得阈值的选项$_k$ 纳入校准预测集。为处理**选项 A** 和 **选项 B**不同的概率分布，我们定义两个校准集：$\mathcal{Z}^A_{\text{cal}} = \{z_k = (\text{‘A’}\ p_k) \mid k = 1、\ldots\ldots、N_k\}$ 以及 $\mathcal{Z}^B_{\text{cal}} = \{z_k = (\text{‘B’}\ p_k) \mid k = 1、\ldots\ldots、N_k\}$。为收集校准集，我们从 20 个多样化的 HM3D 场景中抽取（已观察物体、目标物体）对，并生成真实标签。为处理可能有多个物体相关的场景，我们按照[12]中的流程，正式将置信度最高的单个选项定义为校准用的真实值。对于每个已标注的物体对，我们应用 V.A 节所述的大语言模型推理过程，获得概率值 p_A 和 p_B，分别构成校准集 $\mathcal{Z}^A_{\text{cal}}$ 和 $\mathcal{Z}^B_{\text{cal}}$。这些校准数据集代表家庭环境中空间共现判断所对应的底层分布 $\mathcal{D}^A_{\text{cal}}$ 和 $\mathcal{D}^B_{\text{cal}}$ 的样本。设$\mathcal{D}^A_{\text{test}}$ 和 $\mathcal{D}^B_{\text{test}}$ 表示新场景中选项-概率对的未知分布，并假设它们与 $\mathcal{D}^A_{\text{cal}}$ 和 $\mathcal{D}^B_{\text{cal}}$可交换。

在我们的 MM-EQA 设置中，我们认为校准样本和测试样本可以视为独立同分布（i.i.d.）。令 $(s_i,\ p_i)$ 表示第 i 个样本，其中 s_i 表示空间关系（A 或 B），p_i表示大语言模型的概率输出。我们的独立同分布假设如下：

$$P((s_1,p_1),(s_2,p_2),\ldots,(s_n,p_n)) = \prod_{i=1}^{n} P(s_i,p_i), \quad (1)$$

其中 $(s_i,p_i) \sim F$ 对所有 i均成立，且 F 是空间关系与概率值上的固定联合分布。此假设成立，是因为校准过程满足独立同分布的要求：1）校准场景从 HM3D家庭数据集中随机采样，不存在时间或空间依赖；2）每次（观测物体、目标物体）评估都代表一次独立的语义判断；3）大语言模型针对不同物体对输出的概率之间不存在因果关系；4）空间共现模式代表典型的家庭组织原则；5）大语言模型的语义理解在不同场景中保持一致；6）物体空间关系的底层分布 F 在整个数据集中保持不变；7）测试期间，每个机器人仅依据当前视野为目标物体和相关物体生成语义值，不考虑过去的帧，从而确保满足保形预测所需的空间和时间独立性。满足这些条件即可确保符合应用保形预测的先决条件。虽然共形预测通常只要求数据具有可交换性，但我们的独立同分布假设提供了更强的理论保证。要实现可靠的覆盖率，约需 $1/\delta$ 个校准样本；20 个场景足以提供充分数据，以达到常见的置信水平（例如 $\delta = 0.05$）。

对于新的测试样本 $z_{\text{test}} = (\text{“A”或“B”}\ p_{\text{test}})$，我们使用共形预测（CP）通过经过校准的分数阈值 p_{thres} 构造预测集 $\mathcal{C}(z_{\text{test}})$。该阈值设为校准集中的分数的 $1-\epsilon_1$ 分位数，其中 ϵ_1 是用户定义的失覆盖率（例如，$\epsilon_1 = 0.05$ 对应 95% 的置信度），它会影响预测集 $\mathcal{C}(\cdot)$ 的大小。随后，将所有分数 p_{test} 达到或超过该阈值的选项纳入预测集。该构造提供了

形式上的边际保证，即以高概率确保真实标签包含在预测集中：

$$P(z_{\text{test}} \in \mathcal{C}(z_{\text{test}})) \geq p_{\text{thres}} \tag{2}$$

