

MapBert: 用于实时语义地图生成的按位掩码建模

Yijie Deng^{*1,2,3,4}, Shuaihang Yuan^{*1,2,4}, Congcong Wen^{1,2,4}, Hao Huang^{1,2,4},
Anthony Tzes^{1,2}, Geeta Chandra Raju Bethala^{1,2,4}, Yu-Shen Liu⁵, Yi Fang^{†1,2,3,4}

¹NYUAD Center for Artificial Intelligence and Robotics (CAIR), Abu Dhabi, UAE.

²New York University Abu Dhabi, Electrical Engineering, Abu Dhabi 129188, UAE.

³New York University, Electrical & Computer Engineering Dept., Brooklyn, NY 11201, USA.

⁴Embodied AI and Robotics (AIR) Lab, NYU Abu Dhabi, UAE.

⁵School of Software, Tsinghua University, Beijing, China.

Abstract

空间意识是具身智能体的一项关键能力，因为它使它们能够预测和推理未观察到的区域。主要的挑战在于学习室内语义的分布，这因稀疏、不平衡的对象类别和多样的空间尺度而变得复杂。现有方法难以在实时生成未观察区域方面表现稳健，并且在新环境中泛化效果有限。为此，我们提出了 MapBERT，这是一种旨在有效建模未见空间分布的新框架。我们首次利用一种不需要查找表的 BitVAE，将语义地图编码为紧凑的按位标记，受这种将语义地图的独热编码与位编码的二进制结构自然对齐的观察启发。在此基础上，使用一个掩码变换器来推测缺失区域并从有限的观察中生成完整的语义地图。为了增强对象中心的推理，我们提出了一种对象感知掩码策略，该策略同时遮盖整个对象类别并将其与可学习嵌入配对，从而捕捉对象嵌入和空间标记之间的隐式关系。通过学习这些关系，模型能够更有效地捕捉对于实际机器人任务至关重要的室内语义分布。在 Gibson 基准测试上的实验表明，MapBERT 在实现最先进的语义地图生成方面取得了平衡计算效率与精确重建未观察区域的优异表现。

1. 介绍

由于室内机器人在现实环境中的应用越来越多，像这样的实体代理受到相当多的关注 [20, 23, 36]。由于这些机器人必须在未结构化和不熟悉的空间中有效运行，因此赋予它们强大的感知能力是至关重要的 [12, 28]。语义理解是机器人感知的基本组成部分 [14–16, 37–39, 42]，主要集中于直接可见的物体和区域，通常忽视了代理的即时视野之外的遮挡或未知区域 [1, 7, 17]，例如角落周围或障碍物后面的区域。然而，弥合所见与所未见之间的差距不仅仅是语义感知所能实现的

[30, 40, 41]；它需要一种空间意识，即推理观察到和未观察到区域的能力，从而使代理能够预测环境中尚未探索的区域。

赋予代理空间意识的关键在于学习如何在室内环境中分布对象，这引入了一组独特的挑战，因为特征的稀疏性、类别分布的不平衡性以及不同对象类别的可变尺度在室内环境中特别显著。当前旨在学习室内语义分布的研究工作可以大致分为两种方法。首先，基于完成的方法旨在完成部分观察的场景，以解决机器人有限视野常常导致的局部和不完整信息。这些方法 [10, 22, 27] 能够精细化缺失或遮挡区域，但仍受限于可见环境的范围。第二种更具挑战性的方法 [18, 45, 46] 则涉及生成未观察区域的空间一致语义，基于已检测到的对象和空间布局。近期工作 SGM [45] 利用掩码自动编码器 (MAEs) [13] 来学习完整语义地图中语义的分布，并从当前观察中推断未观察区域。然而，MAEs 通常表现出低样本效率，导致在面对不熟悉的环境布局时泛化和生成能力较差。相比之下，基于扩散的技术 [18] 提供了更强大的生成能力，使得智能体可以预判各种可能的布局。尽管具有这种潜力，但此类技术面临高计算开销和缓慢推理的问题，这阻碍了实际机器人应用的部署。尽管有几项工作 [9, 26, 32] 已努力加速基于扩散的方法，但推理时间和生成质量之间存在权衡。为了解决上述问题，我们提出了 MapBERT，这是一种有效建模未见空间分布的新生成框架。我们的方法包括一个两阶段的流程。首先，我们观察到语义地图通常是独热编码的，这自然与位令牌相一致。利用这一见解，我们首次探索了使用基于比特的表示来编码语义的可能性，而非依赖传统的离散令牌。第二，我们使用一种受 BERT 启发的掩码变换器来推断缺失区域并生成完整的语义地图。为了增强变换器对感兴趣对象的推理能力，我们引入了一种基于对象感知的蒙版策略，以改进标准的随机蒙版。我们的方法并不是随机地蒙版单个块，而是同时蒙住所有与特定对象类型相关的块。这通过可学习的对象嵌入来补充 - 当一个对象类型被完

^{*}Equal contribution.

[†]Corresponding author: Yi Fang, yfang@nyu.edu.

全蒙住时，它的嵌入会与蒙住的输入令牌串联。这种设计帮助转换器更好地学习对象嵌入与其对应空间令牌之间的隐式关系。通过将紧凑的基于比特的表示与这一增强的转换器架构相结合，我们的方法同时实现了计算效率和稳健的性能，从而在实时中实现精确的语义地图生成。我们方法的贡献总结如下：

1. 一种用于语义地图生成的按位掩码建模框架。我们提出了 MapBERT，这是一种新颖的框架，通过无查找的 BitVAE 将语义地图编码为二进制 tokens，并通过基于 BERT 的掩码 transformer 重建缺失区域。
2. 对象感知掩码和嵌入。与传统的随机块掩码不同，我们的方法对整个对象类别进行掩盖，并结合相应的可学习嵌入，从而增强了变压器捕捉固有对象关系的能力。
3. 在语义地图生成方面实现了最先进的表现。我们的方法在 Gibson 室内场景中达到了最先进的性能，展示了在完成未观察区域时的计算效率和卓越的预测准确性。

