arXiv:2611.30171v1 [cs.CV] 23 Nov 2026

视听重排

Ruohao Guo[1],Jiatong Y e[1], Chenwei Dong[1],Jinxing Zhou[2],Dantong Niu[3],Guangyao Li[1],Zhenbo L i[1]
[1]中国农业大学[2]穆罕默德·本·扎耶德人工智能大学
[3]加州大学伯克利分校

摘要

我们提出了视听重排（*AVR*）这一新任务：智能体接收口头请求，取回合适的物体，并将其交付给请求用户或指令指定的目的地。*AVR* 要求智能体理解口头请求，并通过声音定位说话者以完成交付。为支持这一任务，我们构建了一个数据集，涵盖三类查询：显式目标指定、相对于说话人的空间指示和隐式用户意图；这些查询包含八种子类型，涉及多样的物体选择与交付要求。此外，我们提出了一个框架，将运动感知空间音频模块集成到音视频大语言模型中，以编码口头请求中的空间线索，并在智能体移动时对其进行更新。我们进一步通过多任务指令微调，引入针对语音理解、空间音频感知和视觉定位的辅助监督，以提升动作预测能力。实验在已观测和未见环境中建立了基线性能，为未来具身智能和人机交互研究奠定了基础。代码将很快公开。

1. 引言

家用机器人必须结合所见与所闻进行推理，才能将人们日常的需求转化为物理动作。例如，当用户说“我渴了”时，机器人应推断用户需要饮料，找到并取来一杯水，再将水递给说话者。完成这类交互，机器人既要理解用户需要什么帮助，也要确定预期接收者的位置。语音提供关于用户意图的语义线索，以及关于说话者的空间线索；视觉观测则有助于物体定位和导航。这一场景将视听具身研究拓展至面向用户的物体重排，通过把隐含意图理解、说话者定位和物体操作纳入统一任务来实现这一目标。



图1. **视听重排概述**。给定一条口头请求，智能体必须（1）寻找并拾取合适的物体；（2）导航至请求者所在处并将物体交给对方。

这需要理解语音中要检索的内容，在视觉上确定目标物体的位置，并在多人中通过声音定位说话者，同时还需要进行导航和操作。

具身重排的最新进展使智能体能够将房间恢复到目标配置 [15],推断常识性物体放置 [13, 16], 并执行语言引导的物体检索与放置 [11, 14, 18]。与此同时，音视导航探索了智能体如何利用视觉观测和空间声学线索来定位声源 [3, 4, 17, 19]。这些研究方向涉及家庭辅助中相互补充的方面，但通常被分开研究。音视导航侧重于到达声源，而对检索物体并将其交付给说话者关注有限。具身重排基准测试主要通过文本指令或视觉目标状态来指定任务，而不要求智能体利用语音定位请求者。然而，在日常辅助中，这些能力紧密相连：同一句话既传达用户所需，也提供关于用户所在位置的声学线索。这就要求智能体将语音理解、说话者定位与视觉感知相结合，以检索合适的物体并将其交付给目标人物——这一设定在现有基准测试中仍未得到充分探索。

鉴于这些局限，我们提出视听重排（AVR），这是一项新任务：具身智能体根据口头请求检索合适的物体，并将其交付给请求用户，或放置在指定的支撑表面上（图1）。在有多人的陌生室内环境中，AVR要求智能体推断哪个物体能满足口头请求，并定位请求者。它带来了意图理解与空间推理方面密切相关的挑战：智能体需要推断哪个物体能满足请求，并定位说话者；说话者既可能是预期接收者，也可能是选择物体或交付位置时的空间指代。由于请求只说一次，智能体还需要记住说话者的位置，并在整个探索与操作过程中更新自身相对于说话者的位置。因此，AVR为研究智能体如何整合口头意图、空间感知与物理交互以提供家庭辅助，提供了一个任务设定。

为应对 AVR，我们基于 Qwen2.5-Omni [12] 开发了一个全模态框架，该框架融合了口语语言理解、运动感知空间听觉和闭环视觉技能预测。该框架利用初始口头请求，同时推断说话者的意图和位置。具体而言，运动感知空间（MAS）音频编码器从双耳语音中提取方向和距离线索。随着智能体移动，运动适配器根据智能体相对于初始聆听姿态的姿态更新这些空间特征，在无需重复语音的情况下保留相对于说话者的信息。空间投影器将适配后的特征映射为标记，语言模型则将其与指令、视觉观测和执行反馈相结合，以预测导航和操作技能。该设计将面向物体的探索与后续朝向说话者的导航联系起来。我们通过三阶段课程训练该框架：空间音频预训练培养定位能力，运动适配学习在智能体移动时更新空间特征，多任务指令微调则使这些表征与指令理解和动作预测保持一致。

为支持 AVR 研究，我们构建了一个仿真基准测试，其中包含口头请求、双耳空间音频，以及布置有家具的室内环境中的多个人。该基准测试涵盖明确目标、空间参照和隐式意图请求，训练与验证共包含 2,000 个回合，评估则包含 49 个场景中的 400 个回合。我们的方法在已观测场景和未观测场景中的总体成功率分别达到 20.5% 和 17.0%，在全部三种请求类型上均优于评估基线。消融研究表明，空间音频编码和运动适配能够提升总体任务成功率，而多任务监督与响应损失加权的结合使用则能进一步提高表现。更多分析刻画了在不同初始说话者距离下的成功执行情况以及存在并发环境声音时的表现。这些结果凸显了整合口语语义、空间听觉和物理交互以实现家庭辅助的潜力，同时指出声学干扰仍是一项关键挑战。