更高的 p_{thres} 会使通信过程中分享更有信心的选项，反之亦然。正确选项的相应$O^i_{\text{relevant},j}$ 和$O^i_{\text{target},j}$ 以及正确选项的大致位置被用于根据以下模板生成消息 ζ^i ：“我看到 {个可能与您的目标 {相关的相关物体} ，并且{个可能的目标物体} 可能是您在 {位置}的目标。”如果未识别出任何物体，机器人将不会发送消息。

C. 通信探索

为指导探索，我们首先按照 [2], 计算二维栅格地图的语义值（SV）$\text{SV}^i_{\text{no-com},j}$ ，不考虑从其他智能体接收到的消息。我们将 P 表示为视觉语言模型在当前位姿下识别出的一组前沿点——已探索区域和未探索区域边界上的位置 [27]。$\hat{}p \in P$ 的SV记为$\text{SV}^i_{\hat{p},j}$。在接收到消息 ζ^i后，根据相关物体和目标物体的数量，并结合温度 τ_1 和 τ_2进行缩放，调整每个$\hat{}p \in P$ 基于通信得到的SV：

$$\text{SV}^i_{\text{com},\hat{p},j} = \log\left(\tau_1 \text{Num}(O^i_{\text{relevant},j}) + \tau_2 \text{Num}(O^i_{\text{target},j})\right), \tag{3}$$

其中， τ_1 和 τ_2 分别对相关物体和目标物体的数量进行加权，在传递的语义值中平衡间接语义线索与直接任务信息。对于每项任务，SV 定义如下：

$$\text{SV}^i_{\hat{p},j} = \max(\text{SV}^i_{\text{no-com},\hat{p},j}, \text{SV}^i_{\text{com},\hat{p},j}) \tag{4}$$

位置 $\hat{}p$ 处的语义值设为智能体自身估计值与接收消息中的值两者的最大值，以优先采用信息量最高的来源。计算所有 SV^i_j 值的平均值，以确定最终语义值：

$$\text{SV}^i_{\text{final},\hat{p}} = \frac{1}{N_q}\sum_{j=1}^{N_q} \text{SV}^i_{\hat{p},j} \tag{5}$$

如果 $\text{SV}^i_{\text{final},\hat{p}} = 0$，机器人会随机选择一个前沿点继续探索。移动后，机器人应用体积截断符号距离函数 (TSDF) 融合 [28], [29] ，以更新 M 中体素的占据状态，并根据深度图像 I^i_d 将其标记为已探索或未探索。随后，将这些数据投影到与 SV 地图大小相匹配的二维栅格地图上。接着，机器人使用基于前沿的探索 (FBE) [27] 进行路径规划，其中 SV 值较高的点表明周边区域可能包含更多信息。为提高探索效率，高斯平滑会将每个 SV 点的值扩散到周边区域，从而实现更平滑的导航，并帮助机器人优先探索信息价值较高的区域。

D. 置信度检查

机器人会执行置信度检查，以确定自己是否有足够把握回答具身问题。这里，我们提示视觉语言模型生成四种可能答案的预测概率：{‘A’、‘B’、‘C’、‘D’}，以及问题-图像相关性分数。该相关性分数反映机器人当前视野中包含回答问题所需信息的可能性。视觉语言模型对提示“请考虑这个问题。根据当前视野，你是否有信心回答这个问题？”回答“Yes”的概率，被视为问题 q^i_j 的问题-图像相关性分数 Rel^i_j 。答案概率集合定义为 { $\text{Ans}^i_j(L) \mid L \in$ {‘A’、‘B’、‘C’、‘D’}}。Rel^i_j 和 $\text{Ans}^i_j(L)$ 均取值于区间 $[0, 1]$。如果满足以下条件，则认为答案可信，后续探索将排除问题

$$\{ \quad q^i_j \quad \} \quad {}^i_j L \times \quad {}^i_j > 1 - \epsilon_2 \quad \exists L \in : \quad \epsilon_2$$

对于所有 ‘A’、‘B’、‘C’、‘D’，Ans ( ) Rel，其中， 是用户定义阈值，反映了终止所需的置信度水平。在判断答案是否可接受时，所有机器人均使用此阈值 ϵ_2 。