2. 相关工作

2.1. 语义地图推理

语义地图将物体级语义嵌入到空间布局中，在帮助智能体解读场景和进行环境感知决策中起着关键作用。语义地图大致可以分为两类：语义地图补全 [10, 22, 27] 和语义地图生成 [18, 45, 46]。语义地图补全通常集中在重建未观测或部分观测的区域。一个常用的方法依赖于结合占用估计与语义分割的自上而下投影。另一种方法是重建现有语义地图中的遮挡部分。然而，这些方法中的学习能力有限，这激发了对生成模型的关注，催生了一种有时称为生成语义映射的概念。相比之下，语义地图生成通常通过累积智能体轨迹中的多次观测来迭代地完善预测。这种自监督策略减轻了对大量标注数据的需求，正如 Zhang 等 [45] 所证明的那样。然而，生成能力仍然有限。为此，我们提出了一种新的方法，通过向量量化和掩码 Transformer 架构来利用离散潜在表示，以产生更加丰富和多样的未观测区域预测。通过学习语义标记的代码簿，并在基于 Transformer 的框架中对这些标记进行掩码和预测，我们的方法能够捕捉多种可能的补全，从而应对之前基于 MAE 系统的不确定性输出。

2.2. 生成性掩码建模

由 BERT [8] 开创的掩码语言建模方法涉及以恒定速率随机遮掩部分输入标记，并训练一个双向变压器来重建这些隐藏元素。虽然作为文本编码器效果显著，但这种技术缺乏生成新内容的能力。为了解决这一局限性，Maskgit [2] 引入了一种由调度函数控制调整速率的动态掩码机制，使得通过控制的掩码模式能够迭代生成样本。在此基础上，MAGE [21] 开发了一个结合表示学习和图像合成功能的综合框架。该概念随后由 Muse [3] 得到了进一步扩展，该方法针对文本引导的

图像创建和操作进行了适应，而 Magvit [35] 提出了一个可灵活处理各种视频生成任务的掩码框架。这些生成性掩码建模技术的成功激发了在众多领域的应用。例如，MoMask [11] 和 MMM [24] 将这些原理应用于生成人体运动序列，而 Chen 等人 [6] 则将类似概念应用于图像补全任务。受到这些发展的启发，我们的工作将生成性掩码建模扩展到了室内语义地图生成领域。

3. 方法

在本节中，我们介绍了 MapBERT，这是一种用于语义地图生成的新方法。我们的方法通过采用两阶段架构来应对生成完整语义地图的挑战：首先，我们使用 BitVAE [29] 学习紧凑的离散地图表示，然后我们利用类似 Bert [8] 的掩码转换器从部分观测中进行地图生成。我们方法的综合架构如图 1 所示。

3.1. 问题表述

我们的目标是从室内环境的部分观测生成完整的语义地图。形式上，令 $M \in \mathbb{R}^{H \times W \times C}$ 表示一个完整的独热编码语义地图，其中 H 和 W 表示空间维度， C 表示语义类别的数量。给定这个地图的部分观测，记作 $M_{\text{partial}} \in \mathbb{R}^{H \times W \times C}$ ，其中未观测区域用零掩盖，我们的目标是预测整个环境的语义布局。部分观测 M_{partial} 可以来自各种来源，例如具有语义分割功能的 RGB-D 传感器，但我们的方法对部分观测的具体来源是无关的。

目标是学习一个生成函数 $f: M_{\text{partial}} \rightarrow \hat{M}$ ，该函数不仅能从部分观测中预测完整的语义地图 $\hat{M}$ ，还能够捕捉未观测区域中语义布局的底层分布。通过紧密逼近真实地图 M 的语义分布，该函数具有强大的生成能力——能够从任何给定的部分观测中合成多样且合理的语义地图生成。这是一个具有挑战性的问题，要求深入理解室内环境中固有的结构和语义规律，以对未观测空间进行准确且连贯的预测。

3.2. 使用 BitVAE 学习地图表示

为了有效地表示室内语义地图，我们采用了 BitVAE 架构 [29]，它将语义地图编码到二进制离散潜在空间中。传统的类似 Bert 的掩码建模方法 [6, 11, 24] 通常采用离散码本表示，这需要一个维度为 $N \times C$ 的查找表，其中 N 表示离散码的数量， C 表示每个码的维度。虽然这种基于码本的方法在编码信息密集的信号（如彩色图像和运动序列）方面表现优异，但我们认为它们对于室内语义地图而言并不理想。我们的关键见解是，室内语义地图在其独热编码表示中本质上是稀疏的，仅包含二进制值：0（表示缺失）和 1（表示存在）。这种二进制特性使 BitVAE 特别适合我们的任务。编码过程如图 2 所示，具体过程如下：

一个补丁编码器 $E(\cdot)$ [13] 首先将语义图 $M \in \mathbb{R}^{H \times W \times C}$ 转换为特征图 $m = E(M) \in \mathbb{R}^{h \times w \times b}$ ，其中 b 表示用于编码的比特数量， h 和 w 表示特征图的大小。