我们的主要贡献总结如下：

1）我们提出了视听重排（AVR）任务：智能体根据口头请求取回合适的物体，并将其送到多人中的请求者身边，或送至指令指定的目的地。该任务需要协同完成口头意图理解、说话人定位和物体操作。

2）我们开发了一种全模态基线框架，从单次初始话语中提取空间特征，并根据智能体的相对运动对这些特征进行调节，再将其与口头意图和视觉观测相结合，以预测可执行的导航和操作技能。

3）我们建立了一个仿真基准，涵盖已见和未见场景中的三种请求类型。实验对比了基线系统；消融实验评估了框架各组成部分及训练配置的贡献；此外，我们还分析了该框架对并发环境声音的敏感性。

2. 相关工作

2.1. 音视导航

音视导航 [3, 5]要求智能体结合视觉观测与声学线索，在陌生环境中定位并抵达发出声音的物体。为支持这项任务，Chen 等人 [3]推出了 SoundSpaces，这是一个用于在 3D 环境中进行几何声学仿真的平台。语义视听导航 [4, 17]将这项任务扩展到具有语义意义且间歇发声的物体，智能体必须推断目标物体的类别和位置，并利用过去的观测，即使声音停止后也能找到它。SAVN-CE [19] 借助 SoundSpaces 2 的实时声学渲染，将这项任务扩展至连续环境。0 [6], 这使智能体能够在 3D 场景中自由移动，同时接收空间和时间一致的视听观测。与这些导航任务不同，我们的任务要求智能体遵循口头指令，利用语音语义和声学定位，取回物体并将其交给说话者或送至指令指定的位置。

2.2. 具身重排

具身重排 [1] 旨在通过导航和物理交互，将环境转变为所需配置。在视觉房间重排中， [15], 智能体将物体位置和状态恢复至先前观测到的房间配置。其他研究 [13, 14, 16, 18]则探讨家庭环境重排和移动操作

，且目标状态没有明确的视觉参照。这些任务要么需要根据常识知识或用户偏好推断适当的物体放置位置，[13, 16],要么遵循明确指定预期结果的自然语言指令 [10, 14, 18]。前者侧重于判断物体应放置的位置并据此重新定位，后者则侧重于将指令与视觉观测建立对应，以指导物体搜索、导航和操作。除单智能体执行外，人机协作 [2, 11] 还引入了与共享环境中伙伴的协调。Habitat 3。0 [11] 通过人形机器人仿真和人在回路交互支持这一设定。其社交重排任务让机器人与人形机器人伙伴协作，高效地将物体移动到指定位置。与基于文本的指令跟随任务相比，我们的任务不仅涉及理解口头请求，还要求定位说话者并将其请求的物体递送给对方，更贴近人们在日常生活中寻求机器人协助的方式。

3. 视听重排

我们提出了视听重排（AVR）这一新颖任务：具身智能体遵循口头请求，取回合适的物体并将其交付给请求用户，或将其放置在指令指定的支撑表面上。在此任务中，口头请求既提供有关用户意图的语义线索（例如“给我拿个杯子”），也提供有关用户位置的空间线索。这一设定为具身重排带来了新的挑战 [1]。智能体需要将语音传达的意图与视觉观察到的物体对应起来，并利用空间音频线索，在多人环境中识别并定位请求者。

3.1. 仿真环境

我们基于 Habitat 3.0 [11],，使用配备家具的 HSSD [9] 室内场景构建仿真环境。每个环境包含多个来自 Habitat 人体资产的人形化身，以及一个配备自我中心相机、建模为 Fetch移动操作机器人的具身智能体。其中一个化身被指定为请求用户，其余化身则充当干扰项。我们使用 Qwen3-TTS [8] 为每个请求合成语音，并将生成的波形与请求用户关联。为模拟智能体接收到的音频，我们使用集成于 Habitat-Sim 中的 SoundSpaces 2.0 [6] 声学渲染器。具体而言，我们将声源放置在请求用户的头部，并使听者的位置和朝向与智能体相机保持一致。随后，我们将合成的语音波形与模拟的双耳房间脉冲响应进行卷积，生成空间化语音，从而将请求的语义内容与说话者位置的声学线索结合起来。

