# 摘要

移动智能代理的一个关键能力是整合来自环境中多个感官输入的证据，并采取一系列行动以达到其目标。在本文中，我们尝试解决音频-视觉体验导航的问题，即在仅提供原始自我中心视觉和音频感官数据的情况下，规划从场景中的随机起始位置到室内环境中的声源的最短路径的任务。为了完成这个任务，代理需要从各种模态中学习，即将音频信号与视觉环境关联起来。在这里，我们描述了一种利用视觉和音频证据的音频-视觉体验导航方法。我们的解决方案基于三个关键思想：一个构建环境空间记忆的视觉感知映射模块，一个从代理中推断声源相对位置的声音感知模块，以及一个根据音频-视觉观察和环境空间记忆规划一系列行动的动态路径规划器以导航达到目标。在使用模拟的多模态环境收集的新的视听房间数据集上的实验结果表明，我们的方法相对于几个竞争基线表现出了有效性。

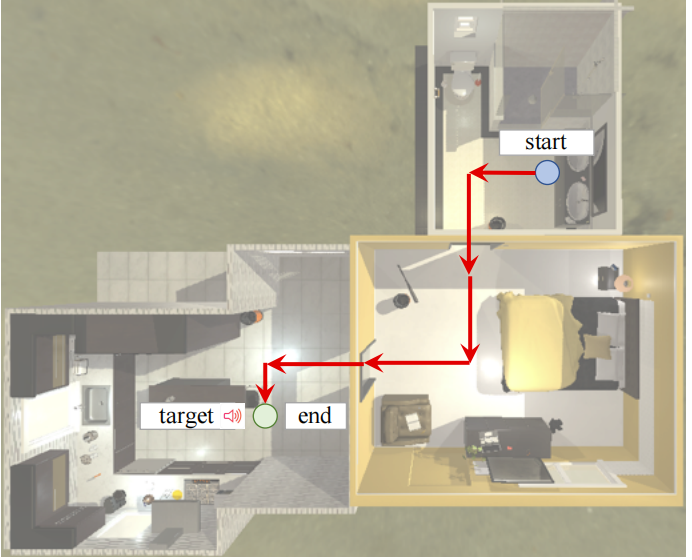
# I. 引言

我们人类通过整合多个感官输入，包括但不限于声音和视觉表示，来感知和穿越世界。认知心理学和神经科学的大量证据表明，即使在幼儿时期，人类也非常能够捕捉不同信号模态之间的对应关系，以全面理解自然现象 [1]，[2]。

在这项工作中，我们的目标是建立一个机器模型，以实现类似于人类在音频-视觉场景分析方面的能力。特别是，我们关注一个尚未充分探索的问题：如何教导一个配备摄像头和麦克风的智能代理与环境进行交互并导航到声源。这个任务比理解静态图像更复杂和具有挑战性。我们将其称为音频-视觉体验导航。代理必须根据多个感官观察依次做出决策，以在环境中导航。图1说明了音频-视觉体验导航问题的示例。代理从室内环境中随机开始，并提供一个声音脉冲（例如，手机铃响）。然后，代理需要找出从其起始位置到声音源（目标位置）的最短轨迹。为了完成这个任务，代理必须将声音与视觉环境相关联，同时从原始感官输入映射到路径规划的行动。这个任务非常具有挑战性，因为它实质上要求代理逆向工程其感知到的信号的因果效应。它不仅必须理解视觉环境，还必须推理潜在的声源。如果声源在另一个房间，代理还必须学会移动到另一个房间。在这个任务上的进展将促进家庭机器人在许多实际应用中的应用。例如，寻找声源的能力可以

图一

赋予机器人帮助人类找到手机或关闭水龙头。

我们通过借鉴人类如何完成这个任务的见解来解决这个问题。想象一下用户在新环境中寻找声源的过程。他/她首先对位置、周围环境以及到声源的距离和方向有一定的感觉。当他/她四处移动时，房间布局以及他/她关于声音和房间的记忆在他/她做出未来行动的决策中也起着至关重要的作用。为此，在本文中，我们设计了一个音频-视觉网络，让机器代理模仿我们人类在现实生活中使用的导航机制。我们系统地研究了两种设置：探索与行动以及非探索。在第一种设置中，代理被允许在发出声音警报之前探索环境一定数量的步骤；因此，它可以通过在探索过程中的视觉观察建立对房间的不完整空间记忆，并在解决随后的导航任务中参考这个内部表示。在第二种设置中，代理必须在导航过程中建立一个动态的空间地图。当代理被要求搜索声源时，类似于人类，代理必须首先推断声源的相对位置，然后根据其空间记忆和视听观察采取顺序行动。它还可以在每个时间步更新其空间记忆。

