分　类　号：　TP391　　单位代码：10183
研究生学号：　2022562014　　密　　级：　公开

吉 林 大 学

硕士学位论文

(学术学位)

面向具身智能的三维室内场景生成方法研究

Research on 3D Indoor Scene Generation Methods for Embodied Artificial Intelligence

作者姓名：姜渤巍

专　　业：计算机科学与技术

研究方向：计算机图形学

指导教师：马锐　副教授

培养单位：人工智能学院

2025 年 05 月

面向具身智能的三维室内场景生成方法研究

Research on 3D Indoor Scene Generation Methods for Embodied Artificial Intelligence

作者姓名：姜渤巍

专业名称：计算机科学与技术

指导教师：马锐 副教授

学位类别：工学硕士

答辩日期： 2025 年 05 月 27 日

摘 要

面向具身智能的三维室内场景生成方法研究

随着 3D 技术在具身智能领域的快速发展，人们对高质量室内场景生成的需求日益增长。室内场景生成技术不仅能够为游戏和虚拟现实构建逼真环境，还可广泛应用于家居设计、室内布局规划等多个领域。尽管基于深度学习的场景生成方法已取得显著进展，该领域在方法和数据集层面仍存在诸多挑战。具体而言，三维场景生成方法本身存在一些瓶颈：首先，现有算法在用户指令理解与场景逻辑一致性保障方面存在显著不足；此外，多数方法生成的布局虽在视觉上合理，却难以满足多机器人协同任务的复杂需求。同时，现有场景数据集与真实室内环境仍存在明显差距。受限于数据标注的局限性，现有数据集难以满足具身智能等前沿领域的研究需求。在机器人技术快速普及的背景下，室内机器人导航将发挥关键作用，而家具布局的合理性将直接影响未来机器人与场景的交互效能。值得注意的是，当前主流数据集如 3D-FRONT 等普遍缺乏针对机器人交互的评估标注。

本文围绕面向具身智能的室内场景生成的方法创新展开深入研究，主要研究内容与贡献如下：

（1）针对现有方法存在的用户指令对齐不足与场景逻辑一致性欠缺问题，本研究提出创新性解决方案 CoT2Scene 框架。该框架基于思维链（Chain-of-Thought，CoT）增强技术，构建了具有强大的空间语义推理能力的场景生成系统，在保证三维空间拓扑合理性的前提下，显著提升了生成结果与用户需求的匹配度。具体而言，本研究首先通过结构化领域知识库提取室内设计规范参数，建立空间约束规则体系；继而利用大语言模型的逻辑推理能力，将抽象用户指令逐步转化为可执行的空间配置方案。特别地，本研究采用思维链提示技术，在初始布局生成阶段保证功能分区合理性，在细粒度配置阶段进行场景的补全，由此生成更贴合用户意图的室内场景，并确保其与用户指令的高度一致性。实验表明，相比基线方法，该方法可确保在严格遵循用户指令指定的空间配置要求条件下，生成逻

辑上更合理的场景布局。

（2）针对现有方法难以生成对机器人友好的室内场景的问题，本研究提出Nav2Scene，一种可部署于现有场景生成框架的即插即用微调机制，通过提升生成场景的导航适配性优化机器人导航效率。创新性地定义路径规划评分（Path Planning Score，PPS）作为评估场景布局导航适配性的新型指标，并训练 ScoreNet 神经网络直接预测给定场景的 PPS 值。鉴于现有场景生成数据集难以满足本研究需求，本方法基于 3D-FRONT 数据集，通过机器人路径规划算法对 4041 个房间进行导航交互性能评估，并标注其是否有利于机器人交互的性能评分，从而构建了机器人路径规划评估数据集。在路径规划评分的引导下，微调后的场景生成器不仅能生成更适配家用机器人导航需求的场景布局，还能在生成质量与多样性方面保持相当或更优的表现。

关键词：

深度学习，室内场景生成，具身智能，大语言模型

Abstract

Research on 3D Indoor Scene Generation Methods for Embodied Artificial Intelligence

With the rapid development of 3D technology in embodied artificial intelligence, there is growing demand for high-quality indoor scene generation. This technology not only constructs realistic environments for gaming and virtual reality, but also finds extensive applications in domains such as home design and interior layout planning. Although deep-learning-based scene generation methods have made significant progress, the field continues to face challenges at both methodological and dataset levels. Current 3D scene generation approaches encounter inherent bottlenecks. First, existing algorithms demonstrate deficiencies in understanding user instructions and ensuring logical scene consistency. Additionally, while most methods produce visually plausible layouts, they fail to address the complex requirements of multi-robot collaborative tasks. Meanwhile, existing scene datasets show significant disparities from real-world indoor environments. Due to limitations in data annotation, current datasets cannot adequately support research needs in cutting-edge fields like embodied intelligence. As robotics technology proliferates, indoor robot navigation will play a crucial role, since the rationality of furniture layouts directly impacts interaction efficiency between robots and environments. Notably, mainstream datasets such as 3D-FRONT lack robot-interaction-specific evaluation annotations.

This dissertation focuses on methodological innovations in indoor scene generation for embodied artificial intelligence. The main contributions are articulated as follows:

(1) To address the limitations of existing methods in user instruction alignment and scene logical consistency, we propose a novel framework designated as CoT2Scene. This framework harnesses Chain-of-Thought (CoT)-enhanced computational architecture to establish a scene generation system endowed with robust spatial-

semantic reasoning capabilities. Through systematic integration of structured domain knowledge bases for extracting indoor design specification parameters, we formulate a spatial constraint rule system that ensures 3D topological rationality while substantially enhancing alignment fidelity with user requirements. The framework specifically capitalizes on the logical reasoning capacities inherent in large language models (LLMs) to methodically transform abstract user instructions into executable spatial configuration plans. Our CoT prompting methodology ensures rational functional zoning during the initial layout generation phase while enabling comprehensive scene realization in the fine-grained configuration phase. This two-stage approach produces indoor scenes demonstrating enhanced alignment with user intent while maintaining superior instructional consistency. Experimental results demonstrate that, compared to baseline methodologies, our approach generates scenographic layouts with enhanced logical coherence under rigorously defined user spatial constraints.

(2) To address the challenge that existing methods struggle to generate robot-friendly indoor scenes, this dissertation proposes Nav2Scene, a plug-and-play fine-tuning mechanism for existing scene generation frameworks, which enhances robot navigation compatibility by optimizing generated scenes' navigational efficiency. We propose the Path Planning Score (PPS) as a novel navigation adaptability metric and develop ScoreNet to predict PPS values for scene evaluation. To address dataset limitations, we create a robot navigation assessment dataset using 3D-FRONT, evaluating and annotating 4,041 rooms with navigation performance scores through path planning algorithms. Guided by PPS, our fine-tuned generator produces robot-navigation-optimized layouts while maintaining competitive generation quality and diversity.

Keywords:

Deep Learning, Indoor Scene Synthesis, Embodied Artificial Intelligence, Large Language Model

目 录

第 1 章　绪论

1.1　研究背景及意义

室内场景生成是指生成具有合理家具摆放和房间布局的过程，具体而言就是将家具摆放在具有固定大小、结构的三维空间中，并满足现实室内场景中的功能约束及物理约束。其中，家具在三维空间中的属性信息可由位置、朝向、类别来描述。室内场景生成任务近年来受到了诸多关注[1]，并被广泛应用于虚拟现实、增强现实、开放世界类游戏以及机器人领域。其中自动化生成减轻了手工制作场景的繁琐重复，设计师们也从中受益，借助自动生成这一工具，可以更快、更方便地生成满足需要的房间布局。因此，自动化室内场景生成任务成为了研究人员关注的热点，并得到了快速发展。

大语言模型（Large Language Model[2]，LLM）的飞速发展催生了诸多下游任务，其中包括利用 LLM 卓越的文本理解和转换能力，实现文本引导的 3D 生成。而思维链（Chain-of-Thought[3]，CoT）的提出也为进一步释放 LLM 的能力提供了指南，CoT 大幅度提高了 LLM 在复杂推理任务上的性能，并且输出的中间步骤方便使用者了解模型的思考过程，提高了大模型推理的可解释性。目前，思维链推理已经成为大模型处理复杂任务的一个常用手段。在上述工作基础上，本文考虑到当前高质量场景数据集规模有限，同时基于监督的学习方法通常需要高昂的训练成本。为此，本研究提出基于目前的 LLM 作为场景布局规划器，进行基于文本的场景生成。同时，这种无需训练的方法还可以进行快速且高质量的室内场景补全。

随着具身智能[4]的概念被提出，将机器人同家居、医疗等生活场景结合，以更好地服务人类，成为人工智能研究发展的热门前沿方向。具身智能（Embodied AI）是指在机器智能领域中，通过将智能算法与物理实体的感知、行动和环境交互相结合，使机器能够以更自然、更智能的方式与环境进行交互和解决问题的能力。其中最常见的是移动机器人，用来为人们提供服务，比如扫地机器人、送餐机器人等。即时定位与地图构建（Simultaneous Localization and Mapping[5]，SLAM）

和路径规划技术作为移动机器人导航中最重要的技术，已经有了多年的发展历程和研究成果。

室内场景的生成在机器人导航领域也扮演着至关重要的角色，尽管关于室内机器人路径规划的研究近年来发展迅速，但切入点均是从机器人路径规划的算法优化出发，令机器人的工作代价值最低，选择路径最短以及行进时间消耗最短等。关于如何优化已有场景，使其更符合具身智能领域中关于机器人路径规划这一关键任务需求的问题，目前尚未得到有效的解决方案。

本项工作采取另辟蹊径的方式，认为生成更适应机器人行走的室内场景不失为一个好的研究方向。在保证室内场景布局合理的前提下，同样达到机器人行进时间消耗最短的效果。甚至，在采用同样路径规划算法驱动时，本研究生成的室内场景在不破坏房间原有基础布局的情况下，能起到更好的导航效果（行进时间消耗更短）。