3.2. 任务分类体系与形式化表述

我们定义了三种查询类型和八种子类型。**显式目标** 查询会直接指出物体类别，包括类别 查询和实例查询。前者接受任意合格实例（例如“给我拿一个杯子”），后者则通过房间或支撑家具来指明特定实例（例如“给我拿厨房台面上的杯子”）。**空间指代** 查询会指定相对于请求用户的空间关系，包括物体查询，要求取回最近的合格物体（例如“给我拿离我最近的杯子”）；以及目的地查询，要求将物体放置在最近的合格支撑物上（例如“把杯子放到离我最近的桌子上”）。这些查询要求识别请求说话者作为空间指代，并比较候选位置相对于该人员的位置。**隐含意图**查询表达某种需求或预期活动，但不指明目标物体，因此需要智能体推断合适的类别。这类查询涵盖四个领域：饮食，包括备餐（例如“我渴了”）；家庭，涉及布置和装饰（例如“给我拿件东西来装饰这个空间”）；工作，涉及工作、学习和整理（例如“我需要为学习活动做准备”）；以及休闲，涉及玩耍和娱乐（例如“我很无聊，想找点东西玩”）。每个隐含意图查询都与一组可接受物体类别相关联，因此根据基准测试的基于类别的标准，不同物体可以满足同一请求。

每个回合都设定一个室内场景、一位指定用户发出的口头请求，以及可接受物体和交付目标。智能体和化身会被初始化在场景可导航区域内随机采样且无碰撞的位置。请求用户仅在回合开始时说出一次请求，任务执行期间不再发言。只有当智能体取回一个可接受物体，并在回合预算内完成所需交付时，该回合才算成功。若要将物体交付给用户，智能体必须持有该物体，处于请求用户的交接范围内，并成功完成该物体的交接动作。接近其他化身不满足这一条件。对于空间指代的目的地子类型，智能体必须将物体放到距离请求用户最近的合格支撑物上，并确保物体在回合结束时仍被稳定支撑。因此，成功既要求正确选择物体，也要求将其交付给正确的接收者或支撑物。

4. 方法

我们提出了一种用于视听重排的基线框架，将语音理解和运动感知空间音频感知与视觉定位相结合，以实现闭环技能预测。我们首先介绍模型架构，重点阐述所提出的运动感知空间音频

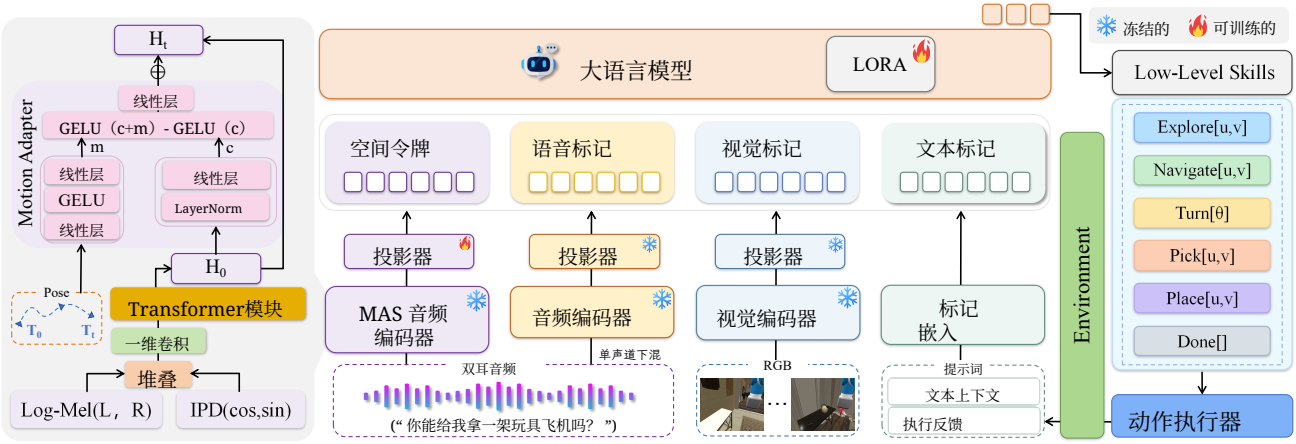


图 2. **基线框架概览**。左侧面板展示了运动感知空间（MAS）音频编码器的细节。原始音频编码器保留了语音理解能力，而并行的 MAS 音频编码器则提取双耳空间线索，并根据智能体的运动对其进行更新。轻量级投影器将这些特征映射到语言模型的标记空间，在此处，它们与语音、视觉和文本标记相结合，以预测可执行技能。动作执行器与环境交互，并返回执行反馈，以支持闭环决策。

编码器（MAS）（第4.1节），随后介绍用于空间感知、运动适配和具身决策的三阶段训练流程（第4.2节）。

4.1. 模型架构

图 2展示了总体架构。我们基于 Qwen- Omni 模型 [12], 构建，同时保留其视觉编码器、单声道音频编码器和 Thinker语言模型骨干网络。我们引入了一个 MAS 音频编码器，用于从双耳音频中提取空间线索，并通过集成的运动适配器，根据智能体的运动对这些线索进行更新。轻量级投影器将其输出映射为空间令牌，供语言模型使用。与单声道分支捕捉的语音内容相辅相成，这些令牌表示说话者相对于智能体的方向和距离。语言模型将空间、语音、视觉和文本标记相结合，以预测可执行技能。