E. 停止条件

在我们的设置中，机器人回答完分配给它的所有问题后便会终止探索。它既可以根据自身的观测和推理进行分析并作答，也可以使用其他机器人的回答。或者，达到允许的最长时间后，探索也会停止。

V. 实验

A. MM-EQA 基准

我们基于 HM3D 数据集 [13] 构建了一个 MM-EQA 基准，用于评估我们的框架和基线方法。该数据集提供了逼真且多样的三维室内场景。相应的问题颇具挑战性，需要进行语义推理。不同于现有每个场景仅包含一个问题的具身问答数据集，我们为每个场景生成六个问题及其对应答案，以支持多智能体多任务设置。这些具身问题可分为以下五种类型：

1) **位置：** 此类型的问题询问物体的位置，例如：“我把靠垫放在哪里了？ A) 卧室角落 B) 走廊里 C) 篮子架附近 D) 客厅电视旁”。2) **识别：** 此类型的问题询问物体的属性，例如：“浴室里是哪种浴室垫？ A) 红色 B) 白色 C) 黑色 D) 灰色”。3) **计数：** 此类型的问题询问物体的数量，例如：“我有没有把球杆或球留在台球桌上？ A) 一个也没有 B) 一个 C) 两个 D) 三个”。4) **存在性：** 此类型的问题询问某个位置是否有物体，例如：“我把餐具和餐巾放在餐桌上了吗？ A) 是 B) 否”。5) **状态：** 此类型的问题询问物体的状态，例如：“洗衣机开着吗？ A) 是 B) 否”。

我们使用 GPT-4V 生成任务问题及对应的目标物体，并由人工监督和完善。我们的基准测试包含 70 个场景中的 420 项具身问题任务，涉及两台机器人，每台机器人分配三项任务进行测试。我们另外分配 20 个场景，用于创建保形预测的校准数据集。我们使用 Habitat[30]作为基础模拟器，用于加载三维场景。

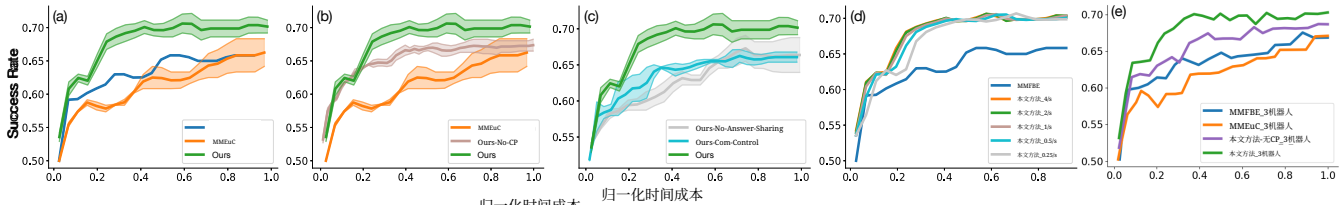

图 3. 我们的 MM-EQA 数据集上 SC 与 NTC 的示意图。(a)–(d) 展示了双机器人团队的结果。(a) 我们的方法与基线方法的比较。(b) 通信模块与保形预测模块的消融比较。(c) 通信消息中的物体数量与答案共享机制的消融比较。(d) 在每秒发送 0.25、0.5、1、2 和 4 条消息的速度下，基线方法与我们的方法的消融比较。(e) 使用三机器人团队比较我们的方法与基线方法的可扩展性分析。

B. 评估指标与基线方法

我们使用两项指标评估方法的性能：**成功率（SR）**，用于衡量分配给机器人的所有问题中正确回答所占的比例；以及**归一化时间成本（NTC）**，用于量化从导航开始到所有机器人完成任务所需的归一化时间。时间成本由两部分组成：机器人移动和消息发送。为分析通信延迟的影响，我们测试了多种不同的消息发送速度。默认情况下，我们将机器人的移动速度设为 1m/s，消息发送速度设为每秒一次，以确保不同实验之间的一致性和可比性。