总结一下，我们的工作做出了以下贡献：

• 我们建立了一个多模态导航环境，以促进音频-视觉体验导航的研究。该环境包含相当复杂的公寓，并具有一个集成的声音模块，观察某些物理定律。

• 我们提出了一个视听房间（VAR）基准，以系统评估多模态导航代理的性能。

• 我们提出了一个音频-视觉体验导航框架，并将其与几个竞争基线进行对比。实验证明了我们方法的有效性，同时揭示了这个任务中的关键挑战，为未来的工作提供了启示。

# II. 相关工作

A. 面向目标的导航

早期的工作通过使用SLAM构建场景地图，然后在该地图中规划路径来解决导航任务[3]。最近，基于深度学习的方法已经被用于直接从原始感官数据规划行动。Zhu等人[4]研究了一种基于深度强化学习的反应性导航方法，以在离散化的3D环境中找到目标对象的图像。Gupta等人[5]通过一个映射器和规划器学习导航。Sadeghi等人[6]研究了一种强化学习方法，可以仅使用模拟数据来教导代理在真实环境中飞行。Mirowski等人[7]通过与循环闭环检测和从RGB获取深度估计等辅助任务联合训练来改进导航结果。Brahmbhatt等人[8]探索了一种用于导航大城市的CNN架构，使用街景图像。Wu等人[9]在第一人称射击游戏环境中将深度强化学习与课程学习相结合。Yang等人[10]提出使用语义先验来改进视觉导航。McLeod等人[11]利用过去的经验来改进机器人在动态未知环境中的未来导航。Katyal等人[12]使用生成网络预测未来机器人运动的占用地图表示。Mousavian等人[13]演示了使用语义分割和检测掩码进行目标驱动的视觉导航。Savinov等人[14]的工作在导航中使用记忆方面对我们具有启发作用。与他们的拓扑图记忆不同，我们使用键-值结构更好地捕捉代理关于环境的内部知识的不完整性，并在非探索设置中引入动态空间记忆。

还有大量基于视觉-语言的体验导航的工作[15]，[16]，[17]，[18]，它直接从语言指令和视觉观察映射到行动。与所有这些方法不同，音频-视觉体验导航任务的目标是找到声源，因此要求代理整合视觉和声学线索以规划一系列行动。与我们的工作同时进行，Chen等人[19]也在Habitat平台[20]上提出了音频-视觉体验导航任务。

B. 声音定位

声音定位的研究问题旨在确定视频中哪些区域发出声音，已有数十年的研究。现有方法分为两大类：基于计算的方法和基于学习的方法。早期的工作使用高斯过程模型[21]，子空间方法[22]，典型相关分析[23]，手工制作的运动[24]和分割[25]来测量像素和声音之间的相关性。最近，研究人员提出训练深度神经网络通过观看和听取许多未标记的视频来定位发出声音的对象[26]，[27]，[28]，[29]，[30]，[31]，[32]，[33]，[34]，[35]。这些方法解决了在视频上定位区域的任务。相比之下，我们的目标是找到声源，这也可能是虚拟环境中的非直线视线。我们的工作还与声源定位的基于声学的方法[36]，[37]，[38]有关。它们通常需要特殊设备（例如麦克风阵列）来记录声音。我们的工作超越了这一点，因为移动代理需要进一步将声音定位的结果与视觉环境连接起来进行路径规划。

# III. 环境

我们在AI2-THOR平台[39]上构建了一个多模态虚拟环境，该平台使用Unity游戏引擎构建了一组场景。我们进一步将一个空间音频软件开发工具包，Resonance Audio API [40]，整合到Unity游戏引擎中，以支持音频-视觉体验导航任务。一个配备摄像头和麦克风的代理可以利用两种不同的感知模态（自我中心的RGB图像和声音）来感知并在场景中导航。

A. 场景

AI2-THOR平台提供了接近照片级别真实感的室内场景。它包含四个类别的120个场景：厨房、卧室、客厅和浴室。每个房间类别包含30个外观各异的房间。由于AI2-THOR中的房间是相互隔离的，它们对于区域目标导航任务来说并不具有足够的挑战性。如图1所示，我们手动将几个房间连接成多房公寓。我们总共能够为这项工作构建七个公寓。与Zhu等人[4]类似，我们还将公寓离散化为网格世界，以便进行量化评估。特别是，每个网格方块为0.5×0.5平方米。总共，每个公寓约有150到200个网格单元。