空间音频特征编码。 为了捕捉与说话者方向和距离相关的声学线索，MAS 音频编码器首先将双耳音频编码为空间特征。给定一段双耳录音，我们将输入特征构造为 Stack$[M_L, M_R, \mathcal{B}(\cos\Delta\phi)\ \mathcal{B}(\sin\Delta\phi)]$, , 其中 M_L 和 M_R 分别是左耳和右耳的对数梅尔特征，$\Delta\phi$ 表示双耳相位差， $\mathcal{B}$ 表示加权幅度的梅尔频带聚合。前端保留绝对信号电平和双耳差异，以保留声学定位线索。这些特征经时间卷积和 Transformer 主干网络处理，随后通过自适应平均池化将输出序列压缩为 16 个空间特征向量 H_0。它们表示初始接收器姿态下的声学观测，并作为后续运动适配的参考。

运动适配。 当智能体从录制姿态 T_0 移动到当前接收器姿态 T_t时，说话者相对于接收器的位置会发生变化。假设说话者保持静止，运动适配器会 H_0 利用从 (T_0, T_t) 得出的相对旋转和平移更新初始声学特征，生成H_t 以当前接收器姿态为条件的空间特征。如图 2左图所示，运动分支将相对姿态编码为运动嵌入 m，而声学分支对 H_0 应用 LayerNorm 和线性投影，以获得条件特征 c。两个分支通过残差瓶颈 $\mathrm{GELU}(c+m)-\mathrm{GELU}(c)$进行交互，其输出经过线性投影后与 H_0 相加，以生成 H_t。每个 H_t 都直接根据 H_0 计算，而不是递归地根据先前的特征计算，从而在无需说话者重复请求的情况下，保持对初始话语进行运动条件化的空间表示。

技能预测。 为了将口头请求转化为可执行技能，语言模型结合空间音频特征、语音内容、视觉观测和文本上下文。可训练的多层感知机投影器将经过适配的空间特征 H_t 映射为语言模型嵌入空间中的空间令牌。单声道分支从通道平均双耳音频中编码语音内容，而视觉观测则提供场景和物体信息。文本上下文包括任务提示、保留的初始请求文本表示以及执行反馈。在后续步骤中，保留的请求和更新后的空间特征无需重复语音即可支持决策，而当前观测和选定的历史视图则支持物体定位和说话人识别。在每个决策步骤，语言模型会预测一项技能

a_t ， 包括探索 $[u, v]$、导航 $[u, v]$、转向 $[\theta]$、拾取 $[u, v]$、放置 $[u, v]$和完成 $[\,]$。对于基于坐标的技能，$(u,\ v)$指定当前自我中心图像中的目标位置，并归一化到范围 $[0, 100]$, ; 而 θ指定离散的旋转角度。放置技能会将手持物体交给指定的接收者或放置到支撑物上，而完成 $[\,]$ 则终止执行。低层控制器利用模拟器深度信息、相机位姿、导航网格和实例观测来执行技能，并返回观测和执行反馈，以进行闭环预测。

4.2. 训练流程

我们采用三阶段训练流程：空间音频预训练、运动适应预训练和多任务指令微调。阶段 1 中，我们预训练 MAS 音频编码器的双耳特征编码路径，使其预测说话者相对于接收者的方向和距离。阶段 2 中，我们冻结预训练的声学编码路径，并训练运动适配器，根据接收者的相对运动更新初始空间特征。我们保留阶段 1 的定位目标，并在当前接收器坐标系中表示目标方向和距离。阶段 3 中，我们联合优化空间投影器和语言模型 LoRA 参数 [7], , 同时冻结 MAS 音频编码器、原始音频和视觉编码器以及基础语言模型的权重。训练混合数据包含十项指令—响应任务：语音转录、声源定位、目标类别提取、视觉目标定位、指令—目标匹配、当前观测中的说话者定位、历史视图中的说话者定位、场景／家具可见性预测、主策略预测和交付策略预测。这些任务提供互补监督，涵盖听觉理解、视觉感知、视听空间关联和具身决策，支持技能预测，并将物品交付给请求用户或指令指定的目的地。

5. 实验

5.1. 实验设置

数据集构建。 我们从 38 个开发场景中的 2,000 个 AVR 基础回合构建指令微调数据，并按场景划分为 34 个场景中的 1,790 个训练回合和 4 个场景中的 210 个验证回合。基于规则的专家利用模拟器特权物体与说话者状态来选择目标，并生成基于当前图像的像素动作。我们保留所有通过筛选的基础轨迹中的决策，并通过验证成功的示例和重采样来补充代表性不足的动作。十任务混合数据集包含 15 万个训练样例和 1.3 万个验证样例。评估使用另外 400 个回合：其中 200 个来自 38 个开发场景（已观测），另外 200 个来自这批场景之外的 11 个场景

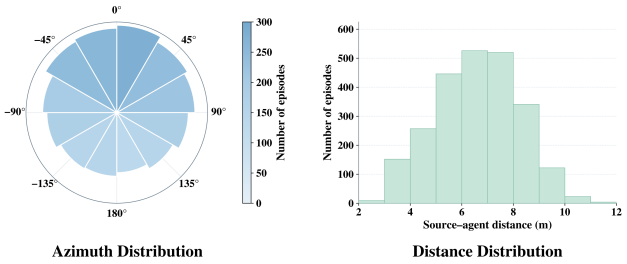


图3。**AVR 回合中初始说话者方位角与声源—智能体欧几里得距离的分布**。**左侧**：极坐标直方图，展示初始说话者相对于智能体的空间分布。**右侧**：直方图，展示声源与智能体之间的距离分布，范围为 2 至 12 米。

