# 学习设置音频-视觉导航的航点

# 摘要

在音频-视觉导航中，一个智能体通过视觉和听觉在复杂未映射的3D环境中移动，以找到声源（例如，在另一个房间响起的电话）。现有模型学习在智能体运动的固定粒度上采取行动，并依赖于对音频观察的简单循环聚合。我们引入了一种基于强化学习的音频-视觉导航方法，具有两个关键的新元素：1）动态设置并在导航策略内端到端学习的航点，以及2）提供结构化、空间基准的声学记忆，记录了智能体在移动过程中听到的内容。这两个新的思想都充分利用了音频和视觉数据的协同作用，揭示了未映射空间的几何结构。我们在两个具有挑战性的真实3D场景数据集Replica和Matterport3D上展示了我们的方法。我们的模型在技术水平上取得了显著的提升，实验证明学习视觉、听觉和空间之间的联系对音频-视觉导航至关重要。

# 1 引言

智能机器人必须能够在物理世界中高效移动。除了几何图和规划之外，实体人工智能领域的研究表明，智能体可以学习绘制地图和导航。通过直接从自我中心的图像中感知，它们共同学习空间记忆和导航策略，以便在新颖的未映射3D环境中快速到达目标位置（Gupta等人，2017b;a; Savinov等人，2018; Mishkin等人，2019）。高质量的模拟器加速了这个研究方向，以至于在模拟中学到的策略有时可以成功地转化为在现实世界中部署的机器人智能体（Gupta等人，2017a; Müller等人，2018; Chaplot等人，2020b; Stein等人，2018）。

目前许多工作集中在由已告知目标位置的PointGoal智能体进行的视觉导航上（Gupta等人，2017a; Sax等人，2018; Mishkin等人，2019; Savva等人，2019; Chaplot等人，2020b）。然而，在最近引入的AudioGoal任务中，智能体必须使用视觉和听觉感知来穿越未映射的3D环境，以找到发出声音的对象，而不知道其位置在哪里（Chen等人，2020; Gan等人，2020）。作为一个学习问题，AudioGoal不仅从认知和神经科学（Gougoux等人，2005; Lessard等人，1998）中获得了强烈的动机，而且在现实世界中具有引人注目的重要意义：某处楼上电话响起；有人在另一个房间呼救；狗在门口挠门要出去。

在学习导航时，音频-视觉输入应扮演什么角色？有两种现有策略。一种采用深度强化学习来学习生成基于两种模态的步骤性行动（向右转，向前移动等）的导航策略（Chen等人，2020）。这种方法统一了感知模态的优点，但在学习进行长序列的单个局部动作时可能效率较低。另一种方法分离了模态——将音频流视为信号目标位置的信标，然后使用视觉映射器规划到达该位置的路径（Gan等人，2020）。这种策略具有模块化的优点，但限制了音频在定位目标方面的作用。此外，现有的两种方法都对应该在多大粒度上预测动作做出了明确假设，要么是每一步的局部预测（0.5到1米）（Chen等人，2020），要么是全局的最终目标位置（Gan等人，2020）。

我们引入了一种新的AudioGoal导航方法，其中代理预测非短视的动作，并具有自适应粒度。我们的关键见解是学习设置音频-视觉航点：代理根据其音频-视觉观察和部分地图动态设置中间目标位置，并且以端到端的方式学习导航任务。直观地说，直接从远处定位远处的声源通常很困难，但可以更容易地确定一般方向（从而是可导航的路径），沿该路径可以更靠近该源。请参见图1。在未映射环境中，音频和视觉模态都对识别航点至关重要。 相反，仅使用视觉输入选择的子目标仅限于已映射的位置或清晰的直线路径。