本研究的工作充分地利用 LLM 的能力，在确保满足用户指令输入的前提下实现了更为高效的室内场景生成，为未来具身智能在人机交互领域的发展提供了一条可行的实现路径。同时，本研究希望能带给机器人的室内导航领域一些新的启发，并间接促进具身智能背景下室内场景生成的多元化发展。

1.2　室内场景生成研究现状

1.2.1　基于统计学的生成方法

目前，室内场景生成领域基于统计学的生成方法已形成多维度技术框架，其核心是通过挖掘物体、空间与人类行为的统计规律构建生成模型。在无样本生成方向，研究者通过概率图模型与空间关系先验实现物理合理的布局，例如利用层次化图网络（如 SceneHGN[6]）分析物体层级关系，或借助马尔可夫随机场优化物体共存概率。然而，这类方法对复杂场景的泛化能力受限于先验数据的质量，且难以融入动态行为影响。

针对人类活动模式的统计建模成为另一热点，通过分析定位数据中的热区分布（如 ActFloor-GAN[7]）或时空行为轨迹（如 iPLAN[8]），生成符合实际使用习惯的功能分区布局，但其对隐私敏感数据的依赖限制了应用场景。

样本驱动方法则聚焦于场景属性的统计分布建模，例如使用高斯混合模型或变分自编码器捕捉物体类型、尺寸的比例规律，或通过“位置指纹”匹配生成风格一致的布局（如 S-INF[9]）。尽管这类方法能快速生成多样化场景，但其生成的场景类型受限于训练样本的覆盖范围。

近年来，隐式统计约束与神经场融合技术崭露头角，例如 S-INF 模型通过解耦全局布局与局部细节的统计特征，结合隐式神经场生成高保真多楼层场景，而 3D 场景图技术则基于点云数据统计实例关系，增强生成结果的结构合理性。这类方法在平衡全局统计规律与局部细节生成上展现出潜力。

当前研究仍面临显著挑战：其一，多数方法依赖大规模标注数据或行为日志，数据获取成本与隐私风险制约了实际应用；其二，动态场景建模能力薄弱，难以适应临时布局调整需求；其三，跨建筑类型的统计先验迁移能力不足，例如住宅与医院场景间的模型泛化存在瓶颈。未来方向可能包括小样本统计学习框架的开发、实时行为数据与生成模型的动态耦合，以及多模态输入（如文本描述）驱动的个性化生成技术。总体而言，统计学方法在提升场景合理性方面已取得实质进展，但需在数据效率、动态交互与跨域适应性上实现突破，以推动室内生成技术向实用化迈进。

1.2.2　基于深度学习的生成方法

目前，室内场景生成领域基于深度学习的方法在场景表示、生成效率和语义理解等方面取得了显著进展。研究主要围绕场景表示形式与生成框架的创新展开。在场景表示上，研究者提出了图结构、矩阵结构、层次结构以及结合全景图与 3D 高斯的混合方法。图结构通过图卷积网络（Graph Convolutional Network[10]，GCN）或 Transformer[11]建模物体间语义关系（如支撑、环绕），增强布局合理性；矩阵结构以稀疏连接网络编码全局布局，虽生成速度快但可解释性较弱；层次结构则模仿人类递进式设计思路，结合变分自编码器（Variational Autoencoder[12]，VAE）分层次生成房间、家具与装饰，优化生成质量。此外，FastScene[13]等研究通过全景图生成多视角图像，结合 3D 高斯泼溅技术重建几何一致的场景，而 DreamScene[14]利用 3D 高斯过滤和相机采样策略进一步提升视角一致性。

生成框架的创新体现在模块化、端到端与多模态协同三方面。例如，SceneX

框架[15]将生成流程拆解为资产生成（PCGHub）和布局规划（PCGPlanner），通过文本指令实现 30 倍效率提升的大规模场景生成；基于 Transformer 或条件 VAE 的端到端模型（如 PlanIT[16]、GRAINS[17]）直接从数据中学习布局规律，无需预定义规则；多模态方法如 FastScene 结合扩散模型生成全景图，再通过深度估计与 3D 重建输出场景，支持文本、草图等多种输入形式。

目前该类方法的技术突破集中在物体关系建模与生成策略优化。传统基于规则或能量函数的布局方法逐渐被深度学习替代，例如图卷积网络自动学习物体间的功能关系，避免了人工定义规则的局限性。渐进式生成策略成为主流，FastScene 将全景图生成分解为小步长位移以减少图像畸变，DreamScene 通过多时间步采样优化 3D 高斯的形状与纹理。此外，结合人类活动生成场景的新兴方法开始探索，例如基于人体姿态或交互轨迹约束家具摆放，但受限于标注数据不足，仍处于初期阶段。程序化生成的可编辑性也得到强化，SceneX 通过 263 个标准化参数实现几何与材质的灵活调控，而 DreamScene 支持通过调整仿射变量或文本提示动态编辑场景对象。

当前方法的优势显著，例如 SceneX 将大规模场景建模时间从数周压缩至数小时，FastScene 单场景生成仅需数分钟，且多模态输入支持提升了用户交互性。物理合理性通过能量函数或碰撞检测等约束得以保障。然而，挑战仍存在：多数方法在非生成视角下存在渲染质量下降，物体间功能关系（如桌椅搭配）的语义理解仍需细化，且模型性能高度依赖数据集规模与标注质量（如 SUNCG 数据集[18]与真实场景的差异）。

未来研究将聚焦于跨模态生成优化、轻量化实时生成、物理引擎集成等方向。例如，结合人体行为数据增强布局的功能性，提升文本、图像与 3D 场景的跨模态对齐精度，或引入物理仿真约束（如重力、碰撞检测）强化真实性。代表性数据集如 3D-FRONT[19]（真实设计师标注）、SceneNN[20]（多模态 RGB-D 数据）的扩展应用，将进一步推动技术从单一布局向复杂交互演进。总体而言，室内场景生成技术正朝着高效化、智能化与可交互化方向发展，但视角一致性、语义理解与数据依赖性等瓶颈仍需突破。

1.2.3　室内场景数据集

室内场景生成技术的发展离不开大量高质量数据的支持，而室内场景数据集作为该技术的基石，起着举足轻重的作用。数据集为机器学习模型提供了丰富的训练样本，使模型能够学习到室内场景的各种特征和规律，从而实现更准确、更自然的场景生成。一个全面、高质量的室内场景数据集，不仅包含各种类型的室内场景，如客厅、卧室、厨房等，还涵盖了不同风格、不同布局以及丰富的物体类别和属性信息。这些数据能够帮助模型学习到不同场景下物体的摆放规则、空间布局特点以及物体之间的语义关系等。例如，通过对大量客厅场景数据的学习，模型可以掌握沙发、茶几、电视等物体在客厅中的常见摆放位置和搭配方式，从而在生成客厅场景时能够生成符合人们生活习惯和审美需求的布局。同时，数据集的多样性和规模也直接影响着模型的泛化能力，丰富多样的数据集可以让模型学习到各种不同的场景情况，使其在面对新的、未见过的场景生成任务时，能够更好地发挥作用，生成合理且逼真的室内场景。因此，深入研究和不断完善室内场景数据集，对于推动室内场景生成技术的发展具有至关重要的意义，是实现更智能、更逼真室内场景生成的关键所在。

目前，室内场景数据集主要可分为真实采集数据集和合成数据集两大类[21]。这两类数据集在数据获取方式、数据特点以及应用场景等方面存在显著差异。

真实采集数据集通过实际的传感器设备对真实室内场景进行扫描和采集，能够真实地反映现实世界中的室内场景情况，具有高度的真实性和可靠性。但数据采集过程往往较为复杂，成本较高，且数据的标注和处理难度较大。Nathan Silberma[22][23]等人利用微软 Kinect 设备，从不同城市的室内场景中采集了 464 个短 RGB-D 序列，经过筛选和标注，最终得到包含 1449 张带有像素级标注的图像数据集 NYU-Depth V2。Jianxiong Xiao 等人提出 SUN3D 数据集[24]，用以解决真实采集数据中物理关系和 3D 结构缺失的问题。它通过开发交互式重建管道，对 41 栋建筑中的 254 个不同空间进行处理，成功恢复了这些空间的 3D 场景结构。在数据采集过程中，不仅获取了 RGB-D 图像，还对其中 8 个场景提供了语义标签，这些语义标签详细标注了 3D 点云的类别信息，以及相机的位姿信息，为后续的研究提供了重要的基础数据。ScanNet[25]是目前规模较大且具有重要影响力

的真实采集数据集。它是由 Angela Dai 等人利用专门设计的易于使用和可扩展的 RGB-D 捕获系统，从超过 1500 次扫描中获取了约 250 万视图的数据。这些数据涵盖了办公室、商店、学校等多种不同类型的室内环境，具有广泛的代表性。该数据集对每个扫描场景都进行了详细的标注，包括 3D 摄像机姿态、表面重建信息以及实例级语义分割信息。

合成数据集则是通过计算机图形学技术和人工设计生成的，具有可控性强、数据规模易于扩展、标注相对简单等优点，但与真实场景相比可能存在一定的差距。苹果公司的 Mike Roberts 等人于 2020 年提出了一个专注于室内场景理解的合成数据集 Hypersim[26]，它利用专业艺术家创作的场景，生成了高质量的合成图像。该数据集提供了逐像素的深度、照度和反射率等标签，以及丰富的几何信息。这些标签和信息为室内场景的多任务学习提供了全面的数据支持。Jia Zheng 等人提出 Structured3D[27]，是一个为结构化 3D 建模提供支持的大型照片级数据集。它包含 3500 个由专业设计师创建的房屋设计，这些设计包含各种真实的 3D 结构注释和生成的照片级 2D 图像。数据集提供了渲染图像和相应的地面真实标注，如语义、反照率、深度、表面法线、布局等。下面介绍本研究采用的开源场景合成数据集，3D-FRONT 数据集[19]（如图 1.1 所示）。