（未观测）。此处，“已观测”同时包括指令微调训练场景和验证场景。两个评估子集均遵循第 3.2 节的设置，其中显式、关系型和隐含请求的比例为 4:3:3。图3 展示了 2,400 个回合中说话者的初始方位角，以及声源到智能体的欧几里得距离。方位角覆盖整个角度范围；距离主要集中在 5–9 米，峰值区间为 6–8 米。

评估指标。我们报告成功率（SR）和平均模拟器步数（步数）。成功要求满足请求中指定的所有物体和目的地约束。ESR、RSR 和 ISR 分别表示显式、关系型和隐式请求的 SR，OSR 表示所有请求类型的总体 SR。步数是所有回合（包括失败和超时回合）的平均执行长度，每回合的步数预算为 5,000 步。由于提前终止也会计入该平均值，步数反映的是总体执行长度，而非成功完成任务时的效率，因此应结合 SR 一并解读。故仅凭较低的步数值并不一定意味着任务表现更好。为区分目标选择与交付，我们还报告目标选择准确率（TSA）和条件递送成功率（CDSR）。TSA 衡量获取的物体是否符合请求，其计算方式为：在获取了任意物体的回合中，获取到正确目标的回合所占比例。未获取物体的回合不计入 TSA。CDSR 衡量获取正确目标后的交付成功情况，其计算方式为：在获取到正确目标的回合中，完全成功的回合所占比例。交付包括向正确接收者进行交接，以及将物体稳定放置在指定支撑面上。

基线。 我们在相同的任务设置和评估协议下，将我们的方法与四种基线进行比较。(1) 随机：一种随机策略，均匀采样六种高层动作类型：探索、导航、转向、抓取、表面放置和交接，并随机采样动作参数。当预言机检测到任务完成时停止（完美停止）。(2) Omni-VLM:

表 1. 在已观测和未观测环境中对音视重排基准进行评估 s.

	Seen					未观测				
	ESR ↑	RSR ↑	ISR ↑	OSR ↑	步数 ↓	ESR ↑	RSR ↑	ISR ↑	OSR ↑	步数 ↓
随机	0.0	0.0	0.0	0.0	5000.0	0.0	0.0	0.0	0.0	5000.0
Omni-VLM	0.0	0.0	0.0	0.0	4927.6	0.0	0.0	0.0	0.0	4921.3
模块化语言智能体	11.3	1.7	3.3	6.0	3729.7	12.5	1.7	5.0	7.0	3851.2
LLM规划器	20.0	0.0	15.0	12.5	2973.9	17.5	1.7	13.3	11.5	2867.8
Ours	**25.0**	**16.7**	**18.3**	**20.5**	**2568.5**	**18.8**	**15.0**	**16.7**	**17.0**	**2685.5**

一个预训练的 Qwen2.5-Omni 模型，无需特定任务微调即可预测动作。它采用与我们的方法相同的观测输入、历史窗口和动作接口。(3) 模块化语言智能体：一种模块化智能体，使用 Qwen2.5-7B-Instruct 进行规划、Whisper 进行语音识别，并使用经过学习的音频定位模块进行声源估计。GroundingDINO 和 SAM 处理 RGB-D 观测，以构建在线语义地图。智能体从没有场景地图的状态开始，根据转录指令和观测到的地图选择高层动作。(4) LLM 规划器：一种基于 LLM 的规划器，以真实场景图初始化，该场景图编码房间归属关系和物体与支撑物关系，并在执行过程中更新。Whisper 提供指令转录文本，音频定位模块则估计声源位置。

实现细节。 音频以 16 kHz 采样。空间编码器包含 32 层 Transformer，隐藏维度为 1280，输出 2048 维特征。瓶颈维度为 128 的运动适配器使用模拟器提供的接收器相对位姿更新这些特征，然后将其投影为 16 个维度为 3584 的空间令牌，供 Qwen2.5-Omni-7B 使用。阶段 1 采用场景不重叠的训练/验证划分，包含 33/2 个场景。阶段 2 采用场景不重叠的训练/验证划分，包含 34/4 个场景。两个划分均排除全部 11 个未见评估场景。阶段 1 训练 10 个 epoch，使用余弦方向损失和平滑的 L_1 对数距离损失，骨干网络/预测头的学习率为 $10^{-5}/10^{-4}$，权重衰减为 0.01。阶段 2 使用 AdamW，在冻结特征上训练运动适配器和定位头，学习率为 $10^{-3}/10^{-4}$，批量大小为 256，且不使用权重衰减。余弦退火采用最小学习率 10^{-5}；我们选用第 40 个 epoch 的检查点。阶段 3 仅在十任务混合数据集上训练投影器和 LoRA 参数。响应令牌交叉熵权重分别为：动作名称和参数为 2.0，Assigned! 为 0.25，其他为 1.0。AdamW 的学习率为 10^{-5}，权重衰减为 0.1，并采用余弦衰减和 3% 的预热。LoRA 使用秩 16、$\alpha = 32$ 和 0.05 的丢弃率。阶段 3 使用 6 张 A100 GPU，以 BF16 精度训练 1 个 epoch，每张 GPU 的批量大小为 1，梯度累积步数为 8（有效批量大小为 48）。