B. 声音

声学引擎是使用Resonance Audio API实现的。这个空间音频SDK可以在虚拟环境中创建高保真的空间音效，并根据对象和房屋的几何形状支持声学光线追踪。Resonance Audio API还支持模拟声音波与具有逼真的频率依赖声音混响特性的各种几何形状和不同材质的对象相互作用。声音波与人耳的相互作用也可以根据耳间时间差、耳间电平差和频谱效应进行准确模拟。配备了Resonance Audio API的场景可以制造一种声音来自虚拟世界特定位置的错觉。因此，我们可以使用该模块为代理提供立体声声音帧。在我们的实验设置中，我们主要考虑连续的声音源（例如，手机铃声和警报提示）。

C. 问题设置

在音频-视觉体验导航任务中，我们考虑两种问题设置：探索与行动以及非探索，用于寻找声源。在第一种设置中，我们让代理与环境进行两个阶段的交互，与Savinov等人[14]的探索设置类似：探索和目标导向导航。在探索阶段，代理随机在环境中走动，直到其轨迹长度达到给定的预算。这种探索经验可以用来建立环境的内部世界模型，用于随后的目标导向导航任务。在第二种设置中，代理必须在导航到目标的同时同时构建对环境的空间记忆。

在测试阶段，我们随机选择代理的起始位置和声源的目标位置。在每个时间步，代理对环境进行自我中心的视觉观察ov和声音观察os，然后从动作集A中选择一个动作。动作集A包括以下动作：向后移动、向前移动、向左旋转、向右旋转和停止。

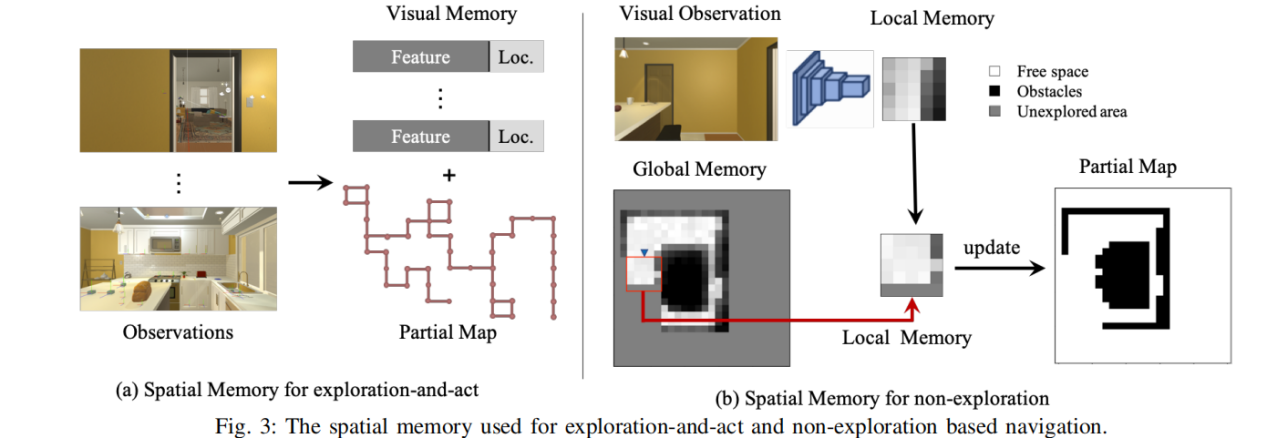
# 图二

# IV. 方法

在本节中，我们介绍了用于音频-视觉体验导航的算法。如图2所示，它包括三个组件：视觉感知映射器、音频感知模块和动态路径规划器。

为了解决音频-视觉体验导航任务，代理可以使用视觉感知映射器从在探索阶段构建的空间记忆中检索，或者使用在导航期间估计的占用图构建环境的部分2D空间地图（带有边的图）。然后，代理利用音频感知模块估计来自其当前位置的声源的方向和距离。最后，代理根据从视觉和音频感知模块以及部分空间地图推断出的结果规划找到达到目标的最短路径。在导航的中间，动态规划器还可以根据任何新的视觉和音频观察更新估计的目标位置以及代理的导航模型。