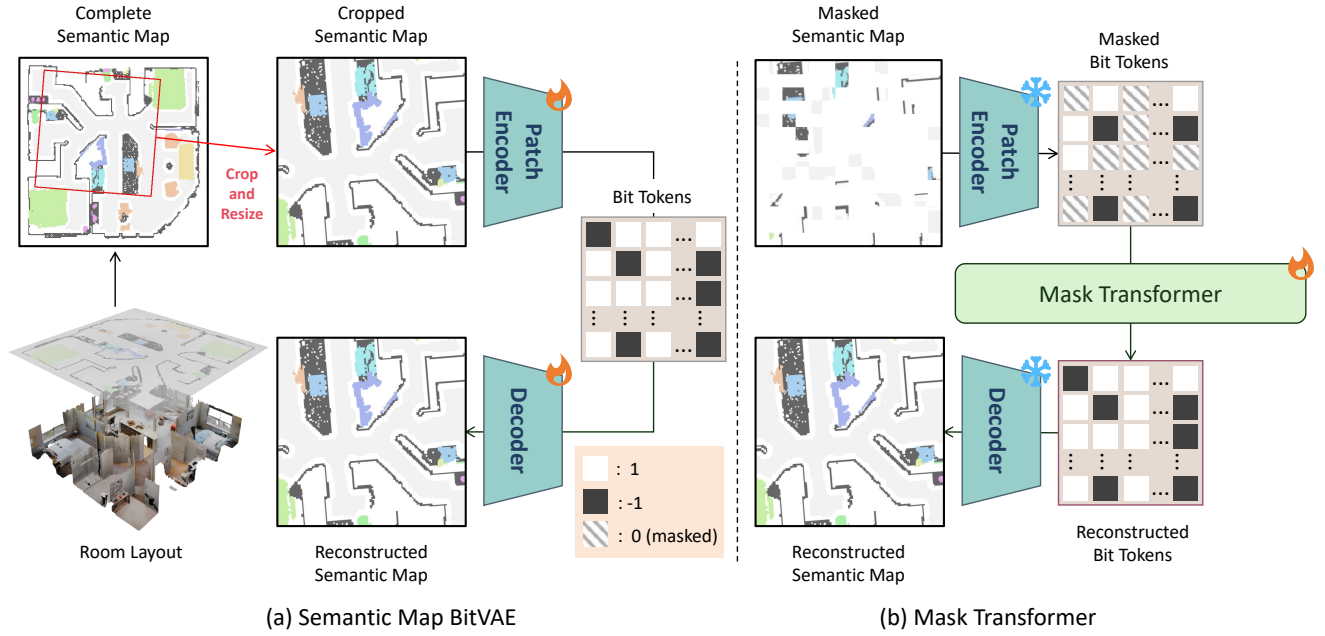


Figure 1. 我们提出的地图生成流程概述。该流程包含两个阶段：(1) 一个无查找的 BitVAE，用于学习语义地图的离散表示，(2) 一个掩码变压器，从部分观测中预测完整的语义地图，可以用于寻找目标。

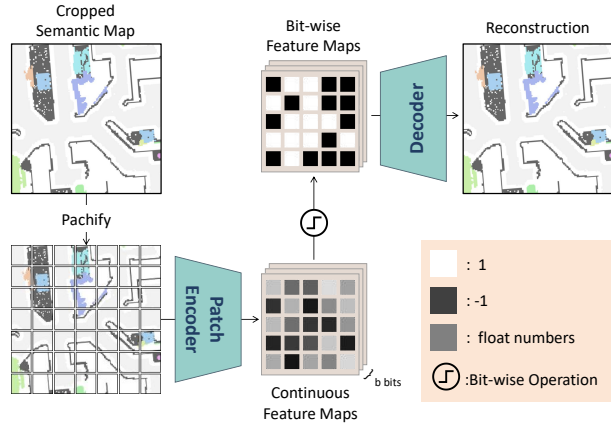


Figure 2. BitVAE 地图重建过程概述。

接下来，特征图经过一个二值化函数 $B(\cdot)$ ，其定义为：

$$B(m_{i,j,k}) = \begin{cases} 1, & \text{if } m_{i,j,k} > 0 \\ -1, & \text{otherwise} \end{cases} \quad (1)$$

，其中 $m_{i,j,k}$ 是位置 (i,j) 处按位特征图 m 的第 k 位。

然后，将得到的二进制特征表示 $B(m) \in \{-1, 1\}^{h \times w \times b}$ 由基于 CNN 的解码器 $D(\cdot)$ 处理，以重建语义图：

$$\hat{M} = D(B(m)) \quad (2)$$

这种 BitVAE 方法提供了两个显著的优点：(1) 由于二进制潜在表示与独热编码地图结构之间的自然对齐，语义地图的重建保真度得到了改善，(2) 与掩码变换器的令牌恢复过程的兼容性增强。二进制表示捕获了更强大的语义特征，这有助于下游学习任务，特别是未观测区域的预测。我们通过在第 4.3 节中的全面消融研究验证了 BitVAE 的有效性。

3.3. 使用掩码变换器的地图生成

在训练 BitVAE 后，我们将其用作语义图的离散标记器。BitVAE 编码器首先将部分语义地图转换为紧凑的二进制表示，这些表示作为我们掩码变换器架构的输入标记。

具体来说，给定对环境的不完全观测，我们通过聚合智能体的观测来构建一个部分语义地图 M_{partial} 。BitVAE 编码器 $E(\cdot)$ 处理该部分地图以生成特征嵌入，然后将其二值化以获得一组索引 $I_{\text{partial}} \in [0, 2^b - 1]$ 。对于部分地图中未观测到的区域，我们分配一个特殊的掩码令牌索引 2^b ，以显式指示缺失的信息。

Mask Transformer 架构遵循 BERT [8] 的成功设计原则，利用双向自注意力来建模室内环境中的复杂空间依赖。它将部分索引 I_{partial} 作为输入，通过学习根据观察到的上下文推断被遮掩的区域来预测完整的索引 I_{complete} 。这使得模型能够捕捉室内空间特有的局部几何约束和全局语义模式。

在训练过程中，我们通过最小化预测索引和真实索

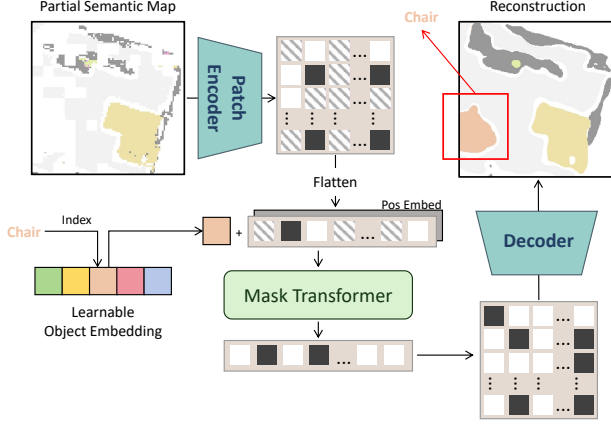


Figure 3. 掩码转换器生成过程概述。

引之间的交叉熵损失来优化遮罩变换器：

$$\mathcal{L}_{\text{MT}} = - \sum_{i,j} \log p(I_{\text{complete}}^{(i,j)} | I_{\text{partial}}) \quad (3)$$