我们将我们的方法与以下基线方法和消融设置进行比较：

1) **MMFBE：**一种在经典 FBE[27] 基础上扩展而来的多智能体多任务前沿探索方法。MMFBE 使用视觉语言模型回答问题，但在探索过程中不使用该模型进行语义映射，也不涉及通信。2) **MMEuC：**对探索直至置信（EUC）框架 [2], 的多机器人多任务扩展，机器人之间独立运行且不进行通信。3) **Ours-No-CP：**一种允许通信但省略保形预测的消融设置，用于评估保形预测在校准大语言模型置信度方面的作用。
4)**Ours-Com-Control：**一种控制通信消息中所含物体数量的消融设置。为匹配保形预测的效果，我们将物体数量固定为与使用保形预测时相同，并从观测物体中随机抽取。5) **Ours-No-Answer-Sharing：**一种消融设置：机器人交换与彼此任务相关的观测和校准后的预测，但不共享答案，从而单独考察答案共享对性能的影响。

C. 实现细节

我们使用与[2]中相同的视觉语言模型（Prismatic-VLM-13B） [31] 。由于我们的设置要求大语言模型输出概率，因此无法使用 GPT-4V 等最先进的闭源模型。我们改用规模较小的开源模型 LLaMA3-8B-instruct。我们将大语言模型的温度设为 0.7。我们分别将 τ_1， τ_2 设为 1.0 和 10.0。非符合性阈值 p_{thres} 是根据校准集概率的（$1-\epsilon_1$）分位数计算得出的，其中 ϵ_1 控制预测集的大小。在我们的设置中，这对应于选项 A 的 0.6 分位数和选项 B 的 0.82 分位数。

我们使用两块英伟达 6000 Ada Generation图形处理器进行所有实验，视觉语言模型和大语言模型分别运行在不同的设备上。

D. 结果与分析

通信效果。我们通过机器人在不同时间范围内进行探索时所取得的平均成功率，展示了通信模块的有效性，如图3(a)所示。我们的方法显著优于基线方法 MMFBE：在 NTC 为 0.4 时，成功率达到 0.68；而 MMFBE 在 NTC 为 0.8 时的成功率为 0.65，效率几乎翻倍。这一显著的效率提升源于我们的方法能够共享相关且经过校准的信息，使智能体能够优先探索重要区域。相比之下，“MMEuC”基线的表现甚至不如 MMFBE，因为它缺少智能体之间的通信。缺乏通信时，机器人无法协调行动或分享见解，从而导致探索路径冗余或低效。此外，“我们的方法”在收敛时取得了最高的最终成功率，这表明我们的通信策略是有效的。就实际耗时而言，“我们的方法”平均用 445 秒完成所有问题，而 MMFBE 用时 594 秒，体现了更快的任务完成速度。

我们方法的卓越性能归功于信息的策略性交换，这使智能体能够协同探索相关对象，而非在缺乏上下文的情况下随机搜索。这显然优于 MMFBE，因为 MMFBE 缺少基于启发式的引导，也无法受益于协作数据共享。还值得注意的是，随着 NTC 增大、探索的对象增多，大语言模型的输出会更加多样，因此结果的标准差也随之增大。MMFBE 的结果没有偏差，因为 FBE 是一种确定性规则方法。

在图3(c)中，与“我们的方法-无答案共享”基线的比较展示了向其他机器人的问题共享答案所产生的影响。若不具备这一能力，机器人会花费额外时间去探索本可由其伙伴解答的问题。这种效率不足体现为更高的 NTC，以及相较于完整“我们的方法”配置更低的SC。因此，缺少该能力会导致收敛更慢、整体任务效率降低，因为每个智能体都必须独立回答问题，而无法充分利用其伙伴所共享的高置信度信息。通过启用答案共享，我们的方法确保高置信度共享信息被