表2. 我们的方法在已见与未见划分中的八种AVR查询子类型上的成功率（%）。

	子类型	Seen	未见
显式目标	类别	30.0	23.3
	实例	22.0	16.0
空间指代	物体	30.0	26.7
	目的地	3.3	3.3
隐含意图	饮食	26.7	20.0
	Home	13.3	13.3
	Work	20.0	20.0
	休闲	13.3	13.3
总体		20.5	17.0

5.2. 与基线的比较

表1 比较了 AVR 上的基线表现。零样本 Omni-VLM 策略未能完成任何任务，这表明仅凭预训练的全模态能力，不足以在该协议下实现可靠执行。模块化语言智能体借助我们的声源定位器取得了非零成功率。使用真实房间归属关系和物体与支撑物关系时，LLM 规划器的整体表现更好，不过其较低的 RSR 可能部分源于场景图缺少相对于说话者的空间信息。我们的方法在已观测和未见场景中的 OSR 分别达到 20.5% 和 17.0%，分别比 LLM 规划器高出 8.0 和 5.5 个百分点。在两个划分中，我们的方法在全部三种请求类型上的 SR 均最高，平均执行步数也最低，分别为 2568.5 和 2685.5。这些结果表明，在检索合适物体并将其送至请求者或指令指定的目的地方面，与评估基线相比，我们的框架取得了更高的任务成功率。

表 2 按请求子类型细分性能。与从命名类别中检索任意物体相比，在两个划分中，从特定位置的物体中进行检索都更加困难。相对于说话者的检索在已观测/未观测场景中分别达到30.0%/26.7%的SR，而最近支撑物放置在两个划分中均仅达到3.3%，这凸显出相对于说话者的放置仍是一个挑战。对于隐含请求，饮食在两个划分中分别达到26.7%/20.0%，工作与学习在两个划分中均为20.0%，家居装饰和休闲各为13.3%，显示出——

了四个领域均取得成功，但不同用户需求之间的表现各异。

5.3. 定性结果

图 4展示了在未观测环境中的两条 AVR 轨迹，蓝色和橙色路径分别表示物体搜索和说话者搜索。在明确请求示例（上图）中，说话者请求卧室桌子上的一个玩偶。指定的类别和位置限定了有效目标，因此需要进行更长时间的探索才能完成检索。随后，智能体携带玩偶导航至请求者身边。在隐式请求示例（下图）中，说话者描述了儿童的玩耍活动，但未指定物体。智能体推断玩具可以满足这一需求。由于房间内有多个玩具，任何合适的玩具都能满足请求，因此物体搜索所需时间更短。由此可见，这两个请求对物体选择施加了不同的约束。两个场景中都包含多个人，因此智能体需要将请求者与其他个体区分开来。由于请求只被说出一次，智能体在整个探索和检索过程中持续保留并更新相对于说话者的空间信息。编号观测结果展示了物体获取过程以及随后接近说话者的过程。总体而言，这些示例说明了请求的具体程度如何影响探索，以及我们的框架如何将口头意图与物体检索和面向说话者的导航联系起来。

5.4. 消融研究

表3 考察了空间音频编码、运动适配以及组合训练策略各自的贡献。前三种配置使用相同的策略训练示例和响应交叉熵损失。在单声道基线模型中加入冻结的预训练空间音频编码器后，OSR 从 3.5% 提升至 7.5%，TSA 从 34.4% 提升至 46.5%，CDSR 则从 15.9% 提升至 25.0%。这些提升表明，空间音频为物体检索和交付提供了有用信息，这与说话者既是空间参照点、又可能是接收者的角色相符。进一步加入冻结的预训练运动适配器后，OSR 提升至 11.0%，CDSR 提升至 34.4%。尽管 TSA 从 46.5% 略微降至 43.5%，总体任务成功率仍持续提高，这表明在结合智能体的相对运动进行条件化后，初始空间特征对任务执行更有用。CDSR 随之提高，这与在探索和操作过程中保留相对于说话者的信息这一动机相符，并有助于智能体离开最初的聆听位置后继续完成交付。

我们的完整模型结合了十任务混合策略学习与辅助监督，并采用响应损失加权，在所评估的配置中取得了最高的 OSR、TSA 和 CDSR

表3. 未见划分上的消融研究。各组件依次添加到单声道基线模型中。我们的方法进一步纳入多任务监督和响应损失加权。

	OSR ↑	TSA ↑	CDSR ↑	步数 ↓
单声道	3.5	34.4	15.9	2621.9
+ 空间音频	7.5	46.5	25.0	2926.4
+ 运动适配器	11.0	43.5	34.4	2752.4
Ours	17.0	47.6	49.3	2685.5

分别达到 17.0%、47.6% 和 49.3%。辅助任务监督指令理解、视觉定位和说话者定位，而响应损失加权则侧重于动作名称和参数。综合这些改进，OSR 比运动适配配置提高了 6.0 个百分点，TSA 和 CDSR 也均有所提升。作为补充性的空间诊断，应用于运动适配特征的定位头在当前接收器坐标系中的环绕方位角平均绝对误差为 19.41°，在阶段 2 验证子集上的水平距离平均绝对误差为 1.058 米。这些结果从不同方面证明了空间定位的准确性，并表明组合训练策略有助于任务完成。