，其中 $p(I_{\text{complete}}^{(i,j)} | I_{\text{partial}})$ 表示模型在给定部分观测条件下在空间位置 (i, j) 处的可能索引的预测概率分布。

3.4. 扩展到对象目标导航

在物体目标导航任务中，代理被提供一个目标物体类别 $c \in \{1, \dots, n\}$ 来定位，其中 n 表示物体类别的总数。我们利用这一目标信息来增强掩膜变换器预测物体位置的能力。如图 3 所示，我们通过维护一个可学习的目标嵌入向量 $E \in \mathbb{R}^{n \times d}$ 来合并目标感知学习，其中 d 是嵌入维度。

在训练过程中，我们使用两种互补的掩码策略处理输入语义图 $M \in \mathbb{R}^{H \times W \times b}$ 。首先，我们随机掩码一部分子区域 $\mathcal{P}_r$ 以促进对整体场景的理解。其次，我们通过随机选择一个类别 c ，并对所有出现该对象的子区域进行掩码 $\mathcal{P}_c$ ($M_{h,w,c} = 1$) 来执行目标特定的掩码。这种双重掩码方法有助于模型学习整体场景结构和特定类别的模式。

然后通过 BitVAE 对掩码映射 M_{masked} 进行编码，以获取位标记 B_{masked} 。这些标记与目标嵌入 E_c 连接在一起，形成输入 $X = [B_{\text{masked}}; E_c]$ 。掩码变换器处理这种组合表示，以预测完整的位标记 $B_{\text{complete}} = \text{MaskTransformer}(X)$ ，其预测是基于掩码观察和目标类别信息进行调节的。

这种目标感知的训练方案使得掩码转换器能够对特定目标类别可能的位置进行预判，从而提高其在导航过程中对目标对象位置做出明智预测的能力。

3.5. 实现和训练

我们在两个步骤中训练我们的 MapBERT，我们将在以下部分详细说明实现细节和训练过程。

对于输入的语义图，我们首先将其调整为 224×224 的分辨率。编码器由一个步幅为 16 的按块 CNN 组成，独立地处理每个块以获取潜在特征。然后这些特征被二值化，每个块为 9 位。解码器实现为一个残差 CNN 结构。我们通过最小化两种损失的组合来训练 BitVAE：

二元交叉熵 (BCE) 损失：给定一个独热编码的语义图 $M \in \mathbb{R}^{H \times W \times C}$ 和网络输出 $\hat{M}$ ，BCE 损失定义为：

$$\mathcal{L}_{\text{BCE}} = - \frac{1}{HWC} \sum_{i,j,c} [M_{i,j,c} \log(\hat{M}_{i,j,c}) + (1 - M_{i,j,c}) \log(1 - \hat{M}_{i,j,c})] \quad (4)$$

地图 IoU 损失：为了直接优化语义分割质量，我们计算原始地图和重建地图之间的交并比：

$$\mathcal{L}_{\text{IoU}} = 1 - \frac{1}{C} \sum_{c=1}^C \frac{|M_c \cap \hat{M}_c|}{|M_c \cup \hat{M}_c|} \quad (5)$$

总损失是加权组合：

$$\mathcal{L}_{\text{total}} = \lambda_{\text{BCE}} \mathcal{L}_{\text{BCE}} + \lambda_{\text{IoU}} \mathcal{L}_{\text{IoU}} \quad (6)$$

我们实现的蒙版转换器是一个多层多头自注意力转换器。输入由位编码器 BitVAE 转换的部分位索引 $I_{\text{partial}} \in \{0, \dots, 2^b - 1\}^N$ 组成，其中 N 是补丁的数量， b 是每个补丁的位数。转换器输出缺失补丁 M 的对数 $L \in \mathbb{R}^{M \times 2^b}$ ，表示所有可能位索引的概率分布。训练目标是 minimize 预测分布与真实索引 I_{gt} 之间的交叉熵损失：

$$\mathcal{L}_{\text{MT}} = - \frac{1}{M} \sum_{i=1}^M \log \left(\frac{\exp(L(i, I_{\text{gt}}^{(i)}))}{\sum_{j=0}^{2^b-1} \exp(L(i, j))} \right) \quad (7)$$

这种表达形式使掩码转换器能够通过离散代码索引预测，学习语义图中观测区域和未观测区域之间有意义的关系。

在掩码变换器的训练中，我们采用了一个两阶段的掩码策略。在第一阶段，持续训练的前四分之一的 epochs，我们仅掩盖一小部分输入位索引 (15 - 20%)。这使得变换器首先学习到全局语义上下文。在第二阶段，我们使用余弦调度逐步增加掩码比例，从 15% 到 75%，这遵循了 BERT 风格的掩码建模中的既定做法 [11, 24]。

4. 实验

4.1. 实验设置

数据集和评估指标。我们在 Gibson 室内场景上评估我们的方法 [33]，遵循以往工作的训练和评估划分 [25, 44]。为了评估从部分观测中生成的语义地图的质量，我们采用：IoU (交集与并集)，它衡量预测的语义地