为了实现我们的想法，我们的第一个贡献是一种新颖的深度强化学习方法，用于带有音频-视觉航点的AudioGoal导航。该模型是分层的，具有生成航点的外部策略和计划达到每个航点的内部模块。对于3D导航的分层策略并不新颖，例如（Chaplot等人，2020b; Stein等人，2018; Bansal等人，2019; Caley等人，2016）。然而，现有的视觉导航方法使用启发式方法来定义子目标，而提出的代理学会以端到端的方式为导航任务设置有用的子目标。这是一种通用的3D视觉导航子目标的新思想，而不仅仅是音频目标（参见第2节）。作为第二个技术贡献，我们引入了一个声学记忆，记录代理移动时听到的内容，与其视觉空间记忆相辅相成。而现有模型纯粹基于非结构化记忆（GRU）聚合音频证据，我们提出的声学地图是结构化的、可解释的，并且在整个强化学习流程中整合音频观察。

我们在Replica和Matterport3D的复杂3D环境中使用SoundSpaces音频（Chen等人，2020）演示了我们的方法。它在AudioGoal导航的技术水平上取得了显著的提升（在听到的声音上的SPL上提高了8到49分），并且在未听到的声音、嘈杂的音频和干扰性声音的挑战性案例中具有更好的泛化能力。我们的结果显示，以端到端的方式学习设置航点优于当前的子目标方法，而提出的声学记忆帮助代理更加智能地设置目标。

# 2 相关工作

在3D环境中学习导航

机器人可以通过使用3D重建算法（如SfM）映射空间，然后规划它们的移动（Thrun, 2002; Fuentes-Pacheco等人，2012）来导航复杂的现实世界环境。尽管许多重要的进展都遵循这一工作方向，但正在进行的工作还展示了直接从自我中心的RGB-(D)观察中学习地图编码和导航策略的潜力（Gupta等人，2017a;b; Savinov等人，2018; Mishkin等人，2019）。目前的方法主要关注所谓的PointGoal任务：代理接收到指向目标位置的2D位移矢量，必须通过自由空间导航到达目标位置。代理依赖于视觉输入和（通常为）GPS测距（Gupta等人，2017a; Mishkin等人，2019; Savva等人，2019; Sax等人，2018; Chaplot等人，2020b）。

相比之下，最近引入的AudioGoal任务要求代理使用视觉和音频导航到声源目标（Chen等人，2020; Gan等人，2020）。重要的是，与PointGoal不同，AudioGoal不提供指示目标的位移矢量。现有的AudioGoal方法要么学习一个策略来选择使用多模态输入的最佳即时下一步动作（Chen等人，2020），要么从音频输入中预测最终目标位置，然后根据视觉输入规划到达该位置的路径（Gan等人，2020）。我们关于音频-视觉航点和声学地图的思想是全新的，并且在性能上取得了显著的提升。

具有中间目标的导航

当前的方法通常学习通过逐步动作空间（例如，TurnRight，MoveForward，Stop）奖励到达最终目标位置的策略（Gupta等人，2017a; Mirowski等人，2016; Mishkin等人，2019; Savva等人，2019）。然而，最近的工作探讨了在PointGoal导航中引入子目标或航点的方法。受分层学习的启发（Bacon等人，2017; Nachum等人，2018），其基本思想是选择一个子目标，使用规划（或本地策略）导航到当前子目标，然后重复（Stein等人，2018; Bansal等人，2019; Chaplot等人，2020b; Nair & Finn, 2020; Wu等人，2020; Caley等人，2016）。例如，Bansal等人（2019）应用CNN对RGB输入进行处理，以预测下一个航点，其地面真实值使用轨迹优化收集，然后应用基于模型的规划。Active Neural SLAM（ANS）（Chaplot等人，2020b）使用环境的部分地图为点目标（或预测的长期探索目标）规划路径，使用解析最短路径规划器将每个子目标生成为距离智能体不到0.25米的位置。我们强调，对于导航，ANS不进行全局策略预测；PointGoal坐标只是简单地作为全局目标输入。这些方法的模块化性质与提出的模型相呼应。然而，存在一些重要的区别。首先，我们处理的是AudioGoal，而不是PointGoal，这意味着我们的顶层模块没有给定目标位置，而是必须根据音频输入学习如何引导代理。其次，我们引入了音频-视觉子目标；而视觉子目标侧重于可见障碍物避免，音频-视觉航点则受益于音频的广泛覆盖。例如，视觉子目标可能认为两个出口门都同样好，而音频-视觉航点更喜欢那个似乎发出更大声音的。第三，我们方法的一个关键新元素是学会以端到端的方式生成导航子目标。相反，先前的工作依赖于启发式方法，如选择前沿（Caley等人，2016; Stein等人，2018）或沿最短无碰撞路径的点（Bansal等人，2019; Chaplot等人，2020b）来定义子目标。我们的航点不是使用启发式方法，而是由策略直接预测。这是对视觉导航的一项新的技术贡献，独立于音频-视觉设置，因为它使代理能够动态识别由最终导航目标驱动的子目标。最近在分层强化学习（HRL）方面的一些工作（Nachum等人，2018; Li等人，2019; Levy等人，2019）探讨了使用高层策略端到端预测子目标的方法，但它们在相同的环境中使用人工低维状态输入训练和测试策略，而我们训练我们的代理以概括到看不见的现实3D环境，并使用视觉和听觉传感输入。