A. 视觉感知映射器

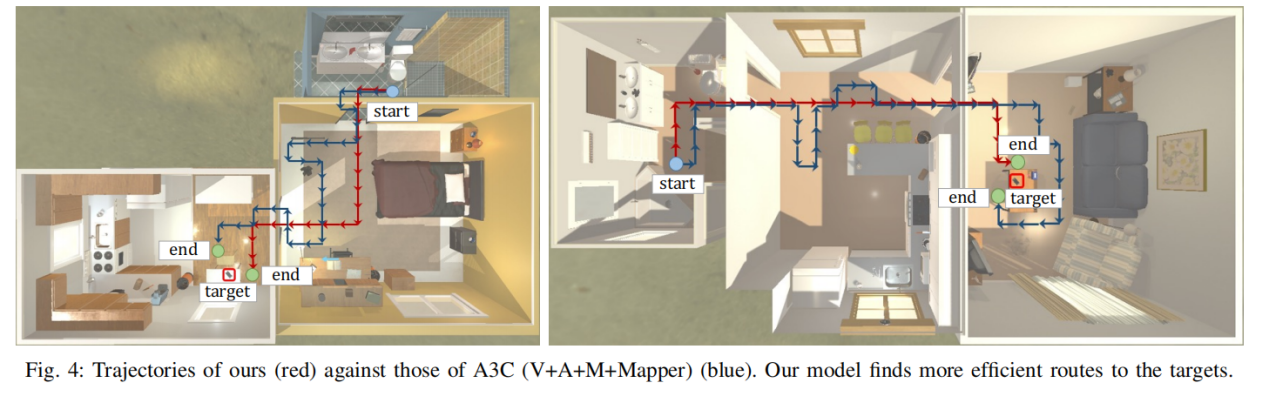
a) 探索与行动视觉映射器：在探索阶段，基于探索的视觉映射器包括一个空间记忆和一个非参数检索模型。在探索期间，我们使用基于键值的空间记忆网络[41]来对代理访问的环境进行编码。视觉观察存储在键部分，而元数据（例如，代理采取的位置坐标和行动）存储在值部分。具体而言，我们通过ImageNet预训练的ConvNet模型ResNet [42]将每个第一人称视图RGB图像传递，以提取视觉特征向量，然后进行L2归一化。每个视觉观察的特征向量和代理的相应坐标和方向形成一个键值对（图3(a)）。

在面向目标的导航阶段，我们使用相同的特征提取流水线从代理的第一人称视图中获取查询特征向量。然后，在空间记忆的键上进行非参数检索步骤。返回与余弦距离下查询最接近的前三个记忆槽。我们平均存储的坐标以估计代理的位置。请注意，在我们的实验中，我们强制代理的起始位置与探索位置没有重叠，因此在早期导航阶段，代理的定位存在相当大的不确定性。

b) 非探索视觉映射器：由于在实际应用中可能不总是允许随机探索阶段，我们进一步引入了一个视觉感知映射器，该映射器可以在代理导航环境时动态构建空间记忆。我们将空间记忆构建为关于网格世界环境的顶部2D占用图M ∈ RN×N（图3(b)）。记忆的每个条目表示环境中的一个网格单元，其值在范围[0, 1]内，表示代理对网格单元是否可穿越的信念。我们将值初始化为0.5，并将代理的起始位置放在记忆M的中心。由于代理只能访问附近区域的视觉观察，视觉感知映射器在每个时间步产生代理前面的局部占用图。然后，代理使用贝叶斯规则[43]更新局部占用图。

为了推断一个单元是否可穿越，我们训练一个卷积神经网络，将RGB图像映射到代理的信念。我们通过利用两个训练公寓中的RGB图像和地面实况2D占用图来收集训练数据。我们将第一人称RGB图像馈送到网络中，以在小窗口大小（我们的实验中为5×5网格）内预测自由空间概率。

B. 音频感知模块

受到人类可以心理推断声音位置的事实的启发，这里我们引入了一个音频感知模块，用于估计目标的坐标，即环境中声源的位置。为了使网络易于学习，我们使用相对位置来估计目标的坐标。我们在网格世界中定义音频源相对位置如下。音频源相对于代理的位置（x，y）表示音频源位于代理右侧x米处，距离代理前方y米处。因此，可以通过同时考虑代理和声音源的绝对坐标以及代理的方向来计算相对位置。在绝对坐标系统中，东轴是正x轴，北轴是正y轴。假设在网格世界中，声源和代理的绝对坐标分别为（16，18）和（4，8），则在代理朝北的情况下，声源的相对位置为（12，10）。如果代理面向西，声源的相对位置变为（10，-12）。