图 1.1　3D-FRONT 的室内布局可视化示意图

3D-FRONT（3D Furnished Rooms with layOuts and semaNTics）数据集是阿里巴巴集团与清华大学、西蒙弗雷泽大学等机构联合研发的室内场景数据集，于2020 年首次发布，旨在填补大规模高质量 3D 场景布局数据集的空白。其核心数据来源于阿里巴巴旗下家装设计平台“躺平设计家”积累的海量真实家居设计方案，结合工业级 3D 建模流程生成。

数据集的构建融合了激光扫描、摄影测量等技术，并通过人工验证确保设计风格与布局的合理性。该数据集源于阿里巴巴淘系的工业生产流程，兼具质量与数量优势。该数据集包含 6,813 个真实户型，平均每个户型涵盖 7 个房间场景，总计 19,775 个精细化设计的室内场景，每个场景包含三个基本要素，分别是家具模型，来自阿里巴巴开源的 3D-FUTURE[28]数据集，覆盖现代简约、古典奢华等多种风格，模型具备高精度几何结构与纹理细节，图 1.2 展示了其中包含的客厅样本的布局；语义标注，包括房间类型（如卧室、客厅）、家具类别、布局关系等结构化信息，支持计算机图形学任务的训练。以及数据格式，以 JSON 文件存储户型设计参数，用户可通过编程工具（如 Pandas）解析。

图 1.2　3D-FRONT 中客厅样本可视化示意图

3D-FRONT 数据集主要服务于计算机图形学研究，用于训练室内场景理解（如物体检测、语义分割）、3D 场景生成等模型。在场景合成任务中，3D-FRONT 数据集的多样性和高质量数据能够为模型提供充足的学习样本，帮助模型学习到不同场景下家具的合理布局和搭配方式，从而生成更符合实际需求和审美标准的室内场景。例如，HouseCrafter 项目[29]基于该数据集实现了平面图到 3D 场景的自动化生成。该数据集也可用于虚拟现实与游戏开发，提供高质量室内场景资源，支持虚拟环境快速搭建。同时，数据本身也支持室内设计自动化，可辅助开发智能设计工具，优化空间布局与风格搭配效率。

表 1.1 展示了 3D-FRONT 与其他室内场景数据集的比较，它将 3D-FRONT

同目前开源的其他 4 个室内场景数据集进行了多个维度的比较。从中能够看出，3D-FRONT 是其中房间规模最大的室内场景数据集，同时包含更多的细粒度特征如纹理、物体类别、场景分类，这为室内场景生成任务提供了重要的数据支持。特别是在物体类别的对比上，该数据集是五个数据集中唯一具备物体类别的标注的数据集。该数据集不仅能帮助研究者们深入理解室内场景的布局设计，还能够为室内机器人的路径规划和避障提供模拟环境，推动具身智能领域在室内导航方向上有更多的研究成果和进展。

表 1.1　3D-FRONT[19]与其他场景数据集对比

数据集	Standford Scenes[30]	SceneNet[31]	ScanNet[25]	Structured3D[27]	**3D-FRONT**
三维网格	✓	✓	✓	×	✓
三维语义网格	✓	✓	✓	×	✓
纹理	✓	×	✓	×	✓
物体类别	-	-	-	-	**34**
场景分类	-	5	19	-	**28**
房屋	None	None	None	3500	**6813**
房间	130	57	1513	21835	**51708**

此外，该数据集优点包括但不限于规模与多样性，它的场景数量远超同类数据集（如 SUNCG、ScanNet），覆盖户型、家具风格、房间类型更全面。同时，数据本身达到工业级质量，它依托阿里巴巴实际业务数据，模型细节（如纹理、几何精度）达到工业应用标准。最后为开源性与易用性，该数据集免费开放，提供结构化标注和配套模型库，降低学术研究门槛。

不过数据集本身也存在诸多限制，比如合成数据偏差：场景为人工设计生成，可能缺乏真实环境中的复杂变量（如光照变化、动态物体），影响模型在真实场景的泛化能力；处理复杂度高：JSON 格式需特定解析工具，对非技术用户存在使用门槛。动态更新不足：数据集自 2020 年发布后更新较少，难以覆盖新兴设计趋势与技术需求（如智能家居布局）。

总而言之，3D-FRONT 凭借其大规模、高精度的特点，已成为室内场景生成与理解领域的标杆数据集。然而，其合成数据特性与静态更新机制仍需结合真实场景数据补充，以进一步提升应用广度。

1.2.4　研究现状总结

现有的室内场景生成主要基于完全的数据驱动来进行的，它存在如下问题：

问题 1：目前的场景生成方法绝大多数为数据驱动的方法，最开始提出的基于规则进行设计的方法，由于其实现成本过高，工程过于复杂而不容易被采用。而基于数据驱动的方法受数据自身限制较大。同时，现有方法在用户指令理解与场景逻辑一致性保障方面存在显著不足，在具身智能框架下，场景生成不仅需要满足视觉真实性，更需融入用户交互性和环境适应性。如何实现布局更加合理而又符合用户输入需求的室内场景生成是一个值得思考的问题。

问题 2：随着具身智能的进展，机器人越来越多地融入到家庭和医疗保健环境中，需要不仅适合人类而且对机器人友好的房间布局。然而，现有的场景生成方法采用的训练数据绝大部分来自设计师们设计的房间布局，主要侧重于人类视角，而忽视了机器人可行高效导航的功能要求，如路径规划和避障。目前场景生成领域采用的数据集大多数来自设计师们预先设计好的布局配置或真实拍摄的 RGB-D 图像，相应标注仅仅为物体的相应属性以及平面图，缺乏关于机器人导航算法的评估结果。这对本研究利用现有数据集生成利于机器人高效导航的室内场景构成了一定挑战。

1.3　主要研究内容

本文针对上述问题对面向具身智能的三维室内场景生成方法研究进行了以下几方面的研究：

（1）针对问题 1，本文提出了 CoT2Scene，一个基于思维链增强的大语言模型的分步式场景生成框架，这种框架可以在保证空间连贯性的同时提升生成质量。本研究利用思维链（Chain-of-Thought，CoT）技术，可以快速引导模型进行指向性输出。具体过程通过一套精心设计的层叠样式表（Cascading Style Sheets，CSS）格式的提示模板完成。为了保证场景生成质量，本研究加入一些空间关系的强约束，并基于 3D-FRONT 构建外部知识库。通过调用知识库中的室内设计先验知识引导 LLM，生成更贴合用户意图的室内场景，并确保其与用户指令的高度一致性。

（2）针对问题 2，本文提出了 Nav2Scene，一种新颖的即插即用的微调机制，它可以部署在现有的场景生成器上，以提高生成的场景对于有效的机器人导航的适应性。本文综合考虑体现机器人在室内场景中的工作效能的两方面因素，即任务时间和任务完成度，首次提出了路径规划得分（Path Planning Score，PPS）作为评价室内场景布局是否适合机器人导航的指标，并训练网络对给定场景的 PPS 进行预测。基于训练后的网络，本方法对现有的场景生成框架进行微调，最终可以生成具有增强的家用机器人导航兼容性以及优良的质量和多样性的场景。本文在 3D-FRONT 数据集基础上，提取每个房间的布局（包含家具及对应房间的平面图），生成相应的二值图（黑色值代表边界和障碍物，白色值代表可活动区域），由此构建了包含 6041 张图像的场景数据集。同时，在问题 2 的背景下，基于现有路径规划算法搭建 ROS[32]系统，实现机器人路径规划算法并给出计算得到的 PPS，以此作为场景数据集的新标签。

1.4　论文组织结构

图 1.3 展示了本文的组织结构，主要展示了本文的技术路线及其相互关系：

第一章，绪论。本章节首先介绍了室内场景生成的研究背景与意义。然后介绍了室内场景数据集、室内场景生成方法、室内路径规划算法的研究现状。最后介绍了本文的主要研究内容以及组织结构。

第二章，相关理论基础。本章节首先介绍了本文相关生成模型的基本原理和关键技术，其中包括 Transformer、扩散模型以及思维链技术。然后列出了场景生成图片、场景家具分布、场景真实性以及路径规划算法的评价指标。

第三章，基于思维链增强大语言模型的室内场景生成方法。本章节首先分析了当前的室内场景生成算法在与用户指令对齐和确保逻辑场景一致性方面的局限性，说明了对于具身智能领域在用户交互方面的制约。然后提出了 CoT2Scene 框架，该框架包含基于 3D-FRONT 数据集的外部知识库构建，以及利用思维链（CoT）这一提示词工程技术增强 LLM，最终生成满足用户输入的文本描述的三维室内场景。最后本章节通过在 3D-FRONT 数据集上的消融与对比试验，验证了所提出方法的有效性。