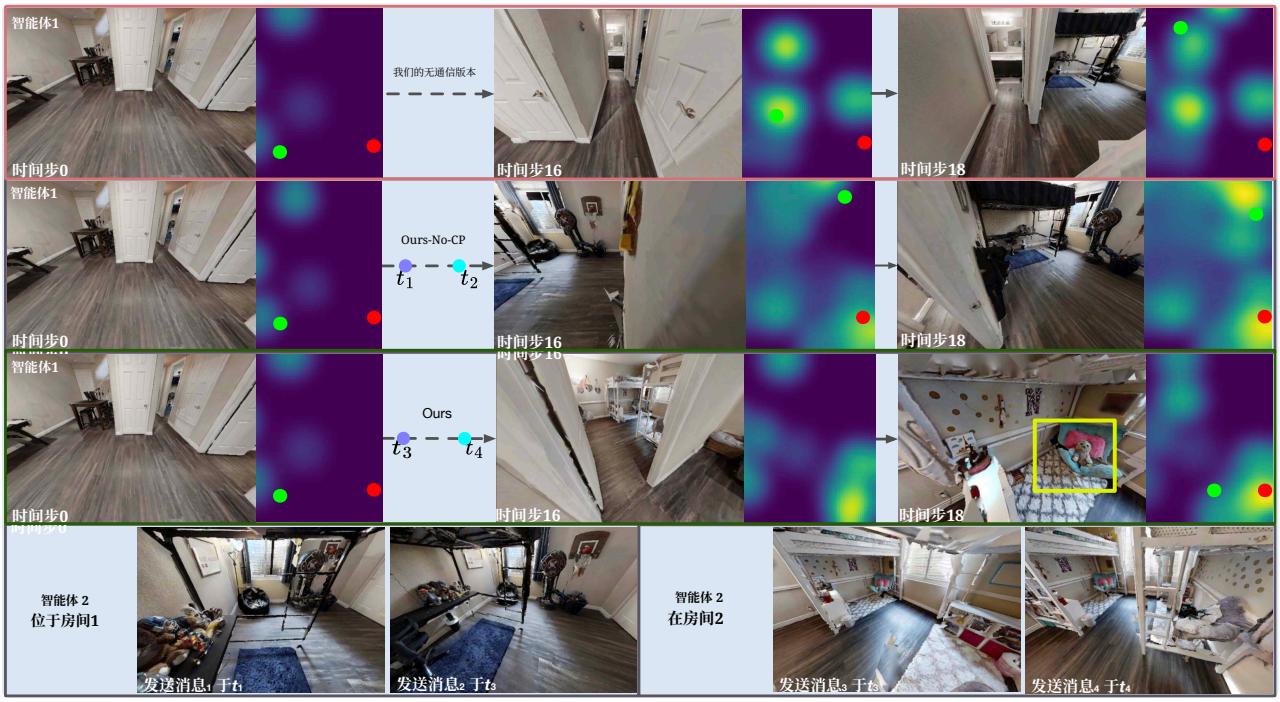

图4. 三种方法的机器人视图和全局SV地图对比。红色点表示目标的位置，绿色点表示机器人的位置。三种方法中的智能体均从相同的位姿出发。该场景中的问题是“红色熊靠垫在哪里？”对于“Ours-No-CP”和“我们的方法”，Robot2分别在不同时间探索相同的房间，并向Robot1发送消息。详细消息如下：MSG_1：我看到了一个篮板、玩偶和黑色椅子，它们可能与您的目标红色熊靠垫有关，其中玩偶可能是您在 $\{position1\}$处的目标。MSG_2：我在 $\{position2\}$处看到了可能与您的目标红色熊靠垫有关的玩偶。MSG_3：我看到了床，以及蓝色椅子上的红色枕头，它们可能与您的目标红色熊靠垫有关，其中蓝色椅子上的红色枕头可能是您在 $\{position3\}$处的目标。MSG_4：我看到了蓝色椅子上的红色枕头，它可能与您的目标红色熊靠垫有关，其中蓝色椅子上的红色枕头可能是您在 $\{position4\}$处的目标。