我们通过使用两个训练公寓中的所有录制的音频剪辑收集训练数据。为了预处理立体声音，我们首先将声音重新采样为16 HZ，然后通过短时傅里叶变换（STFT）将其转换为频谱图。我们将频谱图馈送到一个五层卷积网络中。训练目标是在监督方式下用均方误差（MSE）损失训练音频感知模块，以估计相对于代理的声音位置。可以从相对位置和代理的坐标计算出目标的绝对坐标。音频感知模块还包含噪声，因为房间的布局和表面材料会影响声音。我们还训练了一个单独的声音分类模型，以确定代理是否到达目标；如果是，则将采取停止操作。

C. 动态路径规划器

代理采用动态路径规划器在观察到变化并同时更新内存的情况下在环境中导航。对于探索与行动的设置，代理可以基于代理的探索轨迹（x0，x1，···，xn）和行动（a1，a2，···，an）的序列构建房间的部分图G =（V，E）。记V和E分别为节点和边。图中的每个节点vi ∈ V存储环境中的一个位置（坐标）。如果节点对应于连续的时间步骤，则存在于vi和vj之间的边。

对于非探索的方法，代理可以使用内存构建环境的部分图。根据空间记忆，每个网格可以分为三种类型：自由空间，障碍物和未探索区域。通过假设未探索和未占用的网格可穿越，将可穿越的网格转换为节点，并通过边连接相邻的节点，生成部分图G =（V，E）。当检测到新的障碍时，将删除相应的节点和边。

动态规划器输出一条路径（动作序列），将代理转移到基于环境的图和估计的目标坐标的所需目标位置。在每个时间步t，代理尝试从代理自身的图节点Xtgraph到估计的声源坐标Xttarget最近的图节点Xtgraph 找到最短路径。我们使用Dijkstras算法解决这个问题。具体而言，我们可以使用该算法获得一条最短路径，表示为Ptmin = {at0 , at1 , ...., atm }作为时间t时要执行的动作。

# V. 实验

在本节中，我们首先描述了为寻找声源的音频-视觉体验导航任务创建的数据集和环境，然后展示了我们的方法相对于几种竞争基线的优势。

A. 视觉-音频房间数据集

为了系统地评估多模态导航性能并促进对这一研究方向的未来研究，我们收集了一个新的视觉-音频房间（VAR）基准数据集。我们的最终目标是通过融合视觉和音频观察使机器人能够在室内环境中导航。由于在物理机器人上进行彻底和受控的实验非常具有挑战性，我们改为使用3D模拟环境，构建在AI-Thor平台[4]的基础上。实验中使用了七个公寓。我们将它们分为两个用于训练的公寓和五个用于测试的公寓。我们考虑三个录制的音频类别：铃声、警报警报和时钟滴答声作为工作中的声源。为了生成导航数据，我们预先提取了七个公寓中所有网格的第一人称视图RGB图像。对于训练公寓，我们随机选择了十个位置来放置声源，然后记录了房间中所有位置的代理听到的声音在四个方向。对于五个测试公寓，我们选择了五个网格位置放置声源，并记录了代理在所有位置听到的声音。总共，我们收集了3,728张RGB图像，记录了75,720个音频剪辑。

B. 实验设置

设置。我们的目标是测试代理在新环境中导航的能力。在评估过程中，对于每个房间中声源的五个可能位置，我们随机选择20个代理的位置和方向。因此，我们对每个声音进行了100个测试情节。我们将每个时间步的每个声音的时间持续设置为1秒。在探索与行动的设置中，随机行走步数的数量设置为400。

评估指标。我们使用两个指标来评估模型：成功率和路径长度加权的成功率（SPL）[44]。 SPL衡量了代理在所有测试情景中的导航性能，公式为1 ，其中N表*i*示测试情节的数量，Si是测试情节i中的成功二进制指示器，Pi表示路径长度，li表示到目标的最短距离。