第四章，基于机器人路径规划约束的室内场景生成优化。本章节首先分析了

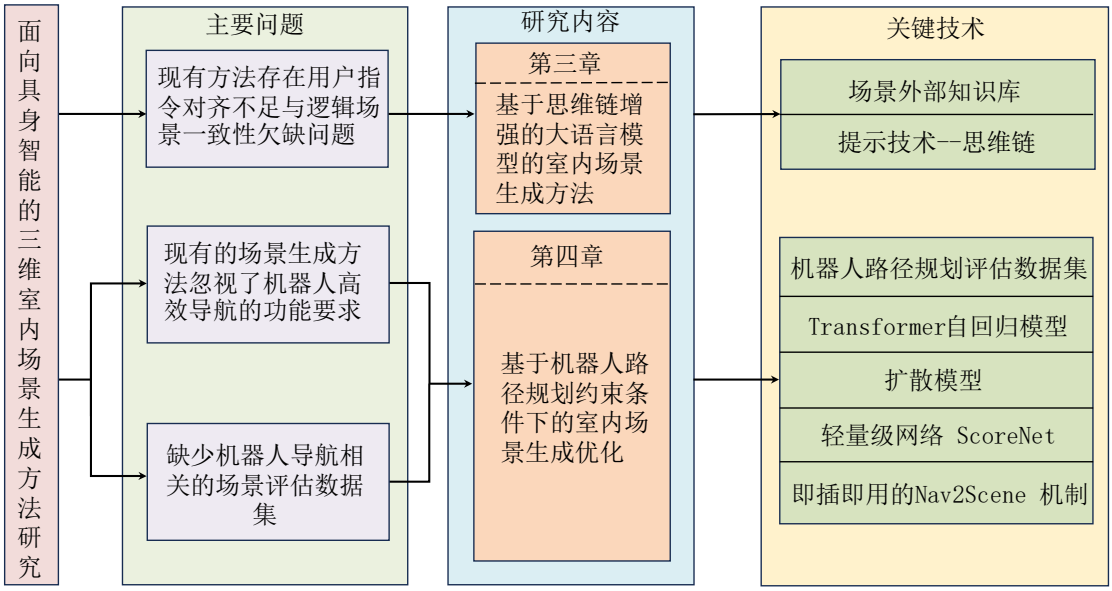

图 1.3　本文的组织结构示意图

现有室内场景生成方法的不足。然后基于两个目前表现最优的场景生成方法提出了优化框架，即插即用的 Nav2Scene 机制，该机制基于机器人规划评估数据集预训练一个能够进行路径规划评估的轻量级网络 ScoreNet，连同生成模型进行微调训练。为此，本章节阐述了机器人规划评估数据集的背景和构建过程。本章节通过在 3D-FRONT 数据集上的定量和定性实验，证明了本研究的方法能够生成对机器人友好的室内场景。最后，本章节通过对 ScoreNet 进行的消融实验，验证了所提出网络的有效性。

第五章，论文工作总结与展望。本章节对本文的全部工作进行总结，并且对未来工作进行展望。

第 2 章　相关理论基础

本章主要概述了与本文后续研究工作相关的理论基础。首先介绍了Transformer的编码器网络、扩散模型DDPM网络以及思维链（Chain-of-Thought，CoT）的基本原理；然后介绍了场景生成图片、场景家具分布、场景真实性以及路径规划算法的评价指标，这些指标用来衡量模型的性能。

2.1　Transformer 基本原理

Transformer 最初于 2017 年在论文《Attention Is All You Need》[11]中提出，旨在解决机器翻译中长序列依赖问题。它摒弃了传统的循环神经网络（Recurrent Neural Network[33]，RNN）和卷积神经网络（Convolutional Neural Network[34]，CNN）架构，以自注意力机制为核心，为自然语言处理领域带来了重大变革。此后，Transformer 在众多领域广泛应用。在自然语言处理中，用于文本生成、问答系统、情感分析等任务，如 GPT 系列模型基于 Transformer 实现了强大的语言生成能力。在计算机视觉领域，用于图像分类、目标检测、图像生成等，如 ViT 模型[35]将 Transformer 应用于图像领域取得了良好效果。在语音处理中，也用于语音识别、语音合成等任务，推动了各领域技术的快速发展。

Transformer 主要由编码器和解码器两部分组成。编码器负责对输入序列进行编码，提取特征。如图 2.1 所示，它包含多个堆叠的编码器层，每个编码器层又由多头注意力机制（如图 2.2 所示）和前馈神经网络组成。自注意力机制能并行计算每个位置与其他位置的依赖关系，确定当前位置的重要性权重，从而捕捉长序列中的语义信息。多头自注意力则是通过多个头的并行计算，捕捉更丰富的特征。前馈神经网络对自注意力的输出进一步处理，进行非线性变换。解码器与

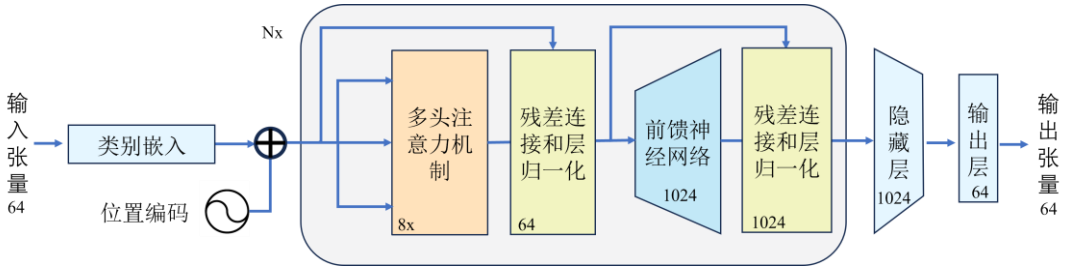

图 2.1　Transformer 的编码器结构示意图

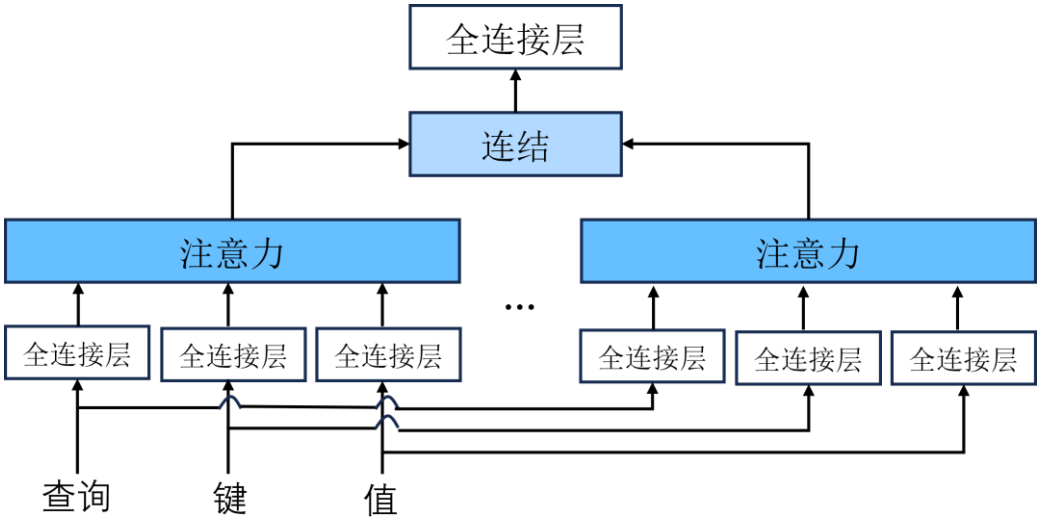

图 2.2　多头注意力机制示意图

编码器结构相似，但多了一个掩码多头自注意力机制，用于防止解码器在生成当前位置时提前看到未来信息。解码器在生成输出时，会结合编码器的输出和自身已生成的历史信息，通过自注意力和交叉注意力机制来预测下一个位置的输出。

Transformer 最初是设计用于自然语言处理（Natural Language Processing，NLP）中的机器翻译任务，之后不局限于 NLP 领域，而是扩展到各种序列生成类任务。同样的，室内场景生成任务的本质也属于序列生成，Transformer 也在其中发挥着重要作用。

首先，可将室内场景的各种元素，如家具、墙壁、门窗等的位置、形状、材质等信息进行编码，连同编码后的房间平面图形成词嵌入，输入如图 2.1 所示的编码器，通过多头注意力机制捕捉这些元素之间的空间关系和语义关联。例如，能理解沙发与茶几通常的相对位置关系等。

然后，在生成阶段，根据这些编码信息和一定的生成规则，通过多层感知机进行属性提取，逐步预测下一个物体（家具）的种类、尺寸、位置等信息。此外，基于 Transformer 中的注意力机制，网络更容易地可以捕捉到房间的家具布局之间的空间关系及属性之间的相关性。最终，通过不断学习大量的室内场景数据，Transformer 可以学习到不同元素组合的规律，从而生成多样化、符合实际语义和空间关系的室内场景。

而且，Transformer 的并行计算能力使得室内场景生成的效率得到大幅提升，能够快速生成高质量的室内场景，为室内设计、虚拟现实等领域提供有力支持。

2.2　扩散模型基本原理

DDPM（Denoising Diffusion Probabilistic Models）即去噪扩散概率模型[36]，源于对非平衡热力学的启发。它为生成式模型领域带来了新的思路和方法，与传统生成对抗网络（Generative Adversarial Network[37]，GAN）、变分自编码器等生成模型有不同的建模方式。作为扩散模型[38]中的经典代表，DDPM 同样包括两个过程：前向过程（Forward Process）和反向过程（Reverse Process），其中前向过程又称为扩散过程（Diffusion Process），如图 2.3 所示。无论是前向过程还是反向过程都是一个参数化的马尔可夫链（Markov Chain[39]），其中反向过程可以用来生成数据。具体而言，若输入为图像，则对清晰图像逐步加噪，再学习去噪过程来实现图像生成，即先将图像逐渐变成高斯噪声，然后再从噪声中逐步恢复出图像。因此，在 DDPM 中，训练的目标就是让模型能够准确预测每一步所添加的噪声，从而实现去噪。