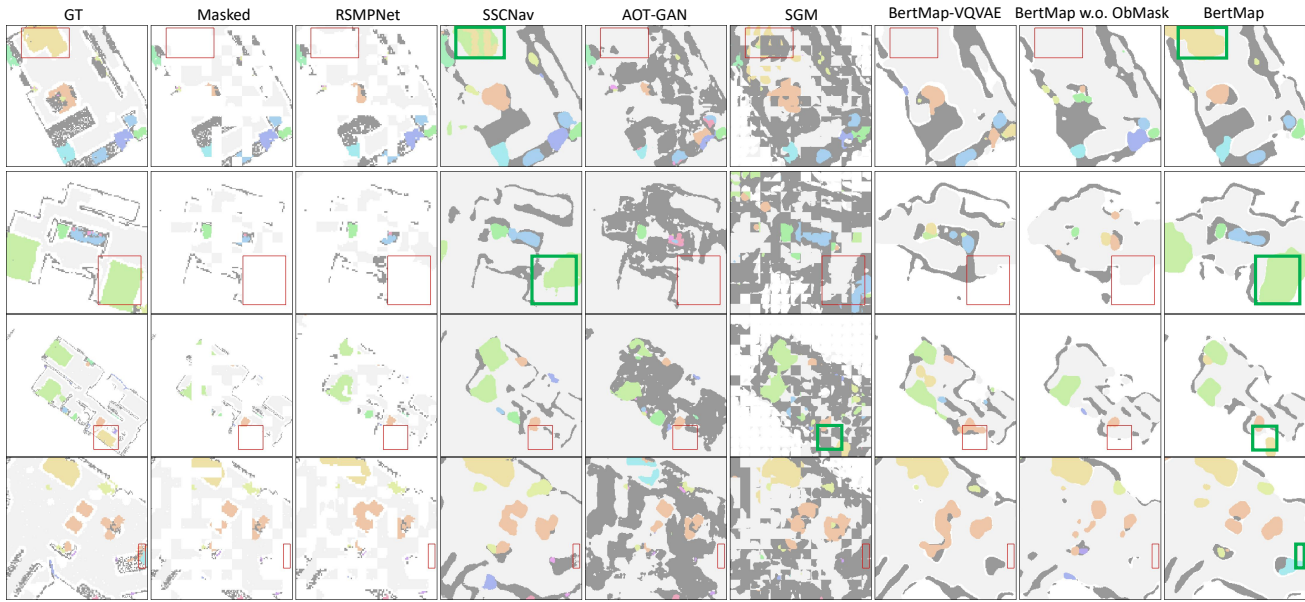


Figure 4. 生成地图质量的定性比较。最左边的两列显示了真实语义地图及其掩模版本，红色框突出显示了使用我们对象感知掩模策略的掩模目标对象。后续列展示了不同方法生成的结果。绿色粗体框表示成功的目标对象定位，而红色框表示定位失败。

图和真实语义地图在包括占用和对象类别的所有通道上的重叠；Recall，正确预测的对象类别与真实类别总数的比率；Accuracy，正确预测的像素（包括对象和非对象）相对于像素总数的比例；F1-score，精度和召回率的调和平均数，强调对象类别预测中平衡的表现。此外，我们引入 sSR（模拟成功率）来评估模型在部分观测语义地图中定位目标对象的能力。对于 sSR，我们应用对象识别屏蔽来遮蔽特定类别的所有像素，然后测量在我们 1,000 个评估子地图中正确识别目标对象位置的成功率。此指标为评估对象定位能力提供了一种高效方法，而无需进行完整的导航模拟。

基线方法和训练设置。为了进行全面评估，我们将我们的方法与几种具有代表性的开源方法进行比较。这些方法包括 RSMPNet [27]（基于 GNN），SSCNav [22]（基于 CNN），AOT-GAN [43]（最初为 RGB 图像修复设计，在此用于语义地图），以及 SGM [45]（基于掩码建模方法）。我们还评估了我们 MapBERT 的两个变体：使用 512 个代码的 VQVAE 量化模块的 MapBERT-VQVAE，以及未使用 ObMask 的 MapBERT，该版本通过随机掩码而不是我们基于对象的策略进行训练。所有基线方法均从头开始训练或使用其发布的检查点进行评估，以确保在相同的掩码条件下进行公平比较。

4.2. 结果

地图生成质量的定量比较。我们使用一个关注对象的遮罩协议来评估语义地图生成的性能。在每个测试案例中，我们通过选择一个对象类别并遮罩所有包含其实例的补丁，以及其他随机补丁来遮罩输入地图的 50%。该方法测试模型从部分观察中生成特定对象位置和更广泛场景结构的能力。如表 1 所示，MapBERT

在关键指标上显示出优越的性能。我们的方法实现了 34.10 % 的 IoU 得分，显著优于之前的方法如 SGM、RSMPNet 和 AOT-GAN。这种改进反映了在生成对象放置和整体场景几何方面的准确性提升。该模型保持了强劲准确率 (41.14 %) 和 F1 得分 (42.30 %)，与如 SSCNav 这样的先进方法相当，表明预测的均衡性很好，误报和漏报都很少。最显著的是，MapBERT 的 sSR 达到了 45.84 %，超过 SGM 的 21.88 % 两倍多。这种语义成功率的显著改善证明了我们方法从部分观察中推断对象位置的卓越能力——这是下游导航任务中的关键能力。

地图生成质量的定性比较。图 1 提供了与表 1 中的定量结果相辅的视觉实例。虽然 SGM [45] 有时能够成功定位目标对象，但其生成的语义地图缺乏连贯的结构和空间一致性。与 SGM 相比，SSCNav [22] 在地图生成和对象定位方面表现出更好的性能。相反，其他基线方法难以有效地完成被遮盖区域或准确预测对象位置。我们的消融研究表明，MapBERT-VQVAE 和没有对象感知遮盖的 MapBERT 可以为被遮盖区域生成可信且一致的图像，但在精确对象定位方面仍有不足。MapBERT 模型在变压器训练期间结合 9 位 BitVAE 和对象感知遮盖，一贯地实现了准确的对象定位。这些结果表明，我们完整版的 MapBERT 模型能够推理并预测场景中的对象位置。

模型效率和规模的定量比较。表 1 展示了模型推理时间和存储需求的比较。我们的 MapBERT 在每个地图输入上实现了 0.011 秒的相对快速的推理速度，同时保持了较小的模型规模 (BitVAE 为 0.1 GB，mask transformer 为 0.31 GB，总计 0.41 GB)。相比之下，SGM 需要显著更多的存储空间，为 1.52 GB。虽然