视觉语义记忆和映射

基于学习的视觉映射算法

学习型的视觉映射算法（Henriques & Vedaldi, 2018; Savinov等人，2018; Gupta等人，2017b;a）展示了克服纯几何地图局限性的激动人心的前景。学得的地图可以在训练过程中编码超出3D点的语义信息，同时适用于代理的最终任务（如导航）。最近的工作探索了空间索引学习的RGB-D特征（Tung等人，2019; Henriques & Vedaldi, 2018; Gupta等人，2017a）、使用可视化明显的节点构建拓扑记忆（Savinov等人，2018; Nagarajan等人，2020; Chaplot等人，2020a），或者使用存储的视觉嵌入上的注意力模型（Fang等人，2019）。在扩展这一工作的基础上，我们引入了第一个多模态空间记忆。它编码了与智能体沿地平面移动注册的视觉和听觉观察。我们展示了多模态记忆对于智能体生成良好动作序列至关重要。

声音定位

机器人系统使用麦克风阵列定位声源（Nakadai & Nakamura, 1999; Rascon & Meza, 2017），而主动控制可以改善定位（Nakadai等人，2000; Wang等人，2014）。房间的几何形状可以通过音频部分感知，如利用回声定位的思想（Dokmanic等人，2013; Christensen等人，2020; Gao等人，2020）。在2D视频帧中，方法学会基于它们与音频一致的视觉关联来定位声音（Hershey & Movellan, 2000; Tian等人，2018; Senocak等人，2018; Arandjelovic & Zisserman, 2018）。与上述任何一种方法不同，我们研究了音频-视觉导航问题，其中代理根据音频和视觉提示在3D环境中有效地朝着声源移动。

# 3 方法

我们考虑的任务是音频目标导航（AudioGoal navigation）（Chen等人，2020; Gan等人，2020）。在这个任务中，智能体在3D环境中移动，每个时间步t从其摄像头（深度）和双麦克风接收传感器观测Ot。在导航开始时，环境未被映射；智能体必须在导航过程中累积观测以了解场景几何结构。与常见的PointGoal任务不同，对于AudioGoal，代理不知道目标的位置（即没有GPS信号或指向目标的位移矢量）。代理必须使用音频源发出的声音来成功定位并导航到目标。

我们引入了一种新颖的导航方法，该方法预测中间航点以有效地达到目标。我们的方法由三个主要模块组成（图2）。给定视觉和音频输入，我们的模型1）使用感知和映射模块对这些线索进行编码，然后2）预测航点，最后3）规划和执行一系列将智能体带到预测航点的动作。智能体重复这个过程，直到它预测已经到达目标并执行停止动作。