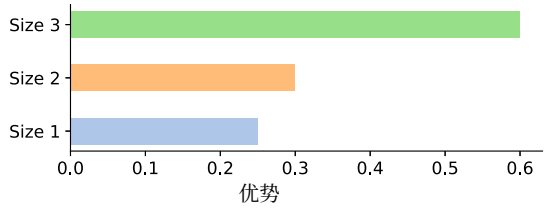

图5. 不同规模环境中的性能提升对比。“优势”表示“我们的方法”的 NTC 与 MMFBE 的 NTC 之差，计算方式为 Advantage $= \text{NTC}_{\text{Ours}} - \text{NTC}_{\text{MMFBE}}$。规模1表示场景面积为 $L \times W < 150\,\text{m}^2$。规模2表示 $150 \leq L \times W < 250\text{m}^2$。规模3表示 $L \times W \geq 250\,\text{m}^2$。

加以利用，从而促进任务完成。

基于保形预测的置信度校准。我们在图3(b)中展示了共形预测的重要性和有效性。没有共形预测时，“Ours-No-CP”基线的表现与“MMEuC”相近，这表明未经校准的输出会以无效或错误的信息误导智能体，导致其在相关性较低的区域进行探索，从而降低效率。相比之下，“我们的方法”采用校准通信，确保消息可靠，进而提升任务成功率。图3(c)显示，“Ours-Com-Control”配置发送消息的频率更高，但相关性较低，未能提升效率，这说明共享信息的质量比数量更为重要。“Ours-Com-Control”中的无效通信提供了更多但相关性较低的信息，导致性能更差、收敛更慢，最终成功率与“MMEuC”配置相近。

我们在图 4 中展示了一个具有代表性的测试案例。在“MMEuC”配置下，语义价值地图的模式扩散缓慢，无法覆盖重要区域，导致机器人找到目标物体的概率较低。这是因为机器人倾向于前往无关区域，而不是探索正确区域，从而降低了找到目标物体的可能性。尽管“Ours-No- CP”中的语义价值地图模式比“我们的方法”扩散得更快，但未经校准的通信会因校准不准确的消息所提供的误导性信息，导致机器人导航至错误的房间。相比之下，“我们的方法”能够有效更新自身机器人的语义价值地图，高效引导机器人，并展现出校准通信的重要性。

场景规模的影响。 图 5 进一步展示了场景规模如何影响“我们的方法”相较于 MMFBE 基线方法的性能。随着场景面积增大，我们的通信方法所带来的优势愈发明显。在更大的场景（规模 3）中，我们的方法平均比 MMFBE 高出 0.6 的 NTC提升，展现出信息共享增强与协同探索所带来的显著效率增益。随着场景规模增大，MMFBE 基线方法面临的挑战也更大，因为它依赖基于规则的非通信式探索，缺乏适应更大、更复杂环境的能力。我们的方法利用经过校准且相关的共享信息，持续提升探索效率和任务成功率，体现了其稳健性与可扩展性。

通信延迟的影响。我们在不同消息发送速度下评估任务成功情况，以分析通信延迟的影响。图 3(d) 显示，

较高的发送速度通过加快信息交换，使成功率在早期阶段更快上升。经过充分探索后，不同速度下的最终成功率几乎相同。值得注意的是，在所有消息发送速度下，我们的方法都优于 MMFBE 基线，证明了其在不同延迟条件下均有效。

可扩展性分析。我们通过扩展到三机器人团队来评估可扩展性。如图3(e)所示，与基线方法相比，“我们的方法”的成功率随 NTC 增长得更快。相反，由于无关信息增多，“Ours-No-CP”的早期成功率有所下降。随着时间推移，所有方法都能从增加智能体数量中获益，但我们的方法始终能更高效地完成任务，并且扩展时计算开销极小。

VI. 结论

为提高基于大语言模型的去中心化多智能体系统的通信效率，我们提出了一种基于大语言模型的高效自然语言通信框架。鉴于大语言模型会产生幻觉，且这些幻觉可能对整体效率产生负面影响，我们的框架引入保形预测来校准输出的置信度，从而过滤掉置信度过高的大语言模型生成的输出，使通信中的信息更加有效。实验结果表明，我们的方法显著提高了探索效率和总体成功率，尤其是在较大的场景中。在本研究中，我们通过高效通信，在双智能体和三智能体协作场景中开发了我们的方法。未来的工作将把我们的方法扩展到更复杂的部署场景，其中包含规模大得多的机器人团队和环境。