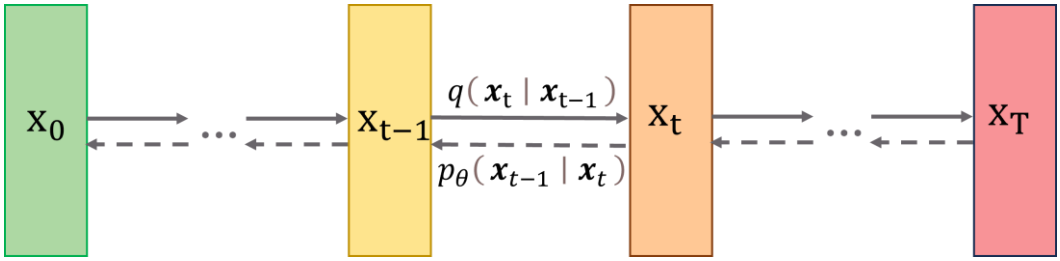

图 2.3　前向过程和反向过程示意图

如图 2.4 所示，DDPM 的网络结构主要基于 U-Net[40]架构，具有编码器和解码器的对称结构，中间通过跳跃连接相连。宏观来看，U-Net 结构中的编码器负责提取不同尺度的特征，解码器则利用这些特征和跳跃连接的信息逐步恢复图像细节，实现去噪和生成的功能。

具体而言，扩散生成模型由一个非参数正向加噪过程和可学习的反向去噪过程组成。前向过程将数据点从$q(x_0)$逐步加噪为一连串噪声越来越大的隐变量：$q(\boldsymbol{x_{1:T}} \mid \boldsymbol{x_0}) = \prod_{t=1}^{T} q(\boldsymbol{x_t} \mid \boldsymbol{x_{t-1}})$。之后，可以训练一个神经网络，通过反复对它们进行去噪处理，来逆转这一过程：$q(\boldsymbol{x_{1:T}} \mid \boldsymbol{x_0}) = \prod_{t=1}^{T} q(\boldsymbol{x_t} \mid \boldsymbol{x_{t-1}})$。其中，$c$是指导逆向过程的操作条件，用于根据需要引导逆向过程。这两个过程在足够大的T条件下，这两个过程应该可以接受$p(\boldsymbol{x_T}) \approx q(\boldsymbol{x_T}|x_0)$。然后，生成模型通过最

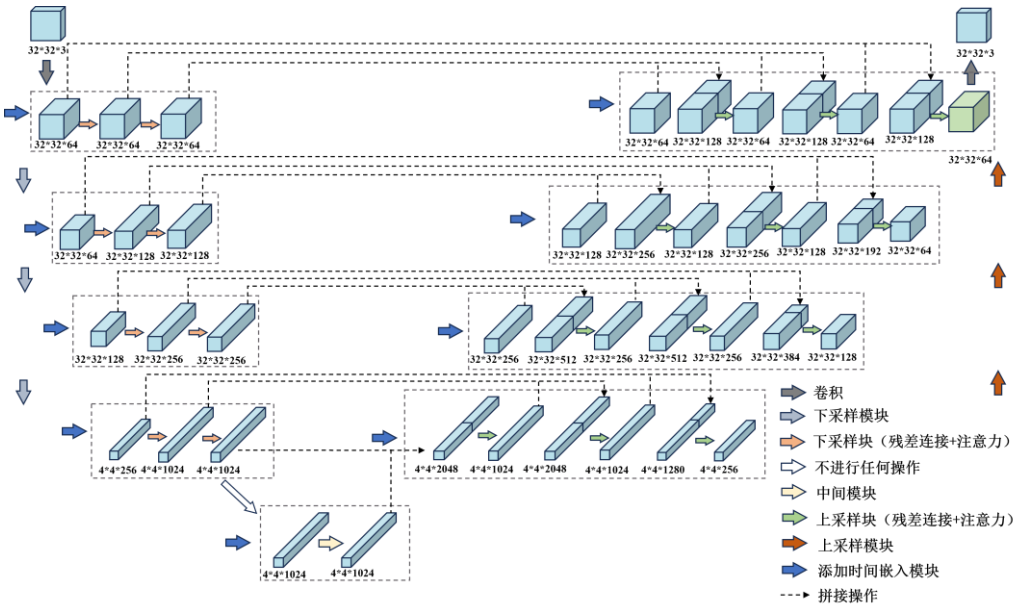

图 2.4　U-Net 结构示意图

小化$E_{q(\boldsymbol{x_0})}[-log\, p_\psi\,(\boldsymbol{x_0})]$的变化上界来进行优化。这里可将求解过程整理为如下公式：

$$\mathcal{L}_{vb} \coloneqq \mathbb{E}_{q(\boldsymbol{x}_0)}[D_{KL}[q(\boldsymbol{x}_T \mid \boldsymbol{x}_0) \parallel p_\theta(\boldsymbol{x}_T)] + \sum_{t=2}^{T} D_{KL}\big(q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)\big) - \mathbb{E}_{q(\boldsymbol{x}_1 \mid \boldsymbol{x}_0)}[log\, p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)]] \quad (2.1)$$

DDPM 在多个领域都有广泛应用。在图像生成领域，能够生成高分辨率、高质量的自然图像，如人脸、风景等，生成的图像具有丰富的细节和多样性。在医学图像领域，可用于生成合成医学图像，辅助医学研究和诊断，比如生成特定病变的医学影像，帮助医生更好地理解病情。在视频生成领域，也有一定的应用潜力，尝试生成连贯、流畅的视频内容。此外，在文本生成、音频生成等领域，DDPM 的思想也在被探索和应用，为多模态数据生成提供了新的途径。

在室内场景生成任务中，DDPM 可以对室内场景的图像或三维表示进行建模。对于图像形式的室内场景生成，首先将室内场景图像作为原始数据$\boldsymbol{x}_0$输入到正向扩散过程中，逐步添加噪声。在反向去噪生成时，网络通过学习大量的室内场景图像数据，能够捕捉到室内场景的各种特征和规律，如不同家具的样式、布局，色彩搭配等。利用 U-Net 结构的多尺度特征提取和融合能力，从加噪的图像中逐步恢复出清晰、合理的室内场景图像，生成具有真实感的室内装修效果图等。对于三维室内场景生成，可以将三维场景的体素表示或点云表示等作为输入数据。

DDPM 在正向扩散过程中对这些数据进行噪声添加，在反向去噪时，通过学习三维空间中物体的位置关系、形状特征等，逐步生成完整、合理的三维室内场景，为室内设计和虚拟现实等应用提供高质量的三维场景模型，帮助用户提前直观地感受室内空间布局和设计效果。

2.3　思维链基本原理

提示词工程中的思维链（Chain-of-Thought，CoT）技术是一种通过分解问题步骤、引导模型逐步推理的革新性方法，其发展与应用深刻改变了语言模型的逻辑推理能力。

CoT 的雏形可追溯至 2022 年谷歌团队的研究，研究者发现传统“少样本提示”仅提供输入-输出示例，难以解决多步骤推理问题。为此，他们提出在提示中嵌入中间推理步骤（即思维链），例如数学题解答时加入公式推导过程。这一方法显著提升了模型在复杂任务中的表现，如 GSM8K 数学基准测试[41]中，PaLM 模型的准确率从 17.7%跃升至 78.7%。后续研究进一步扩展了 CoT 的边界，例如苏黎世联邦理工学院提出的思维图，将线性链升级为图结构，支持思维回溯与融合，使模型在排序任务中的准确率比传统 CoT 提升 70%。至此，CoT 从单一链式推理发展为支持动态路径规划的通用框架。

CoT 的核心价值在于解决需多步骤逻辑推演的任务，主要应用场景包括：

数学与符号推理：通过分解运算步骤，模型可处理含加减混合运算的数学题，例如“食堂苹果数量增减问题”的准确率提升至 90%以上。

常识与逻辑推理：在“数三退一”等游戏中，CoT 帮助模型模拟人类淘汰逻辑，避免直接输出错误结论。

机器人任务规划：将动作指令分解为环境感知、路径规划、执行反馈等子步骤，提升决策可靠性。

数据聚合任务：例如文档摘要时，CoT 可引导模型分阶段提取关键词、重组语义单元，最终合成精简文本。

在基于文本的室内场景生成中，CoT 通过结构化推理显著提升生成质量与可控性。首先通过 CoT 将模糊描述（如“现代简约客厅”）分解为空间布局、家具类型、材质色调等子目标，逐一生成对应元素。例如，模型首先生成“L 型沙发

+圆形茶几”的布局方案，再根据“简约”风格过滤雕花装饰等冗余设计。其次，采用 CoT 的图结构思维，可以允许模型回溯调整冲突设计。例如当用户指定“小空间放置大书柜”时，模型可触发“空间缩放”子任务，自动优化家具比例。同时，通过思维链串联材质、色彩、照明等属性，本研究能够确保北欧风场景中“原木色地板”与“布艺沙发”的视觉统一性，避免风格混杂。

CoT 与室内生成的深度融合仍需突破两大挑战：一是动态交互场景中实时推理效率的优化，二是对主观审美偏好的量化建模。随着思维链等技术的迭代，未来有望实现“用户语言描述-多方案推理树-3D 场景实时渲染”的端到端生成系统，推动个性化设计普惠化。总之，CoT 通过模拟人类渐进式思考模式，为语言模型赋予了结构化推理能力，其在室内生成等复杂任务中的应用，正不断拓展 AI 创造力的边界。

2.4　网络模型评价指标

2.4.1　场景生成图片评价指标

FID（Frechet Inception Distance[42]）和 KID（Kernel Inception Distance[43]）都是用于评估生成图片质量，尤其是在场景生成等领域衡量生成图像与真实图像相似程度的重要指标。它们能从不同角度量化生成图像的质量和多样性，为评估生成模型的性能提供了客观依据，在生成对抗网络等多种生成模型的评价中广泛应用。

FID 基于两个高斯分布之间的 Wasserstein 距离，主要计算生成图像和真实图像在特征空间中的距离。具体来说，先使用预训练的残差网络提取真实图像集x和生成图像集g的特征，假设这些特征服从高斯分布，分别记为$p(x)$和$p(g)$。首先计算两个分布的均值μ_x、μ_g和协方差θ_x、θ_g。FID 的计算公式为：

$$FID = ||\mu_x - \mu_g||^2 + Tr(\Sigma_x + \Sigma_g - 2(\Sigma_x\Sigma_g)^{\frac{1}{2}}) \cdots\cdots\cdots\cdots (2.2)$$