3.1 3D环境和音频-视觉模拟器

我们使用AI-Habitat模拟器（Savva等人，2019），搭配公开可用的Replica（Straub等人，2019）和Matterport3D（Chang等人，2017）环境以及公开的SoundSpaces音频模拟（Chen等人，2020）。Replica环境是通过对公寓、办公室、酒店和房间的真实世界扫描构建的网格。Matterport3D环境是真实世界家庭和其他室内环境，具有3D网格和图像扫描。代理可以在接收实时的自我中心视觉和音频观测的同时穿越这些空间。使用SoundSpaces的房间脉冲响应（RIR），我们可以在3D环境中放置音频源，然后在场景中的每个位置模拟现实声音，空间分辨率为Replica的0.5m和Matterport3D的1m。这些最先进的渲染捕捉了声音如何传播以及与周围几何和表面材料的相互作用，模拟了RIR的所有主要特征：直接声音、早期镜面/漫反射反射、混响、双耳空间化以及来自材料和空气吸收的频率依赖效应（有关详细信息，请参见Chen等人（2020））。我们尝试使用102种日常声音进行实验（详见Supp）。

模拟器维护环境的可导航图（对代理不可见）。如果两个节点之间存在连接边，并且代理面朝该方向，代理才能从一个节点移动到另一个节点。动作空间A有四个动作：MoveForward、TurnLeft、TurnRight和Stop。我们使用这些真实世界的图像扫描和高度真实的音频模拟来在可重现的评估环境中测试我们的想法。请查看Supp视频以评估其逼真程度。我们的实验通过考虑分心声音和嘈杂传感器（详见Sec. 4）进一步推动了逼真度。将策略转化为真实世界机器人是未来的工作，我们受到最近的模拟到实际尝试的鼓舞（Gupta等人，2017a; Müller等人，2018; Chaplot等人，2020b; Stein等人，2018）。

3.2 感知和映射

视觉感知

在每个时间步t，我们从代理的第一人称深度视图中提取视觉线索，这对于地图构建比RGB更有效（Chaplot等人，2020b; Chen等人，2019）。首先，

我们使用摄像机的内参数将深度图像反投影到世界坐标中，以计算局部场景的3D点云。然后，我们将这些点投影到一个2D顶视自我中心局部占用图Lt中，其大小为3×3米，对应于真实世界传感器可靠的典型距离。该地图有两个通道，一个用于占用/自由空间，一个用于已探索/未探索区域。如果一个地图单元有一个高于0.2米且低于1.5米的3D点，被视为占用；如果任何3D点投影到该单元，则被视为已探索（结果对嘈杂深度具有宽容性；请参见Supp）。我们通过将Lt相对于代理的上一个姿势变化进行转换，然后与Gt−1的相应值平均，来更新一个地心几何地图Gt。值大于0.5的单元格被视为占用或已探索。见图2的顶部分支。

声学感知

在每个时间步，代理接收由右耳和左耳的声谱图表示的双耳声音Bt，声谱图是音频信号频率随时间变化的矩阵表示（图2的第二个分支；有关声谱图详情，请参见Supp）。除了编码当前声音外，我们还引入了声学记忆。声学记忆是一个与Gt相似的地图At，以结构化的方式聚合了随时间变化的音频强度。它记录了代理访问的位置上的直接声音强度的移动平均值。见图2的第三个分支。请注意，音频强度地图既揭示了有关环境和目标的空间信息，墙壁和其他主要表面影响代理在任何给定位置接收到的声音，而目标处的声源在代理远离目标时提供了对方向的粗略感知。随着代理接近目标，这种定向线索变得越来越精确。