参考文献

[1] A. Das、S. Datta、G. Gkioxari、S. Lee、D. Parikh 和 D. Batra，《具身问答》，载于*IEEE* 计算机视觉与模式识别会议论文集，2018年，第 1–10 页。[2] A. Z. Ren、J. Clark、A. Dixit、M. Itkina、A. Majumdar 和 D. Sadigh，《探索直至置信：具身问答的高效探索》，载于机器人学：科学与系统，2024年。[3] A. Majumdar、A. Ajay、X. Zhang、P. Putta、S. Yenamandra、M. Henaff、S. Silwal、P. Mcvay、O. Maksymets、S. Arnaud 等，《OpenEQA：基础模型时代的具身问答》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2024年，第 16 488–16 498 页。[4] L. Yu、X. Chen、G. Gkioxari、M. Bansal、T. L. Berg 和 D. Batra，《多目标具身问答》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2019年，第 6309–6318 页。[5] J. Yao、X. Zhang、Y. Xia、A. K. Roy-Chowdhury 和 J. Li，《通过保形不确定性处理实现人群导航中的可泛化安全性》，载于机器人学习会议（*CoRL*），2025年。[6] H. Zhang、W. Du、J. Shan、Q. Zhou、Y. Du、J. B. Tenenbaum、T. Shu 和 C. Gan，《利用大语言模型模块化构建协作式具身智能体》，载于第十二届国际学习表征会议，2024年。[7] Y. Zhang、S. Yang、C. Bai、F. Wu、X. Li、X. Li 和 Z. Wang，《面向具身多智能体协作的高效大语言模型落地》，*arXiv*预印本 *arXiv:2405.14314*，2024年。[8] S. Rasal，《LLM Harmony：用于问题求解的多智能体通信》，*arXiv*预印本 *arXiv:2401.01312*，2024年。[9] S. J. Mielke、A. Szlam、E. Dinan 和 Y.-L. Boureau，《通过语言校准降低对话智能体的过度自信》，计算语言学协会会刊，2022年。[10] G. Hao、J. Wu、Q. Pan 和 R. Morello，《量化复杂自适应社会网络中大语言模型幻觉传播的不确定性》，《科学报告》，第 14 卷，第 1 期，第 16375 页，2024年。