Table 1. 在 Gibson 室内场景 [33] 数据集上的地图生成质量和模型效率的比较。

Method	Map Generation Quality					Efficiency and Model Size	
	IoU(%) ↑	Recall(%) ↑	Precision(%) ↑	F1(%) ↑	sSR(%) ↑	Inference Time (s)	Model Size (GB)
RSMPNet [27]	26.49	30.09	53.99	37.15	0.00	0.016	0.44
SSCNav [22]	33.56	45.50	40.12	42.64	6.14	0.046	0.20
AOT-GAN [43]	21.58	36.24	40.88	30.72	0.01	0.007	0.06
SGM [45]	25.88	57.06	29.56	35.43	21.88	0.013	1.52
MapBERT-VQVAE	26.55	35.60	31.09	32.61	5.35	0.012	0.42
MapBERT w.o. ObMask	29.32	40.00	36.01	37.90	4.56	0.011	0.41
MapBERT (Ours)	34.10	45.65	41.14	42.30	45.84	0.011	0.41

Table 2. 在 Gibson [33] 数据集上关于量化结构设计选择的消融研究。

Quantization	Code Dim	Codes/Bits	FID ↓
VQVAE	16	256	1.74
		512	1.75
		1024	1.70
BitVAE	—	8	1.75
		9	1.72
		10	1.65

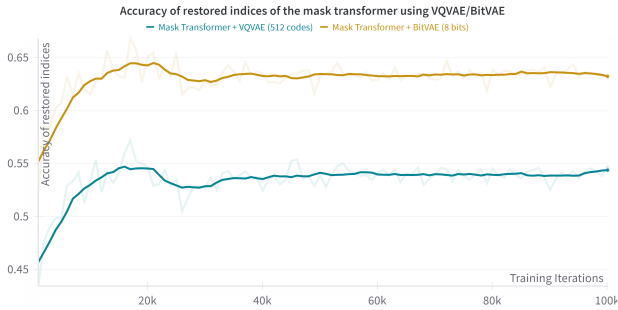


Figure 5. 使用 VQVAE/BitVAE 在采样的 Gibson [33] 数据集上的掩码变换器的恢复 token 索引的准确性训练曲线。蓝色曲线表示使用 VQVAE 训练的掩码变换器在验证数据集上的平滑准确性曲线，而棕色曲线则表示使用 BitVAE 训练的准确性曲线。

AOT-GAN [43] 和 SSCNav [22] 效率更高，但在地图生成质量上有所不足。这些结果表明，MapBERT 不仅在语义地图生成质量上表现优越，还在计算效率和内存占用方面具有实际优势，使其更适合于实际部署。

4.3. 消融研究

VQVAE 与 BitVAE。我们比较了 VQVAE 和 BitVAE 架构，以验证我们选择位编码特征用于语义图的正确性。我们测试了 VQVAE，码本大小分别为 256/512/1024(码维度 16)，并与 BitVAE 的相应 8/9/10 位进行了对比。正如表 2 所示，BitVAE 采用 10 位编

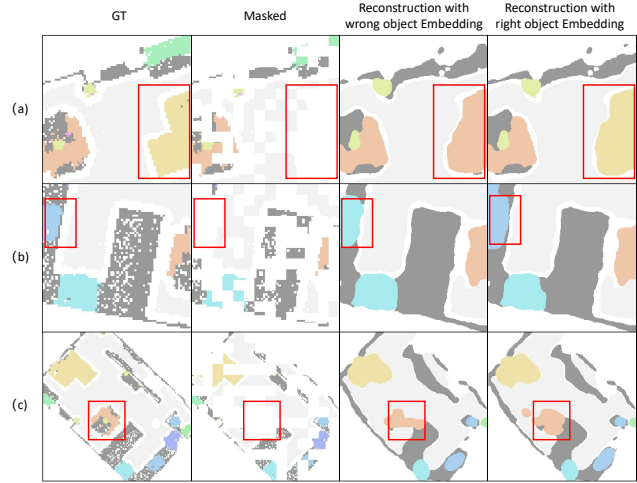


Figure 6. 正确/错误对象嵌入对地图生成的影响。

码时实现了最佳生成质量，证明了直接按位编码的有效性。

对于下游任务，我们使用预训练的 VQVAE (512 码) 和 BitVAE (8 位) 模型，通过我们的掩码转换器来评估这两种架构。图 5 中的训练曲线显示，BitVAE 在恢复被掩盖的标记时使准确率提高了 10%，表明其二进制表示更好地捕捉了对象之间的语义关系。

遮罩变压器模型的大小。我们的实验表明，将变压器模型大小从 ViT-B (768 通道, 12 头/层) 增加到 ViT-L (1024 通道, 16 头/层) 在地图生成质量上仅有微小的改善 (表 3)，这表明较小的架构足以捕捉室内地图语义。

遮罩策略。我们的目标感知遮罩策略相比于随机遮罩显著提高了地图生成质量，如表 3 所示。对于 VQVAE 和基于 BitVAE 的 transformer，分别提升了 3-4 % 的地图质量指标，并将 sSR 分数提高了高达 10 倍，展示了对物体位置更好的推理能力。这一改进源于整合了学习到的物体嵌入——通过遮罩目标物体同时提供其嵌入，transformer 学会了嵌入与 token 索引之间的关系。如图 6 显示，正确的嵌入提高了物体定位的准确性 (行 a-b)，而即使在提供足够上下文的情况下使用错误嵌入，模型仍然保持稳健 (行 c)。这验证

Table 3. 图生成质量的消融实验。VIT-B 表示基础尺寸的掩码转换器，而 VIT-L 表示较大尺寸的掩码转换器。表格中掩码列的 R 表示使用随机掩码策略进行学习，O 表示使用对象感知掩码策略进行学习。

Quantization	MT	Codes/Bits	Mask	Map Generation					
				IoU(%)	Recall(%)	Precision(%)	F1(%)	sSR(%)	
VQVAE	VIT-B	256	R	29.01	39.50	35.59	36.27	8.22	
		512	R	26.55	35.60	31.09	32.61	5.35	
		1024	R	26.73	36.75	32.90	33.44	9.50	
		1024	O	29.47	40.42	36.20	36.81	41.58	
	VIT-L	256	R	28.96	39.41	35.62	36.26	11.39	
		512	R	26.51	35.55	30.96	32.57	5.44	
		1024		R	26.77	36.78	33.00	33.49	10.50
		BitVAE	VIT-B	8	R	29.32	40.00	36.01	37.90
9	R			30.43	40.72	37.22	37.83	6.04	
9	O			34.10	45.65	41.14	42.30	45.84	
10	R			30.10	39.45	36.25	37.78	4.92	
VIT-L	8		R	29.11	39.24	35.81	37.45	4.42	
	9		R	30.43	40.72	37.25	37.81	5.54	
	10		R	29.88	38.44	35.42	36.87	4.43	