表 4. 未见划分上对空间干扰项的鲁棒性。

	OSR ↑	TSA ↑	CDSR ↑	步数 ↓
干净	17.0	47.6	49.3	2685.5
干扰项	6.5	37.2	27.1	2738.1

5.5. 对空间干扰项的鲁棒性

为评估对声学干扰的鲁棒性，我们在距说话者至少2米、距智能体1米的可导航位置引入并发的环境声音，其有功功率与目标信号相匹配。在相同的200个未见回合中，干扰源将OSR从17.0%降至6.5%，将TSA从47.6%降至37.2%，将CDSR从49.3%降至27.1%，而步数增至2738.1（表4）。TSA的下降表明，在成功完成物体获取的回合中，目标选择更差；而较低的CDSR表明，即使已获取正确目标，交付仍然困难。因此，声学干扰同时影响目标获取及后续交付，凸显了当前框架在多源环境中的局限性。



图4。**未见环境中的AVR轨迹**。上方和下方的示例分别对应明确请求和隐含请求。每个示例展示了俯视场景图、占据栅格地图，以及探索（1）、目标获取（2）和接近说话者（3）过程中的自我中心观测。蓝色和橙色路径分别表示物体搜索和声源搜索。路径沿导航网格连接记录的决策位置。部分视图附有模拟器渲染快照以供说明。

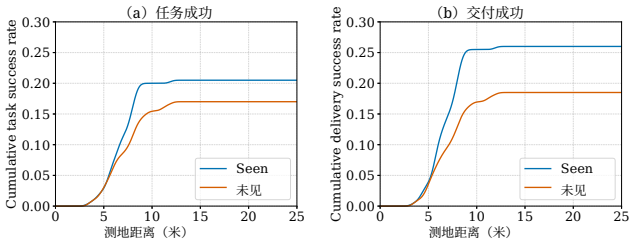


图5。**不同初始测地距离下的累计成功率**。（a）完整AVR任务成功率。（b）将物品交接给正确的接收者。蓝色和橙色分别表示已观测和未见回合。距离基于静态场景导航网格计算。

5.6. 初始距离的影响

我们的数据集将智能体和说话者初始化在2–25米的测地距离处。图5 报告了相对于初始测地距离的累计成功率，其中考虑了障碍物和房间布局。在所评估的距离分布内，任务成功与接收者交接成功率均在约10–12米处接近其总体比率，更大距离上几乎没有额外成功。该模式在已见与未见划分上均出现，并与远距离说话者定位和导航需求的增加相一致。更大的间距及其间的障碍物会削弱直达路径语音，使空间线索更不可靠，而更长的导航路线则会带来更多执行错误的机会。由于请求仅说出一次，智能体还必须维持与说话者相对的in-

在这些较长轨迹中的信息。

6. 结论

我们提出了 AVR，这是一项新任务，将口头请求与多人室内环境中的物体检索和交付联系起来。我们的全模态框架结合 MAS 音频编码器与运动适配器，从单条初始话语中提取空间线索，并在智能体移动时持续更新这些线索。这些线索与口头意图和视觉观测相结合，以支持闭环导航和操作。实验表明，我们的框架优于评估基线，在针对说话者的相对位置请求上尤其表现突出。结果还揭示了在未见环境中可靠执行任务，以及应对并发声学干扰方面仍面临的挑战。未来工作将探索多源音频理解，以及向真实世界移动操作机器人的迁移，以拓展该框架在实际家庭辅助中的应用，将人们口头表达的需求与物理动作联系起来。

大语言模型使用声明

作者仅使用大语言模型来提升论文的语言表达和清晰度。这些工具未用于研究构思、实验设计、数据分析或得出科学结论。作者已审阅并核实所有修订内容，并对本文内容负全部责任。

参考文献

[1] Dhruv Batra、Angel X. Chang、Sonia Chernova、Andrew J. Davison、Jia Deng、Vladlen Koltun、Sergey Levine、Jitendra Malik、Igor Mordatch、Roozbeh Mottaghi、Manolis Savva 和 Hao Su。重排：具身人工智能的一项挑战。*arXiv*预印本 *arXiv:2011.01975*，2020年。2，3[2] Matthew Chang、Gunjan Chhablani、Alexander Clegg、Mikael Dallaire Cote、Ruta Desai、Michal Hlavac、Vladimir Karashchuk、Jacob Krantz、Roozbeh Mottaghi、Priyam Parashar、Siddharth Patki、Ishita Prasad、Xavier Puig、Ak- shara Rai、Ram Ramrakhya、Daniel Tran、Joanne Truong、John Turner、Eric Undersander 和 Tsung-Yen Yang。PARTNR：具身多智能体任务中的规划与推理基准测试。发表于第十三届国际学习表征会议，2025年。3[3] Changan Chen、Unnat Jain、Carl Schissler、Sebastia Vi- cenc Amengual Gari、Ziad Al-Halah、Vamsi Krishna It hapu、Philip Robinson 和 Kristen Grauman。SoundSpaces：三维环境中的音视导航。发表于欧洲计算机视觉会议（*ECCV*），2020年。1，2[4] Changan Chen、Ziad Al-Halah 和 Kristen Grauman。语义音视导航。发表于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第15516–15525页，2021年。1，2