3.3 音频-视觉航点预测器

音频和视觉输入都携带了互补的信息，以设置通往音频目标的良好航点。虽然音频信号Bt（双耳输入）和At（声学记忆）告诉代理目标的大致方向并提示房间几何形状，但以占用图Gt的形式的视觉信号允许航点的空间定位，并有助于避开障碍物。回顾图1，代理在卧室中需要到达另一个房间里响起电话的地方。我们学习了三个编码器来表示这些输入：gt = fg(Gt)、bt = fb(Bt) 和 at = fa(At)。函数fg和fa首先变换几何和声学地图（Gt和At），使代理位于地图的中心朝上，并将它们裁剪到尺寸sg × sg和sa × sa，分别。每个函数最后都有一个卷积神经网络（CNN）来提取特征（详见Supp）。我们将三个向量gt、bt和at连接起来，获得完整的音频-视觉特征，并将其传递到门控循环神经网络（GRU）（Chung等人，2015）。见图2。

我们的强化学习航点预测器具有演员-评论家结构。它获取GRU的隐藏状态ht并预测可能航点的概率分布π(Wt|ht)。Wt是大小为sw × sw的动作图，表示以代理为中心的区域内的候选航点。我们使用局部占用图对策略的输出进行蒙版处理，以确保模型选择在自由空间中的航点。我们根据策略的预测概率分布从Wt中采样一个航点wt = (∆x, ∆y)。航点相对于代理的当前位置，并传递给规划器（见Sec. 3.4）。

这个航点策略是我们方法设计中的一个重要元素。它允许代理根据当前看到和听到的信息动态调整其中间目标。与现有的AV导航方法不同，我们的航点以可变的粒度引导代理，而不是将其动作固定为短视的下一步（Chen等人，2020）或最终目标的预测（Gan等人，2020）。与现有的视觉子目标方法不同，后者依赖于基于边界的启发式方法或最短路径上的点（Chaplot等人，2020b; Stein等人，2018; Bansal等人，2019; Caley等人，2016），我们的航点与导航任务紧密集成。我们的结果展示了这种优势。

3.4 路径规划器

给定生成的航点wt，最短路径规划器试图生成一系列从A中选择的低级执行命令，以将代理移动到该航点。规划器维护基于几何地图Gt的场景图，并使用Dijkstra算法估计从代理的当前位置到wt的路径。在规划过程中，将地图中的未探索区域视为自由空间（Chaplot等人，2020b）。基于最短路径，通过解析计算出一个低级执行命令。代理执行该动作，获得新的观测Ot，更新Gt和At，然后重复上述过程，直到退出规划循环。

规划循环在三种情况下中断：1）代理到达航点，2）规划器无法找到到达航点的路径（在这种情况下，代理在中断循环之前执行随机动作），或3）代理达到规划步数限制。规划步数限制设置为减轻糟糕的航点预测（由于嘈杂的占用估计）或难以到达的航点（如在另一房间的墙后）使代理偏离目标的影响。如果模型选择wt = (0, 0)（即代理的当前位置），这意味着代理认为已经到达最终目标；然后执行Stop动作，该情节终止。

3.5 奖励和训练

遵循典型的导航奖励（Savva等人，2019; Chen等人，2020），如果代理成功到达目标并在那里执行Stop动作，则奖励代理+10，此外，为减少到目标的测地距离奖励+0.25，增加则有等效的惩罚。最后，我们对每个执行的动作进行时间惩罚，为了鼓励效率，奖励为-0.01。对于每个航点预测步骤，代理根据上一轮规划执行期间收集的累积奖励值获得奖励。总的来说，奖励鼓励模型选择可达、远离当前代理位置并在通往目标的路线上的航点，或者选择目标本身如果它在可达范围内。所有可学习的模块都是使用Proximal Policy Optimization（PPO）（Schulman等人，2017）进行联合训练和更新的，每150个航点预测步骤。PPO损失包括值网络损失、策略网络损失和熵损失，以鼓励探索。请参见Supp以获取所有实现细节。