Table 4. Gibson 上的目标导航比较。SR(%) = 成功率，SPL(%) = 按路径长度加权的成功，DTS(m) = 到成功的距离。

Method	SR(%) ↑	SPL(%) ↑	DTS(m) ↓
Random	0.4	0.4	3.89
DD-PPO [31]	15.0	10.7	3.24
EmbCLIP* [19]	68.1	39.5	1.15
FBE [34]	48.5	28.9	2.56
ANS [5]	67.1	34.9	1.66
SemExp [4]	71.1	39.6	1.39
PONI [25]	73.6	41.0	1.25
3D-aware [44]	74.5	42.1	1.16
SGM [45]	75.4	39.3	1.26
MapBERT (Ours)	75.8	38.6	1.26

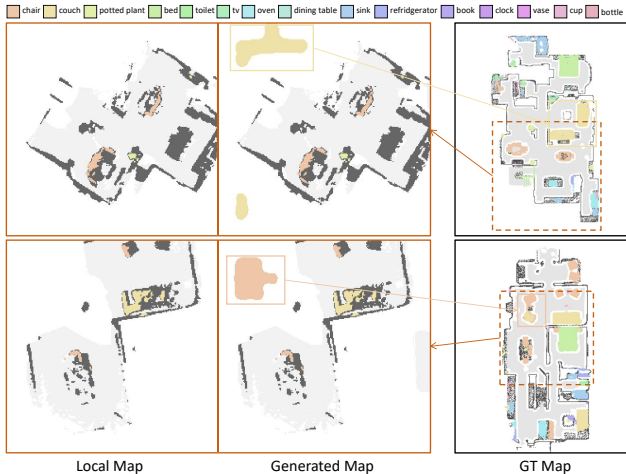


Figure 7. Gibson [33] 上的目标导航结果。从左到右：由智能体的观察构建的局部地图、我们方法生成的语义地图以及真实的全球地图。我们的方法成功地从部分观察中预测出准确的目标位置，从而实现有效的导航。

了我们的策略在增强物体理解的同时，在不完美条件下保持了可靠性。

我们通过将 MapBERT 集成到 SGM 的开源实现中来评估其性能，替换了他们基于 MAE 的地图生成器。如表所示，MapBERT 在导航性能上达到了与之前方法相当的水平。图中展示了在 Habitat 模拟器中成功的物体定位。尽管地图生成质量有所提升，但我们观察到导航性能仅有轻微的提升。这一差距源于两个主要挑战：(1) 随机掩码训练与导航时代理内外探索模式的不匹配，以及 (2) 清晰训练地图与由 RGBD 观察构建的噪声实时语义地图之间的表示差异。未来的工作应侧重于开发更符合导航探索模式的训练策略，并通过增加的训练数据提高对噪声实时观察的鲁棒性。

5. 结论

在这项工作中，我们提出了 MapBERT，这是一种用于在室内环境中从部分观测生成完整语义地图的新框架。通过将用于二元离散表示的无查找表 BitVAE 与受 BERT 启发的掩码变换器相结合，MapBERT 高效地对未观测区域进行建模，并在 Gibson 数据集上取得了最先进的效果。我们的对象感知掩码策略和可学习的对象嵌入进一步增强了变换器捕获复杂对象关系的能力，有助于合理的场景生成。这些进展表明，比特级潜在表示与有效的掩码方案相结合，可以提供高效的推理和高质量的未见区域重建。在未来，我们将探索我们方法的外画生成能力，并提高其对噪声观测的鲁棒性，最终提高对对象目标导航等下游任务的性能。

References

- [1] Seunghyeok Back, Joosoon Lee, Taewon Kim, Sangjun Noh, Raeyoung Kang, Seongho Bak, and Kyoobin Lee. Unseen object amodal instance segmentation via hierarchical occlusion modeling. In 2022 International Conference on Robotics and Automation (ICRA), pages 5085–5092. IEEE, 2022.
- [2] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11315–11325, 2022.
- [3] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704, 2023.
- [4] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258, 2020.
- [5] Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. Neural topological slam for visual navigation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 12875–12884, 2020.
- [6] Haiwei Chen and Yajie Zhao. Don’t look into the dark: Latent codes for pluralistic image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7591–7600, 2024.
- [7] Venkata Naren Devarakonda, Ali Umut Kaypak, Shuaihang Yuan, Prashanth Krishnamurthy, Yi Fang, and Farshad Khorrami. Multitalk: Introspective and extrospective dialogue for human-environment-llm alignment. arXiv preprint arXiv:2409.16455, 2024.
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pages 4171–4186, 2019.
- [9] Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. arXiv preprint arXiv:2410.12557, 2024.
- [10] Georgios Georgakis, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, and Kostas Daniilidis. Learning to map for active semantic goal navigation. arXiv preprint arXiv:2106.15648, 2021.
- [11] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1900–1910, 2024.
- [12] Ruihua Han, Shuai Wang, Shuaijun Wang, Zeqing Zhang, Jianjun Chen, Shijie Lin, Chengyang Li, Chengzhong Xu, Yonina C Eldar, Qi Hao, et al. Neupan: Direct point robot navigation with end-to-end model-based learning. arXiv preprint arXiv:2403.06828, 2024.
- [13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 16000–16009, 2022.
- [14] Hao Huang, Shuaihang Yuan, Congcong Wen, Yu Hao, and Yi Fang. 3d-trans: 3d hierarchical transformer for shape correspondence learning. In 2024 10th International Conference on Automation, Robotics and Applications (ICARA), pages 536–540. IEEE, 2024.
- [15] Hao Huang, Shuaihang Yuan, Congcong Wen, Yu Hao, and Yi Fang. Noisy few-shot 3d point cloud scene segmentation. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 11070–11077. IEEE, 2024.
- [16] Hao Huang, Shuaihang Yuan, Congcong Wen, Yu Hao, and Yi Fang. Weakly scene segmentation using efficient transformer. In 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 9784–9790. IEEE, 2024.
- [17] Yixuan Huang, Jialin Yuan, Chanh Kim, Pupul Pradhan, Bryan Chen, Li Fuxin, and Tucker Hermans. Out of sight, still in mind: Reasoning and planning about unobserved objects with video tracking enabled memory models. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 3108–3115. IEEE, 2024.
- [18] Yiming Ji, Yang Liu, Zhengpu Wang, Boyu Ma, Zongwu Xie, and Hong Liu. Diffusion as reasoning: Enhancing object goal navigation with llm-biased diffusion model. arXiv preprint arXiv:2410.21842, 2024.
- [19] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14829–14838, 2022.
- [20] Keqin Li, Jin Wang, Xubo Wu, Xirui Peng, Runmian Chang, Xiaoyu Deng, Yiwen Kang, Yue Yang, Fanghao Ni, and Bo Hong. Optimizing automated picking systems in warehouse robots using machine learning. arXiv preprint arXiv:2408.16633, 2024.
- [21] Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2142–2152, 2023.
- [22] Yiqing Liang, Boyuan Chen, and Shuran Song. Sscnav: Confidence-aware semantic scene completion for visual semantic navigation. In 2021 IEEE international conference on robotics and automation (ICRA), pages 13194–13200. IEEE, 2021.