其中$||\cdot||$表示欧几里得距离，Tr表示矩阵的迹。该公式衡量了两个高斯分布在均值和协方差上的差异，FID 值越低，说明生成图像的分布与真实图像的分布越接近，生成图像的质量越高。直观理解就是，如果生成图像和真实图像在特征空间

中的均值和协方差相似，那么它们的 FID 值就会较低，意味着生成图像在视觉上与真实图像更相似，能够捕捉到真实图像的主要特征和分布规律。

KID 基于核方法和残差特征，通过计算核矩阵之间的差异来衡量分布的相似性。对于真实图像集$x = \{x_1, x_2, \cdots, x_N\}$和生成图像集$g = \{g_1, g_2, \cdots, g_M\}$，首先使用残差网络提取特征，然后计算核矩阵。常用的核函数如高斯核：

$$k(x_i, x_j) = exp(-\frac{||\phi(x_i)-\phi(x_j)||^2}{2\sigma^2}) \cdots\cdots\cdots\cdots\cdots\cdots (2.3)$$

其中$\phi(x)$是残差网络提取的特征，σ是核带宽。计算真实图像的核矩阵K_{xx}，其元素$K_{xx}(i,j) = k(x_i, x_j)$，生成图像的核矩阵$K_{gg}$，其元素$K_{gg}(i,j) = k(g_i, g_j)$；以及交叉核矩阵$K_{xg}$，其元素$K_{xg}(i,j) = k(x_i, g_j)$。KID 的计算公式为：

$$KID = \frac{1}{N^2}\sum_{i,j=1}^{N} K_{xx}(i,j) + \frac{1}{M^2}\sum_{i,j=1}^{M} K_{gg}(i,j) - \frac{2}{NM}\sum_{j=1,i=1}^{M,N} K_{xg}(i,j) \cdot (2.4)$$

KID 值同样越低，表示生成图像与真实图像的分布越相似，生成模型的性能越好。KID 从核矩阵的角度，考虑了特征之间的相似性分布，能更细致地捕捉图像特征分布的差异，对生成图像的质量和多样性评估提供了另一种视角。

FID 和 KID 作为场景生成图片的重要评价指标，从不同方面为评估生成图像的质量提供了有效的量化手段。FID 基于高斯分布的距离，直观地衡量了特征空间中均值和协方差的差异；KID 通过核矩阵的计算，更细致地捕捉了特征之间的相似性分布。在实际应用中，通常结合使用这两个指标，能更全面、准确地评估生成模型在场景生成任务中的性能，推动生成模型不断优化和发展。

2.4.2　场景家具分布评价指标

在室内场景的家具分布分析中，KL 散度（Kullback-Leibler Divergence[44]，KL）作为一种经典的概率分布差异度量工具，能够有效量化实际分布与理论分布之间的偏离程度。KL 散度的定义为：对于两个离散概率分布P和Q，其 KL 散度：

$$D_{KL}(P \parallel Q) = \Sigma_x P(x) log \frac{P(x)}{Q(x)} \cdots\cdots\cdots\cdots\cdots\cdots (2.5)$$

其中P为观测分布，Q为参考分布。在室内布局研究中，P可表示实际场景中家具的空间位置分布，Q则可定义为符合人体工程学[45]或设计规范的理想分布。KL 散度的非对称性特性使其能够区分“真实分布相对于期望分布的偏离”与“期

望分布相对于真实分布的偏离”，这为分析家具布局的合理性提供了方向性依据。例如，当 KL 值显著大于零时，表明当前布局存在明显不符合预期模式的结构性问题。

然而，传统 KL 散度在面向多类别家具分布评估时存在局限性。由于不同类别家具（如座椅、桌子、储物柜）在功能属性和空间约束上存在显著差异，简单的全局分布对比可能掩盖类别间的特异性规律。为此，研究者提出了类别加权 KL 散度（Categorical KL Divergence[46]，CKL），其核心思想是对各家具类别分别构建独立分布模型，再通过加权求和实现整体评估。具体而言，CKL 可定义为：

$$D_{KL}^{weighted}(P \parallel Q) = \sum_{i} w_i \cdot P(i)\log\frac{P(i)}{Q(i)} \cdots\cdots\cdots\cdots\cdots\cdots (2.6)$$

其中$P(i)$和$Q(i)$分别是真实分布和近似分布在家具类别i上的概率。w_i是分配给类别i的权重，需满足$w_i \geq 0$且通常归一化（例如$\Sigma w_i = 1$）以保证可解释性。权重系数可根据家具的功能重要性或空间占比动态调整，例如将床的权重设为高于茶几，以反映其对卧室布局的主导作用。这一改进使得评估过程既保留 KL 散度的分布对比优势，又强化了对类别异质性的建模能力。

从计算实现角度，CKL 散度的应用需解决三个关键问题：首先，需建立合理的空间离散化方法，将连续的三维空间划分为网格单元以统计分布概率。实验表明，采用自适应网格划分（如基于家具尺寸的等比缩放）相比固定网格能提升分布表征的精度。其次，参考分布Q_C的构建需要融合领域知识，例如通过专家规则定义沙发应位于客厅主要墙面附近，或基于大规模数据集学习典型分布模式。最后，权重系数ω_c的确定需结合具体评估目标，若关注功能性则依据使用频率设定权重，若侧重美学评价则可按视觉显著性分配权重。

与 KL 散度相比，CKL 散度的优势体现在两方面：其一，通过解耦不同类别分布，能够精准定位布局异常的具体家具类型。例如，当整体 KL 值偏高时，CKL 分析可进一步识别是餐桌偏离用餐区还是书柜遮挡视线。其二，权重机制的引入增强了评估的灵活性，使算法能够适配不同场景需求。但需注意的是，CKL 的计算复杂度随类别数量线性增长，在面向超多类别场景（如包含上百种装饰品的展厅）时需结合降维技术优化效率。

此外，为了进一步探索家具布局生成效果。本研究也采用了近期工作提出的越界率作为评价指标。在室内场景分析中，越界率（Out-of-Bounds Rate[47]，OBR）

是一项用于评估家具分布合理性的重要指标。越界率指的是室内家具在三维空间中的位置是否超出了预设的空间边界或设计范围，这一指标可以有效地反映家具的摆放是否符合空间布局的规范和使用要求。具体而言，越界率是指家具部分或全部超出设定边界区域的比例，通常用家具超出平面边界的场景百分比来表示。

在实际应用中，越界率的计算不仅考虑到家具本身的物理尺寸，还需结合房间的实际功能和设计目的。例如，在一个卧室中，床头柜若占据了窗户附近的区域，可能会影响到室内的采光和通风，从而影响居住者的舒适度。此外，越界率还可能涉及家具与家具之间的相对位置关系，如沙发与茶几、餐桌与椅子的相互位置。如果家具的摆放导致过多的空间被占用或通道被堵塞，越界率值就会较高，这通常表明空间使用不合理或布局不当。

从数据分析的角度来看，越界率作为一种空间质量评估指标，能够为自动化室内设计、虚拟现实场景生成以及室内布局优化提供量化依据。通过分析越界率的变化，研究人员能够进一步理解室内设计方案的实际适用性，并为后续的设计改进提供数据支持。因此，越界率不仅是家具布局合理性的一个关键评估指标，也是室内空间设计优化和人机交互研究中的一个重要参考值。

总结而言，越界率作为衡量家具分布合理性的指标，能够直观反映室内空间的使用效果，促进室内设计的精细化和智能化发展。在实际应用中，通过优化越界率，可以有效提升空间的利用率与居住舒适度，推动室内设计领域的创新与进步。

2.4.3　场景真实性评价指标

场景真实性评价指标，即场景分类准确率（Scene Classification Accuracy[48]，SCA）是一种基于图像分类模型的客观评估方法，用于量化生成场景与真实场景之间的真实性差异。其核心思想是通过对比生成场景与真实场景在语义层面的可区分性，判断生成结果是否接近真实数据的分布。具体而言，SCA 将生成的场景图像与真实场景图像混合后输入预训练的场景分类模型，通过统计模型对两类图像的分类准确率来评估生成质量。若分类准确率接近 50%（即随机猜测水平），则说明生成场景与真实场景在视觉特征和语义表达上高度相似，分类模型无法有效区分两者，表明生成结果具有较高的真实性。

SCA 的有效性依赖于两个关键因素：一是分类模型的训练质量，二是生成场景与真实场景的数据分布匹配度。首先，分类模型需要在真实场景数据集上充分训练，具备对场景类别（如客厅、厨房、卧室等）的高精度识别能力，以确保其能够捕捉场景的语义特征。其次，生成场景需覆盖真实场景的多样性，包括合理的空间布局、物体间几何关系、材质光照等物理属性。当生成场景与真实场景的差异较大时，分类模型能够轻易区分两者，导致准确率显著高于 50%；反之，若生成场景在细节上逼近真实，分类模型将难以判断图像来源，准确率趋近于理想阈值。

这一指标的优势在于其客观性和可解释性。相较于主观人工评价或低层次特征相似性指标（如 FID），SCA 通过高层语义分类任务直接反映生成场景的合理性。例如，在室内场景生成中，若生成的沙发与墙面比例失调或灯具位置违反物理规律，分类模型可能因场景结构异常而将其判定为生成图像，从而降低 SCA 评分。同时，SCA 的计算效率较高，预训练模型可快速完成批量图像的评估，适用于生成模型的迭代优化。