[5] Changan Chen、Sagnik Majumder、Ziad Al-Halah、Ruohan Gao、Santhosh Kumar Ramakrishnan 和 Kristen Grauman。学习为音视导航设置路径点。载于国际学习表征会议（ICLR），2021。2[6]Changan Chen、Carl Schissler、Sanchit Garg、Philip Kobernik、Alexander Clegg、Paul Calamia、Dhruv Batra、Philip Robinson 和 Kristen Grauman。SoundSpaces 2.0：用于视觉声学学习的仿真平台。载于神经信息处理系统进展，2022。2，3[7] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。LoRA：大语言模型的低秩适配。载于国际学习表征会议（ICLR），2022。5[8] Hangrui Hu、Xinfa Zhu、Ting He、Dake Guo、Bin Zhang、Xiong Wang、Zhifang Guo、Ziyue Jiang、Hongkun Hao、Zishan Guo、Xinyu Zhang、Pei Zhang、Baosong Yang、Jin Xu、Jingren Zhou 和 Junyang Lin。Qwen3-TTS技术报告。arXiv预印本 arXiv:2601.15621，2026。3[9]Mukul Khanna、Yongsen Mao、Hanxiao Jiang、Sanjay Haresh、Brennan Shacklett、Dhruv Batra、Alexander Clegg、Eric Undersander、Angel X. Chang 和 Manolis Savva。Habitat合成场景数据集（HSSD-200）：物体目标导航中三维场景规模与真实感权衡的分析。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第16384–16393页，2024。3[10] Aishwarya Padmakumar、Jesse Thomason、Ayush Shrivastava、Patrick Lange、Anjali Narayan-Chen、Spandana Gella、Robinson Piramuthu、Gokhan Tur 和 Dilek Hakkani-Tur。TEACh：能够对话的任务驱动具身智能体。载于AAAI人工智能会议论文集，第2017–2025页，2022。3[11] Xavier Puig、Eric Undersander、Andrew Szot、Mikael Dallaire Cote、Tsung-Yen Yang、Ruslan Partsey、Ruta Desai、Alexander Clegg、Michal Hlavac、So Yeon Min、Vladim´ır Vondru˘s、Theophile Gervet、Vincent-Pierre Berges、John Turner、Oleksandr Maksymets、Zsolt Kira、Mrinal Kalakrishnan、Jitendra Malik、Devendra Chaplot、Unnat Jain、Dhruv Batra、Akshara Rai 和 Roozbeh Mottaghi。Habitat 3.0：人类、化身与机器人的共居环境。载于第十二届国际学习表征会议，2024。1，3[12] Qwen团队。Qwen2.5-omni技术报告。arXiv预印本 arXiv:2503.20215，2025。2，4[13] Gabriel Sarch、Zhaoyuan Fang、Adam W. Harley、Paul Schydlo、Michael J. Tarr、Saurabh Gupta 和 Katerina Fragkiadaki。TIDEE：利用视觉语义常识先验整理新颖房间。载于欧洲计算机视觉会议（ECCV），2022。1，2，3[14] Mohit Shridhar、Jesse Thomason、Daniel Gordon、Yonatan Bisk、Winson Han、Roozbeh Mottaghi、Luke Zettlemoyer 和 Dieter Fox。ALFRED：用于理解日常任务落地指令的基准测试。载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，第10740–10749页，2020。1，2，3

[15]Luca Weihs、Matt Deitke、Aniruddha Kembhavi 和 Roozbeh Mottaghi。视觉房间重排。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2021。1，2[16]Jimmy Wu、Rika Antonova、Adam Kan、Marion Lepert、Andy Zeng、Shuran Song、Jeannette Bohg、Szymon Rusinkiewicz 和 Thomas Funkhouser。TidyBot：利用大语言模型提供个性化机器人辅助。《自主机器人》，2023。1，2，3[17] Zeyuan Yang、Jiageng Liu、Peihao Chen、Anoop Cherian、Tim K. Marks、Jonathan Le Roux 和 Chuang Gan。RILA：用于零样本语义音视导航的反思与想象型语言智能体。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2024。1，2[18] Sriram Yenamandra、Arun Ramachandran、Karmesh Yadav、Austin S. Wang、Mukul Khanna、Theophile Gervet、Tsung-Yen Yang、Vidhi Jain、Alexander Clegg、John M. Turner、Zsolt Kira、Manolis Savva、Angel X. Chang、Devendra Singh Chaplot、Dhruv Batra、Roozbeh Mottaghi、Yonatan Bisk 和 Chris Paxton。HomeRobot：开放词汇的移动操作。载于第七届机器人学习会议论文集，第1975–2011页。PMLR，2023。1，2，3[19] Yichen Zeng、Hebaixu Wang、Meng Liu、Yu Zhou、Chen Gao、Kehan Chen 和 Gongping Huang。连续环境中的语义音视导航。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第22369–22379页，2026。1，2