- [23] Ismot Sadik Peyas, Zahid Hasan, Md Rafat Rahman Tushar, Al Musabbir, Raisa Mehjabin Azni, and Shahnewaz Siddique. Autonomous warehouse robot using deep q-learning. In *TENCON 2021-2021 IEEE Region 10 Conference (TENCON)*, pages 857–862. IEEE, 2021.
- [24] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2024.
- [25] Santhosh Kumar Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. Poni: Potential functions for objectgoal navigation with interaction-free learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18890–18900, 2022.
- [26] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023.
- [27] Jingwen Sun, Jing Wu, Ze Ji, and Yu-Kun Lai. Rsmptnet: Relationship guided semantic map prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 303–312, 2024.
- [28] Shagun Uppal, Ananye Agarwal, Haoyu Xiong, Kenneth Shaw, and Deepak Pathak. Spin: Simultaneous perception interaction and navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18133–18142, 2024.
- [29] Mark Weber, Lijun Yu, Qihang Yu, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. Maskbit: Embedding-free image generation via bit tokens. *arXiv preprint arXiv:2409.16211*, 2024.
- [30] Congcong Wen, Yisiyuan Huang, Hao Huang, Yanjia Huang, Shuaihang Yuan, Yu Hao, Hui Lin, Yu-Shen Liu, and Yi Fang. Zero-shot object navigation with vision-language models reasoning. In *International Conference on Pattern Recognition*, pages 389–404. Springer, 2025.
- [31] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect point-goal navigators from 2.5 billion frames. *arXiv preprint arXiv:1911.00357*, 2019.
- [32] Zike Wu, Pan Zhou, Kenji Kawaguchi, and Hanwang Zhang. Fast diffusion model. *arXiv preprint arXiv:2306.06991*, 2023.
- [33] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9068–9079, 2018.
- [34] Brian Yamauchi. A frontier-based approach for autonomous exploration. In *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97: Towards New Computational Principles for Robotics and Automation*, pages 146–151. IEEE, 1997.
- [35] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10459–10469, 2023.
- [36] Xianjia Yu, Jorge Pena Queralta, and Tomi West-erlund. Federated learning for vision-based obstacle avoidance in the internet of robotic things. In *2022 Seventh International Conference on Fog and Mobile Edge Computing (FMEC)*, pages 1–6. IEEE, 2022.
- [37] Shuaihang Yuan and Yi Fang. Ross: Robust learning of one-shot 3d shape segmentation. In *proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1961–1969, 2020.
- [38] Shuaihang Yuan, Xiang Li, Hao Huang, and Yi Fang. Meta-det3d: Learn to learn few-shot 3d object detection. In *Proceedings of the Asian Conference on Computer Vision*, pages 1761–1776, 2022.
- [39] Shuaihang Yuan, Congcong Wen, Yu-Shen Liu, and Yi Fang. Retrieval-specific view learning for sketch-to-shape retrieval. *IEEE Transactions on Multimedia*, 2023.
- [40] Shuaihang Yuan, Hao Huang, Yu Hao, Congcong Wen, Anthony Tzes, Yi Fang, et al. Gamap: Zero-shot object goal navigation with multi-scale geometric-affordance guidance. *Advances in Neural Information Processing Systems*, 37:39386–39408, 2024.
- [41] Shuaihang Yuan, Muhammad Shafique, Mohamed Riyadh Baghdadi, Farshad Khorrami, Anthony Tzes, and Yi Fang. Zero-shot object navigation with vision-language foundation models reasoning. In *2024 10th International Conference on Automation, Robotics and Applications (ICARA)*, pages 501–505. IEEE, 2024.
- [42] Shuaihang Yuan, Anthony Tzes, and Yi Fang. Reference convolutional networks for 3d deep point signature learning. In *2024 10th International Conference on Automation, Robotics and Applications (ICARA)*, pages 526–530. IEEE, 2024.
- [43] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Aggregated contextual transformations for high-resolution image inpainting. *IEEE transactions on visualization and computer graphics*, 29(7): 3266–3280, 2022.
- [44] Jiazhao Zhang, Liu Dai, Fanpeng Meng, Qingnan Fan, Xuelin Chen, Kai Xu, and He Wang. 3d-aware object goal navigation via simultaneous exploration and identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6672–6682, 2023.
- [45] Sixian Zhang, Xinyao Yu, Xinhang Song, Xiaohan Wang, and Shuqiang Jiang. Imagine before go: Self-supervised generative map for object goal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16414–16425, 2024.

- [46] Xinxin Zhao, Wenzhe Cai, Likun Tang, and Teng Wang. Imaginenav: Prompting vision-language models as embodied navigator through scene imagination.