2.4.4　路径规划算法评价指标

在第二篇工作中，本研究提出了路径规划算法评价指标，即路径规划得分（Path Planning Score，PPS）。受到在大语言模型中使用基于人类反馈的强化学习的启发（Reinforcement Learning from Human Feedback[49]，RLHF），本研究利用 PPS 作为反馈来优化室内场景的生成。下面介绍如何得到 PPS。

假设T表示覆盖路径规划算法执行的整个路径的运行时间，t表示遍历重复路径所消耗的时间，s表示完成的覆盖扫描所覆盖的区域面积，S表示总的可访问区域面积。为了评估扫描机器人在室内场景中覆盖路径规划的时空效率，本研究将 PPS 定义如下：

$$PPS = \alpha \cdot s/S + \beta \cdot (1 - t/T) \cdots\cdots\cdots\cdots\cdots\cdots (2.7)$$

$$\alpha + \beta = 1 \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (2.8)$$

显而易见，PPS 越高说明室内场景对机器人交互的适应性越强。其中α和β为调节时间和空间两大影响因子的比重。

2.5　本章小结

本章首先介绍了 Transformer 的编码器、扩散模型 DDPM 这两种网络结构以及提示词工程 CoT 的背景知识及原理，然后分别详细阐述了场景生成图片质量、场景分布、场景真实性以及关于路径规划算法的评价指标。这些理论基础为后续章节的模型设计、实验评估提供了重要的指导和依据，为深度探讨三维室内场景生成奠定了基础。

第 3 章　基于思维链增强大语言模型的室内场景生成方法

当前的室内场景生成算法在与用户指令对齐和确保逻辑场景一致性方面面临着显著的局限性，这极大制约了具身智能的场景理解与交互能力。为了应对这些挑战，本章节提出了 CoT2Scene，这是一个利用思维链（Chain-of-Thought，CoT）来增强大语言模型进行三维室内场景生成的框架。首先，本章节利用 LLM 处理用户指令，并从知识库中检索最匹配的布局示例。然后，使用检索到的示例，通过框架使用本章节精心设计的 CoT 模板查询 LLM，以流程化的方式一步一步地进行生成意图对齐和实现逻辑结构化的空间布局。随后，从 3D-FUTURE 数据集中检索高质量的 3D 模型，以构建最终的 3D 室内场景。实验结果表明，与之前的方法相比，本章节的方法不仅生成了与用户意图紧密一致的场景，而且在场景质量和逻辑一致性方面优于现有方法。

3.1　引言

室内场景生成需要创建符合逻辑且符合常识的三维空间布局。作为一项基本的计算机图形学任务，它旨在确保符合人体工程学的空间布局来满足以人为中心的要求，以及实现与室内设计原则一致的视觉语义一致性。

这一研究领域已经成为跨学科应用的关键焦点，包括虚拟/增强现实系统、游戏环境的程序内容生成和自动化机器人中的空间推理。大语言模型的快速发展促使其在各个科学领域得到广泛采用，同时也促使人们广泛探索其在改善人类生活体验方面的实际应用。在自动化室内设计领域，现有的场景生成方法在很大程度上依赖于预定义的空间模式和传统的家具布置原则。例如，狭窄的矩形空间经常施加布局约束，阻止标准的家具布置，如梳妆台的放置，导致在任意放置时违背空间边界约束。这种传统的方法从根本上限制了创造性空间利用的潜力，因为有效的室内布局优化旨在实现合理的空间分配和和谐的家具关系管理，这两个要求都是当前基于规则的系统在不同的住宅环境中难以调和的。

本章节中强调利用思维链来引导大语言模型，目的是利用 LLM 强大的生成能力来指导场景生成过程。本章节提出了 CoT2Scene，一个新颖的场景生成框架，提高了合成场景的可控性和合理性，在不需要后处理的情况下获得了吸引人的结

果。本章节的框架的核心概念在于基于 CoT 的提示工程，结合来自知识库的示例驱动指导。这种方式使本章节的方法能够生成逻辑上连贯且与用户指令一致的结果。本章节通过定义家具排列模式的层叠样式表（Cascading Style Sheets，CSS）格式的文本模板实现了从知识库中检索的示例驱动的指导。

为了构建知识库，本章节对 3D-FRONT 数据集进行预处理，以促进高效匹配，使 LLM 能够在推理过程中利用结构化的领域知识。这种集成显著增强了系统处理复杂领域特定查询的能力。最后的场景生成阶段包括从资产库中检索相应的家具模型，并根据 LLM 生成的布局规范将其放置。最终，生成一个完整的室内场景。

3.2　方法介绍

正如第一章所阐述，许多场景设计工作要么从现有的 3D 场景数据库中学习空间知识，要么利用用户输入并迭代改进 3D 场景。最近，大语言模型已被证明对生成 3D 场景布局很有用[47]19[50][51]。然而，他们直接使用 LLM 输出值的方法可能会导致与物理上的合理性相反的布局（例如，家具之间的摆放位置存在重叠）。相比之下，CoT2Scene 使用设计好的提示引导大语言模型逐步生成布局，并使用一系列空间约束来优化布局，以确保物理上合理的场景布置。实验结果表明，与 LLM 端到端生成的布局相比，CoT2Scene 生成的布局更合理。

3D 生成的早期工作侧重于从特定类别的数据集中学习 3D 形状或纹理的分布。随后，像 CLIP[52]这样的大型视觉语言模型的出现，使得 3D 纹理和物体的零样本生成成为可能。这些工作在生成 3D 物体方面表现出色，但在生成复杂的 3D 场景方面却受到一定限制。最近的工作如 NeRF[53]通过将预训练的文本到图像模型与深度预测算法相结合来生成纹理网格，从而生成 3D 场景。然而，这些方法受到模型理解文本能力的限制，并且缺乏模块化的可组合性和交互功能。相比之下，CoT2Scene 则充分利用了 LLM 的文本理解能力和转换能力，将文本表示出来，并匹配相应的 3D 模型来填充场景。本章节介绍的工作可用于未来具身智能的研究。

下面针对本章节提出的方法进行细致的阐述，具体内容包括问题陈述、面向大语言模型的外部知识库构建以及基于思维链增强大语言模型的场景生成流程

的细节陈述。

图 3.1 展示了本章节框架的实现流程。首先，本章节利用 LLM 来解释用户指令并提取其中蕴含的家具之间的核心空间关系。随后，本章节从外部知识库中检索与用户意图一致的最相关的布局示例，并将其合并到 LLM 的请求提示词模板的上下文中作为指导。最后，本章节采用精心设计的基于 CoT 的模板来指导 LLM 一步一步地生成合理的场景布局。

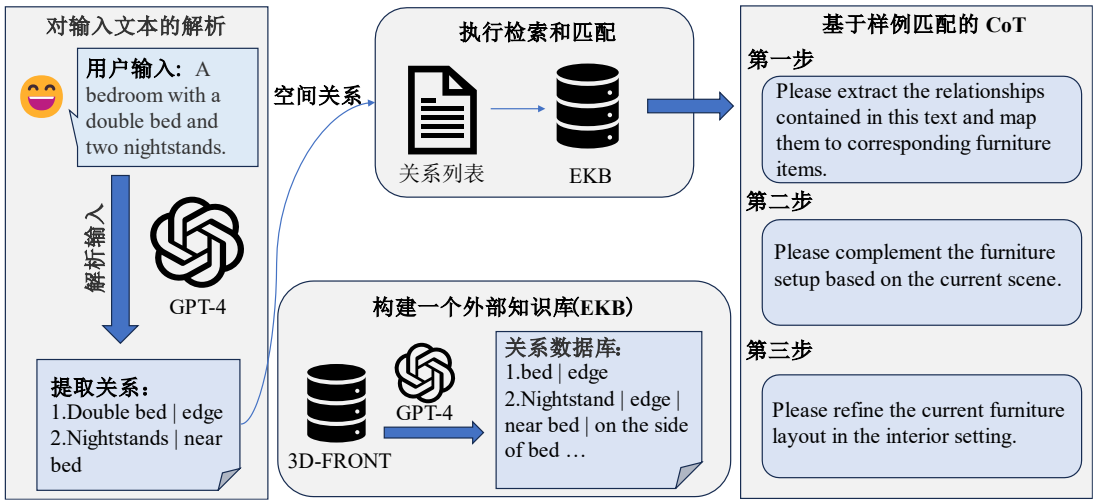

图 3.1　CoT2Scene 框架总览

3.2.1　问题陈述

本章节将场景生成问题定义为预测家具的属性，其中室内场景是由各种家具物品和房间平面图组成的。然后，本章节将家具的属性表征为不同维度向量的组合。这些属性包括类别（用 1D 向量表示）、尺寸（3D 向量）、方向（3D 向量）和位置（3D 向量）。房间的平面图表示为一个$N * N$的向量矩阵（由 0 和 1 组成，0 代表墙壁或空白空间，1 代表家具占用的空间）。

之后的小节，首先通过构建一个基于 3D-FRONT 的外部知识库，本章节为大语言模型建模提供了丰富的专家知识。接着本章节提出利用 CoT 这一提示词工程技术增强 LLM，最终生成满足用户输入的文本描述的三维室内场景。

3.2.2　面向大语言模型的外部知识库构建

在 3D-FRONT 数据集的基础上，本章节通过对原始数据集的处理，启动工作流，以构建结构化的外部知识库。在接收到用户指令后，本章节利用大语言模

型从用户的输入文本中识别和检索核心空间关系。随后，本章节利用这些关系来查询外部知识库中最相似的布局，将其作为后续流程的参考布局示例。在知识库构建阶段，本章节利用 LLM 提取了家具物品的属性信息和从 3D-FRONT 数据集中得到的关于每个房间内的空间关系。随后，家具属性及其相应的空间关系被系统地存储为知识库中的键值对（如图 3.2 所示）。在布局生成阶段，本章节首先利用 LLM 来理解用户提供的文本输入。接着，本章节使用精心设计的提示模板来提取嵌入在用户输入文本中的核心关系。然后，本章节使用该模板来查询 LLM，通过遵循基于对象及其关系的结构化序列来迭代地识别最关键的关系。最终，本章节在知识库中执行布局匹配，以搜索显示最相似的关系布局，这些布局随后将被合并到下一阶段的提示模板中，作为生成目标布局的上下文指导。