C. 基线

我们考虑了七种基线方法进行评估。

• 随机行走。这是导航任务的最简单启发式方法。代理在每个时间步从动作空间中随机采样一个动作。

• 贪婪搜索（A）。这是一个仅基于声音的基线。代理仅使用声音感知模块在环境中导航。代理贪婪地朝着声源移动，而不观察公寓的结构和布局。

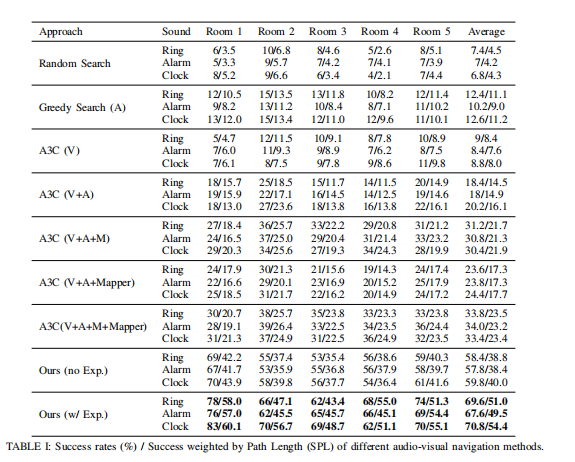
• A3C（V）[45]。异步优势行动者-评论家（A3C）是一种用于导航的最先进的基于深度强化学习的方法。这是一个面向目标的，仅使用视觉的导航基线，没有内存。

• A3C（V+A）。这是一个面向目标的，音频-视觉导航基线，没有内存。在我们的设置中，A3C模型的输入是当前状态的音频-视觉表示。为了公平比较，我们将我们方法中使用的视觉和声音感知网络提取的特征表示级联起来。

• A3C（V+A+Mapper）。在这种情况下，A3C模型的输入是当前状态的音频-视觉表示和我们非探索性方法中使用的显式2D占用图。顶视占用图与音频-视觉表示级联在一起。

• A3C（V+A+M）[14]。与Savinov等人使用的基线类似，我们还实现了一个配备有LSTM内存的A3C作为我们任务的基线。首先，代理在没有给定目标的情况下以相同数量的随机行走步数在探索模式中导航。到达目标没有奖励。在测试时间，我们将探索序列馈送到LSTM代理，然后让其执行目标导航，而不重置LSTM状态。我们期望LSTM模块能够隐含地构建一个内部表示。

• A3C（V+A+M+Mapper）。我们进一步将显式和隐式内存组合到A3C的培训中。



# D. 结果

表I总结了在五个测试公寓中对三种不同声音源的结果。提出的音频-视觉代理在所有指标上均优于所有基线。在探索与行动的设置中，我们的模型（带探险）对于每种声音源的成功率平均超过65％，几乎是表现最佳基线的两倍。图4更详细地呈现了两个测试房间的结果，通过将我们的方法的轨迹与表现最佳的基线（A3C（V+A+M+Mapper））的轨迹进行对比。在非探索性的设置中，我们的方法（无探险）的成功率也可以达到近60％，甚至优于探险性的最强基线。这些结果表明我们的系统更好地利用了房间的布局，并找到了达到目标的最短路径。例如，当声音在另一个房间时，代理倾向于来到门口，从而高效地到达目标。更多定性结果可在补充材料中找到。

从表I中，不难发现基于随机行走的方法的成功率非常低，因为在没有任何先验知识的情况下，在我们的环境中进行导航非常具有挑战性。贪婪搜索略优于随机搜索，因为代理可以从声音中预测目标的位置。然而，成功率仍然低于15％，因为代理对房间布局没有视觉先验知识，并误解最短几何距离为最短路径，而在我们的设置中，这并不总是成立。这些结果表明，连接视觉和声音信息对于音频-视觉导航任务至关重要。

基于A3C的方法的性能明显较之前文献中报告的性能要弱[45]。主要原因是我们关注泛化——将代理放在以前未见过的环境中。这是一个更具挑战性但也更现实的场景。我们还发现我们的方法与A3C之间的性能差距仍然很大，A3C使用了来自探索的隐式空间内存和导航期间的显式空间内存。我们推测我们的方法可以将视觉感知和声音感知与路径规划分开，从而在新的测试环境中提供更好的泛化能力。

# VI. 结论

我们解决了一个新颖的问题，即在多模态虚拟环境中进行音频-视觉体验导航。代理可以利用环境的内部结构化表示，以高效地导航到以前未见过的目标环境。我们的多模态虚拟环境包括具有声音模块和逼真布局的复杂公寓。我们展示了现代深度强化学习方法在泛化方面遇到的困难。我们的方法在新目标和新环境中表现出色。

致谢：本工作部分得到ONR MURI N00014-16-1-2007、Center for Brain, Minds, and Machines (CBMM, NSF STC奖励CCF-1231216)和IBM Research的支持。