[11] G. Shafer 和 V. Vovk，《保形预测教程》，《机器学习研究期刊》，第 9 卷，第 3 期，2008年。[12] A. Z. Ren、A. Dixit、A. Bodrova、S. Singh、S. Tu、N. Brown、P. Xu、L. Takayama、F. Xia、J. Varley 等，《寻求帮助的机器人：大语言模型规划器的不确定性对齐》，载于第七届机器人学习年度会议，2023年。[13] S. K. Ramakrishnan、A. Gokaslan、E. Wijmans、O. Maksymets、A. Clegg、J. M. Turner、E. Undersander、W. Galuba、A. Westbury、A. X. Chang 等，《Habitat-Matterport 3D 数据集（HM3D）：用于具身人工智能的 1000 个大规模三维环境》，载于第三十五届神经信息处理系统大会数据集与基准赛道（第二轮），2021年。[14] C.-M. Chan、W. Chen、Y. Su、J. Yu、W. Xue、S. Zhang、J. Fu 和 Z. Liu，《ChatEval：通过多智能体辩论实现更好的基于大语言模型的评估器》，载于第十二届国际学习表征会议，2024年。[15] X. Zhang、H. Qin、F. Wang、Y. Dong 和 J. Li，《LAMMA-P：利用由语言模型驱动的 PDDL 规划器实现可泛化的多智能体长时域任务分配与规划》，载于 IEEE 国际机器人与自动化会议（ICRA），2025年。[16] S. Hong、M. Zhuge、J. Chen、X. Zheng、Y. Cheng、J. Wang、C. Zhang、Z. Wang、S. K. S. Yau、Z. Lin 等，《MetaGPT：面向多智能体协作框架的元编程》，载于第十二届国际学习表征会议，2024年。[17] J. J. Horton，《作为模拟经济智能体的大语言模型：我们能从 Homo Silicus 中学到什么？》，美国国家经济研究局，技术报告，2023年。[18] W. Hua、L. Fan、L. Li、K. Mei、J. Ji、Y. Ge、L. Hemphill 和 Y. Zhang，《战争与和平（WarAgent）：基于大语言模型的世界大战多智能体仿真》，arXiv预印本 arXiv:2311.17227，2023年。[19] Y. Chen、J. Arkin、Y. Zhang、N. Roy 和 C. Fan，《利用大语言模型实现可扩展的多机器人协作：集中式系统还是分布式系统？》，载于 2024 年 IEEE 国际机器人与自动化会议（ICRA）。IEEE，2024年，第 4311–4317 页。[20] Z. Mandi、S. Jain 和 S. Song，《RoCo：利用大语言模型进行辩证式多机器人协作》，载于 2024 年 IEEE 国际机器人与自动化会议（ICRA）。IEEE，2024年。[21] J. Zhang、Y. Hou、R. Xie、W. Sun、J. McAuley、W. X. Zhao、L. Lin 和 J.-R. Wen，《AgentCF：用于推荐系统的自主语言智能体协作学习》，载于《2024 年 ACM Web 会议论文集》，2024年，第 3679–3689 页。[22] X. Zhou、H. Zhu、L. Mathur、R. Zhang、H. Yu、Z. Qi、L.-P. Morency、Y. Bisk、D. Fried、G. Neubig 等，《Sotopia：语言智能体社会智能的交互式评估》，载于第十二届国际学习表征会议，2024年。[23] J. Wang、G. He 和 Y. Kantaros，《利用保形预测为语言指令驱动的多机器人系统进行安全任务规划》，arXiv预印本 arXiv:2402.15368，2024年。[24] T. Kojima、S. S. Gu、M. Reid、Y. Matsuo 和 Y. Iwasawa，《大语言模型是零样本推理器》，《神经信息处理系统进展》，第 35 卷，第 22 199–22 213 页，2022年。[25] M. Fontana、G. Zeni 和 S. Vantini，《保形预测：理论与新挑战的统一综述》，《伯努利》，第 29 卷，2023年。[26] R. Luo、S. Zhao、J. Kuck、B. Ivanovic、S. Savarese、E. Schmerling 和 M. Pavone，《利用保形预测实现样本高效的安全保障》，《国际机器人研究期刊》，2024年。[27] B. Yamauchi，《自主探索的前沿方法》，载于《1997 年 IEEE 机器人与自动化计算智能国际研讨会论文集 CIRA’97：迈向机器人与自动化的新计算原理》。IEEE，1997年。[28] R. A. Newcombe、S. Izadi、O. Hilliges、D. Molyneaux、D. Kim、A. J. Davison、P. Kohi、J. Shotton、S. Hodges 和 A. Fitzgibbon，《KinectFusion：实时稠密表面映射与跟踪》，载于 2011 年第十届 IEEE 混合现实与增强现实国际研讨会。IEEE，2011年，第 127–136 页。[29] A. Zeng、S. Song、M. Nießner、M. Fisher、J. Xiao 和 T. F unkhouser，《3DMatch：从 RGB-D 重建中学习局部几何描述符》，载于 IEEE 计算机视觉与模式识别会议论文集，2017年，第 1802–1811 页。[30] X. Puig、E. Undersander、A. Szot、M. D. Cote、T.-Y. Yang、R. Partsey、R. Desai、A. Clegg、M. Hlavac、S. Y. Min 等，《Habitat 3.0：人类、虚拟化身与机器人的共同栖息地》，载于第十二届国际学习表征会议，2024年。[31] S. Karamcheti、S. Nair、A. Balakrishna、P. Liang、T. Kollar 和 D. Sadigh，《Prismatic 视觉语言模型：研究视觉条件语言模型的设计空间》，载于第四十一届国际机器学习会议，2024年。