# 4 实验

环境

我们在Habitat模拟器中的Replica和Matterport环境中使用SoundSpaces进行测试（详见Sec. 3.1）。我们遵循SoundSpaces AudioGoal基准测试的协议（Chen等人，2020），在Replica上进行9/4/5场景的训练/验证/测试划分，在Matterport3D上进行73/11/18场景的训练/验证/测试划分。我们强调测试和训练/验证环境是不相交的，要求代理学习可推广的行为。此外，对于相同的场景划分，我们尝试使用不相交的声音进行训练和测试，要求代理推广到未听到的声音。对于已听到的声音实验，电话铃声是声源；对于未听到的声音，我们从102个独特的声音中选择（详见Supp）。

评估指标

我们评估以下导航指标：

1. 成功率（SR）：成功剧集的分数，即代理在网格上准确停在音频目标位置的剧集的分数。

2. 路径长度加权的成功率（SPL）：标准指标（Anderson等人，2018），通过路径长度加权成功剧集。

3. 动作数量加权的成功率（SNA）：惩罚原地旋转动作的成功率，这些动作不会导致路径更改。

请参见Supp以获取这些全面指标的更多详细信息。

现有方法和基线

我们比较以下方法（详见Supp; 请参阅表格3）：

- 随机（Random）：随机选择每个动作，并在到达目标时发出Stop信号的代理。

- 方向跟随者（Direction Follower）：一个分层模型，根据音频预测的到达方向（DoA）设置中间目标K米远，并重复。K通过在验证划分上进行超参数搜索获得， Replica为K = 2，Matterport为K = 4。我们训练一个基于音频输入的单独分类器，以预测此代理何时应该停止。

- 边界航点（Frontier Waypoints）：一个分层模型，将预测的DoA与探测区域的前沿相交，并将该点选择为下一个航点。前沿航点通常在视觉导航文献中使用，例如（Caley等人，2016; Stein等人，2018; Chaplot等人，2020b），这使得它是标准做法的广泛代表性基线。

- 监督航点（Supervised Waypoints）：一个分层模型，使用RGB帧和音频频谱图来在其视野（FoV）内监督（非端到端）学习预测航点。该模型受到Bansal等人（2019）的启发，后者以监督方式学习预测航点。

- Chen等人（2020）：一个端到端的AudioGoal RL代理，使用音频-视觉观察选择动作。它没有任何几何或声学地图。我们运行作者的代码。

- Gan等人（2020）：一个最先进的AudioGoal代理，仅从双耳频谱图中预测音频目标位置，然后使用逐步构建的深度图上的分析路径规划器导航。它使用单独的音频分类器进行停止。我们改进了该模型以提高其在Replica和Matterport上的性能，因为作者最初在游戏引擎模拟器上测试过（详见Supp）。

导航结果

我们考虑两种设置：

1. 已听到的声音：在电话声音上进行训练和测试，按照（Chen等人，2020; Gan等人，2020）。

2. 未听到的声音：使用不相交的声音进行训练和测试，按照（Chen等人，2020）。

在这两种情况下，测试环境始终是未见过的，因此这两种设置都需要泛化。我们的方法在两个数据集的所有指标中都优于其他三个分层模型。这突显了我们直接学习设置航点的优势，而不是当前分层视觉导航模型中使用的启发式方法。即使监督航点模型在未见环境中的性能也不如AV-WaN好。我们认为这是由于监督对于给定导航任务的最佳航点的狭窄定义，相对于我们的模型，后者通过自己的经验学习了在端到端方式中给定导航任务的最佳航点。

在102种未听到的声音设置中，我们的方法再次在两个数据集上几乎在所有指标上都表现出色。唯一的例外是我们在Replica上与Chen等人相比SPL较

低的2.8％，尽管我们的模型仍然在该数据集上在SNA方面超过了Chen等人，这意味着在总动作计数的规范化方面我们的准确性更高。然而，由于不熟悉的音频频谱图模式，所有方法的绝对性能都有所下降。对于这个重要的设置，声学记忆至关重要；它成功地将训练声音的特定内容抽象为更好地进行泛化。

