
SPATIALLM：训练大规模语言模型用于结构化室内建模

Yongsen Mao^{1*} Junhao Zhong^{1*} Chuan Fang² Jia Zheng¹
Rui Tang¹ Hao Zhu¹ Ping Tan² Zihan Zhou¹
¹Manycore Tech Inc. ²Hong Kong University of Science and Technology
<https://manycore-research.github.io/SpatialLM>

Abstract

SPATIALLM 是一个大型语言模型，旨在处理三维点云数据并生成结构化的三维场景理解输出。这些输出包括建筑元素，如墙壁、门、窗户和具有语义类别的定向对象框。与以往利用任务特定网络设计的方法不同，我们的模型遵循标准的多模态 LLM 架构，并直接从开源 LLM 进行微调。为了训练 SPATIALLM，我们收集了一个大规模、高质量的合成数据集，该数据集由 12,328 个室内场景（54,778 间房间）的点云与真实 3D 标注组成，并对各种建模和训练决策进行了仔细研究。在公开的基准测试中，我们的模型在布局估计方面表现出了最先进的性能，并在 3D 目标检测中取得了具有竞争力的结果。由此，我们展示了一条增强现代 LLM 空间理解能力的可行途径，可应用于增强现实、具身机器人等领域。

1 介绍

3D 室内环境在我们的日常生活中无处不在。我们每天在这样的环境中度过大量时间，并进行各种活动。因此，人工智能的长期目标之一是教机器像人类一样感知、推理和与 3D 室内场景互动。在这项工作中，我们专注于结构化室内建模任务，该任务旨在从原始传感输入（例如，RGBD 扫描）中提取结构化的室内场景描述。具体而言，场景描述包括室内环境中的建筑布局（例如，墙壁、门和窗户）以及 3D 对象的边界盒。这种 3D 结构信息已被证明对场景编辑、增强现实和机器人导航等广泛的现实应用有益。

与网格、体素或隐函数等三维格式相比，结构化场景描述提供了一种高度紧凑但灵活的场景表示。在本文中，我们的第一个贡献是将结构化描述视为通用语言（*i.e.*，Python）的脚本，并提出以文本形式预测该语言，如图 ?? 所示。这一设计选择有几个优点：(i) 它是人类可解释且可编辑的；(ii) 它可以轻松扩展以整合任何新类别，而不会影响脚本中的现有内容；(iii) 它允许我们利用预训练的大型语言模型强大的内置编码能力。

因此，我们的目标是通过直接微调开源的大语言模型（LLMs）来实现我们的方法。虽然现代（多模态）大语言模型已经在自然语言处理、二维图像理解和生成等领域带来了革命性的变化，但很少有工作尝试使用大语言模型进行三维结构化场景建模。这部分是由于缺乏适合该任务的大规模、高质量数据集。在本文中，作为我们的第二个贡献，我们收集了一个新的合成数据集，该数据集由 12,328 个不同场景（54,778 个房间）的点云及其相应的三维结构信息组成，并进行了首次实证研究，以研究将点云输入与大语言模型对齐以进行结构化场景建模的最佳策略。

基于我们的发现，我们引入了 SPATIALLM，一个能够处理点云输入并生成结构化场景描述的大型语言模型。作为我们的第三个贡献，我们展示了通过先在我们的大规模数据集上进行训练，然后在较小的下游任务数据上进行训练，我们的模型在布局估计和 3D 物体检测的公共基准测试中分别表现出竞争力。此外，我们提供了概念验证的零样本实验，展示了我们

*Equal contribution.