剔除实验

输入模态和音频-视觉航点组件的消融

表2显示了我们模型的输入模态和音频-视觉航点组件的消融实验。去除几何和声学地图导致性能下降。这是预期的，因为没有At和Gt，模型只有当前音频观察Bt来预测下一个航点。值得注意的是，即使我们的模型的这个严重消融版本在性能上仍然优于最佳的现有模型（Chen等人，2020）（见表1）。这表明我们基于航点的导航框架本身比更简单的RL模型（Chen等人，2020）以及现有的子目标方法更有效。仅去除At也导致性能下降，这表明了所提出的结构化声学记忆对于有效导航的重要性。At和Gt是互补且对于我们的模型达到最佳性能至关重要。最后，我们评估了我们的音频-视觉航点预测理念的影响。我们将模型中的演员网络（见图2中部）替换为一个线性层，该层输出A中四个基本动作的动作分布。一个动作采样器直接从这个分布中采样一个动作并在环境中执行。我们的模型相对于这种消融的增益证实了航点对我们的模型的价值，即使在所有其他组件都固定的情况下。

失败案例

我们对模型的失败剧集进行了分析，发现其中存在两种重复的失败类型（见Supp视频示例）。第一种是音频目标被困在障碍物中或紧靠墙壁。在这种情况下，虽然AV-WaN能够迅速到达目标，但它在目标周围不断摆动，由于目标周围障碍物的强烈音频反射或映射错误而无法准确确定目标位置。在第二种情况中，我们注意到有时代理会在音频目标旁边过早地执行停止动作。我们认为在目标周围音频强度差异较大的区域更难检测，这可能导致这种行为。

嘈杂的音频和干扰音

为了了解我们的模型在嘈杂的音频感知下的鲁棒性，我们考虑两种音频噪声的来源：环境噪声和麦克风噪声。对于环境噪声，我们添加干扰音（例如人类说话，风扇旋转）以干扰代理的音频感知。代理总是被要求找到电话，而干扰器是放置在随机位置的未听到的声音。在每个时间步，代理接收两个声音的组合波形，需要捕捉电话信号并找到其源位置。我们使用了Heard实验（表1）的相同剧集来训练和评估代理。在有干扰的情况下，最佳性能的基线Chen等人在Replica和Matterport上分别获得了71.7％和53.3％的测试SPL，而我们的模型分别达到了83.1％和70.9％。对于麦克风噪声，我们对接收的音频波形添加了逐渐增加的高斯噪声。图4b显示了结果。AV-WaN对音频噪声非常鲁棒，尤其是在有At的情况下，而现有的AudioGoal方法则受到了明显的影响。因此，我们的模型在现实世界中常见的嘈杂环境中的优势仍然存在，并且声学记忆在这种嘈杂的环境中是至关重要的。

动态航点选择

图4a绘制了欧几里得距离到航点的分布，作为代理到目标的测地距离的函数。我们看到我们的代理在离目标较远时选择较远的航点，然后在接近目标时预测较近的航点。详细信息和分析请参见Supp。

航点的放置

为了检查基于周围几何形状选择航点的方式，图4c绘制了在测试Replica环境中的一个顶部地图上的航点分布。航点是在具有起点或终点在a房间或c房间的轨迹上积累的，而目标位置则被排除在外。我们看到航点主要选择在障碍物和门周围，这是决策状态，位于状态空间的关键交叉点，代理可以从中收集最多的新信息，并过渡到新的，可能是未探索的区域（Goyal等人，2019）。最常见的航点通常相距2-3米，接近代理可以选择的最大距离。

# 结论

我们引入了一个强化学习框架，该框架学会为音频-视觉导航设置航点，并配备了声学记忆。我们的方法在具有挑战性的AudioGoal问题上改进了现有技术水平，我们的分析显示了新技术贡献的直接影响。在未来的工作中，我们计划考虑越来越复杂的AV导航任务，例如语义声音、移动声源和实际应用转移。