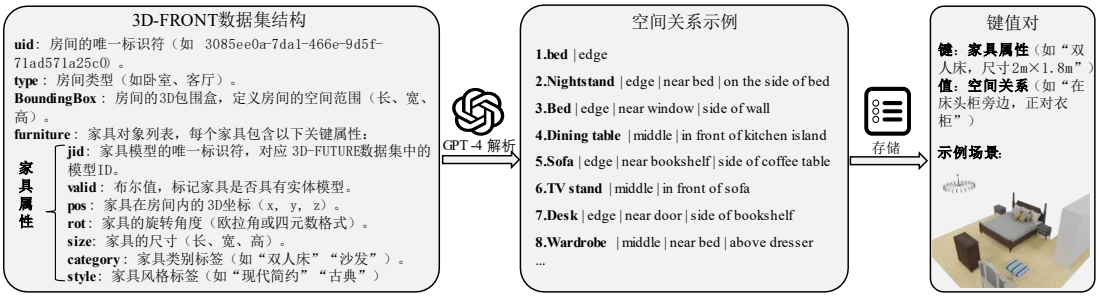

图 3.2　外部知识库构建流程示意图

3.2.3　基于思维链增强大语言模型的室内场景生成流程

在大语言模型领域，思维链是一种新兴的提示词技术，它通过模拟解决问题的思维过程来提高模型生成内容的质量和相关性。CoT 的概念来源于人类在解决问题时的思维方式，即通过一系列逻辑步骤逐步推导和解决问题。为了生成符合室内设计原则的室内场景，本节利用 CoT 推理来指导 LLM 通过结构化的生成过程，从而确保合成场景的物理合理性。

下面，本章节阐述提示词模板的结构。具体包括五个组成部分：第一个组件是系统提示，它将模型角色定义为室内设计师，并要求输出采用层叠样式表格式以便于无缝的程序化解析。第二个组件为用户输入文本与先前提取的核心关系的整合。第三个组件存储了之前从知识库中检索的类似布局示例。第四个组件明确了约束场景生成的专家规则。最后，第五个组件为本章节核心部分之一，即嵌入

的精心设计的 CoT 模板，引导 LLM 逐步生成具有连贯性和逻辑性的布局方案。

为了确保生成的场景符合人体工程学[45]原则，本章节将专家规则合并到 CoT 中。本节根据之前的工作[51]由五大类别构成的十种空间约束条件，分为三组：

（1）全局定位：边缘布局、居中布局;

（2）相对距离：邻近配置、疏远配置;

（3）方位关系：前置方位、侧翼方位、顶部悬置、表面承载;

（4）对齐方式：中轴对齐;

（5）朝向规范：面向关系。

LLM 为每个对象选择这些约束的一个子集，这些约束条件作为指导性建议，当无法完全满足所有约束条件时，允许存在合理范围内的偏差。除上述弹性约束外，本节额外设立了两个刚性约束准则：严格禁止物体间的空间交叠，以及确保所有物体均位于房间边界范围内。

下面介绍约束条件的处理细节。本章节首先将上述空间关系约束表示为数学形式的条件（例如，若两个对象的欧几里得距离大于家具间距平均值 0.7 米，则视为疏远配置）。为生成符合 LLM 采样约束的布局方案，本研究采用自回归优化算法进行家具对象定位。该算法首先通过 LLM 确定锚定对象，继而系统性地探索其可行的布局位置。随后采用深度优先搜索（Depth First Search[54]，DFS）对剩余对象进行有效排布验证。其中，所有刚性约束的完全满足是布局有效的必要条件。例如，床被选定为卧室空间的锚定对象后，床头柜的布局随即展开。算法以固定时间窗口（30 秒）执行多轮迭代，生成候选布局集合，最终选取满足最多约束条件的优化解。本研究在 3.3 节通过实验验证了该约束驱动布局方法的有效性。

综上所述，本研究通过模块化思维链框架将室内布局生成流程拆解为三个递进阶段，在确保逻辑严谨性与方法可复现性的基础上，实现了智能化布局设计的系统性优化。具体实施路径如下：

（1）核心对象生成阶段

基于专家规则与布局参考案例构建查询请求，引导 LLM 精准提取符合预设空间关系的核心功能单元。此阶段通过结构化提示工程，将拓扑约束、人体工学参数等专业要求融入查询指令，使模型在生成过程中既能继承领域知识，又能保持设计创新性。

（2）场景填充与适配阶段

整合预训练模型的多模态先验知识与典型布局样本，运用动态上下文注入技术完成场景细部设计。具体而言，通过示例引导模型理解家具比例、布局规划等隐性规则，同时采用语义强化策略对装饰元素、材质搭配等属性进行参数化描述，实现功能需求与美学表达的深度耦合。

（3）布局优化与验证阶段

进行空间连贯性评估，从拓扑合理性（如通道宽度阈值）、视觉平衡度（如重心分布）及风格一致性三个维度进行迭代调优。最终生成既符合物理约束又具备美学统一性的布局方案。

3.3　实验结果与分析

3.3.1　场景生成任务的实验分析

在实验的初始阶段，本章节实现了一个全面的数据预处理流水线。这包括三个关键步骤：（1）从过滤后的 3D-FRONT 数据集中解析空间配置；（2）用相应的核心关系标注每个房间；（3）为每个处理过的房间生成结构化的层叠样式表。由此产生的数据结构将每个房间组织成一个包含三个核心组件的字典：精炼的家具属性、唯一的房间标识符和用于快速匹配的核心关系。

这里简要介绍采用的数据集 3D-FRONT，该场景数据集包含数万个完备标注的室内场景，其中包括结构化的建筑元素，如房间布局、3D 家具模型和材料参数（包括纹理贴图）。在应用与之前研究一致的预处理流程后，本研究获得了 4273 个合格的卧室场景、273 个餐厅场景和 841 个客厅场景。

为了建立严格的基准框架，本节对基于文本的室内场景生成研究中两种最先进的方法进行了比较评估，即 LayoutGPT[47]和 ATISS。ATISS 是基于 Transformer 构建的自回归室内场景生成模型架构。LayoutGPT 通过解释自然语言输入来生成结构化的视觉场景，从而实现跨模态场景合成。LayoutGPT 说明了在 2D 和 3D 领域的应用，包括 2D 图形渲染和 3D 室内场景合成。

在本章节实验的实现中，本研究选择的 LLM 是 GPT-4[55]，因为 GPT-4 在许多大语言模型中表现良好，并且 API 成本低。此外，关于评估指标，本研究遵循

之前的工作，从四个相机角度为每个场景渲染场景图像，采用弗雷切特起始距离即 FID 进行场景生成图片评估，并报告了生成场景和真实场景的家具类别分布之间的 KL 散度即 CKL。此外，本节还评估了不同方法生成场景的越界率，这代表了家具超出平面图边界的场景的百分比。

场景生成的定量实验结果见表 3.1。在这个实验中，本研究生成了三种具有代表性的房间类型（具体来说是卧室、客厅以及餐厅），并通过三个量化指标系统地评估了生成的场景。

表 3.1　CoT2Scene 与 LayoutGPT 和 ATISS 关于室内场景合成的对比结果

房间类型	方法	越界率	CKL↓	FID↓
卧室	ATISS	53.11	0.0224	32.49
	LayoutGPT	42.43	0.0932	29.53
	CoT2Scene	**23.28**	**0.0168**	**29.07**
客厅	ATISS	86.65	0.1079	85.40
	LayoutGPT	72.66	0.1628	77.60
	CoT2Scene	**40.07**	**0.0779**	**74.16**
餐厅	ATISS	83.26	0.2630	90.33
	LayoutGPT	63.79	0.1605	80.74
	CoT2Scene	**45.62**	**0.1527**	**79.33**

表 3.1 中的对比分析显示，本研究的 CoT 增强方法在越界率度量上显著优于现有方法，验证了思维链增强基于 LLM 的场景生成中的空间推理的假设。此外，结果表明，就测量场景多样性和合理性的两个关键评估指标（FID 和 KL 散度）而言，本章节所提出的框架也超越了当前最先进的方法（LayoutGPT 和 ATISS）。

本方法生成的场景被渲染以产生相应的室内可视化，如图 3.3 所示，本章节的方法不仅考虑了常见的易于理解的 3D 概念，例如“床头灯应该挂在天花板上”和“床头柜应该放在床边”，而且在确保风格一致性的同时，还生成了更多样化的场景布局。考虑到客厅和餐厅的平面尺寸，本章节的方法还会产生更复杂的 3D 布局，一边是餐桌和椅子，另一边是沙发、咖啡桌和电视柜（右下）。在本章节的实验框架中，选择了三种主要的房间类型（卧室、餐厅和客厅）。对于每种房

间类型，指定空间关系和家具配置的描述性文本输入被系统地馈送到生成模型中。

从生成结果可以看出，在同一用户文本输入指令下，本章节的方法和另外两种方法（ATISS 和 LayoutGPT）都生成了基本符合的室内布局。然而，在经过进一步观察后可发现，ATISS 的卧室和餐厅输出中存在明显的家具碰撞问题，而 LayoutGPT 的客厅橱柜显示出违反边界约束的情况。与这些缺点相比，本章节的方法成功地消除了家具碰撞和边界越界。这是由于本方法在提示模板中指定了更为严格的约束条件。总之，本章节的方法生成的室内场景布置在空间组织上更合理，也更符合用户指令。

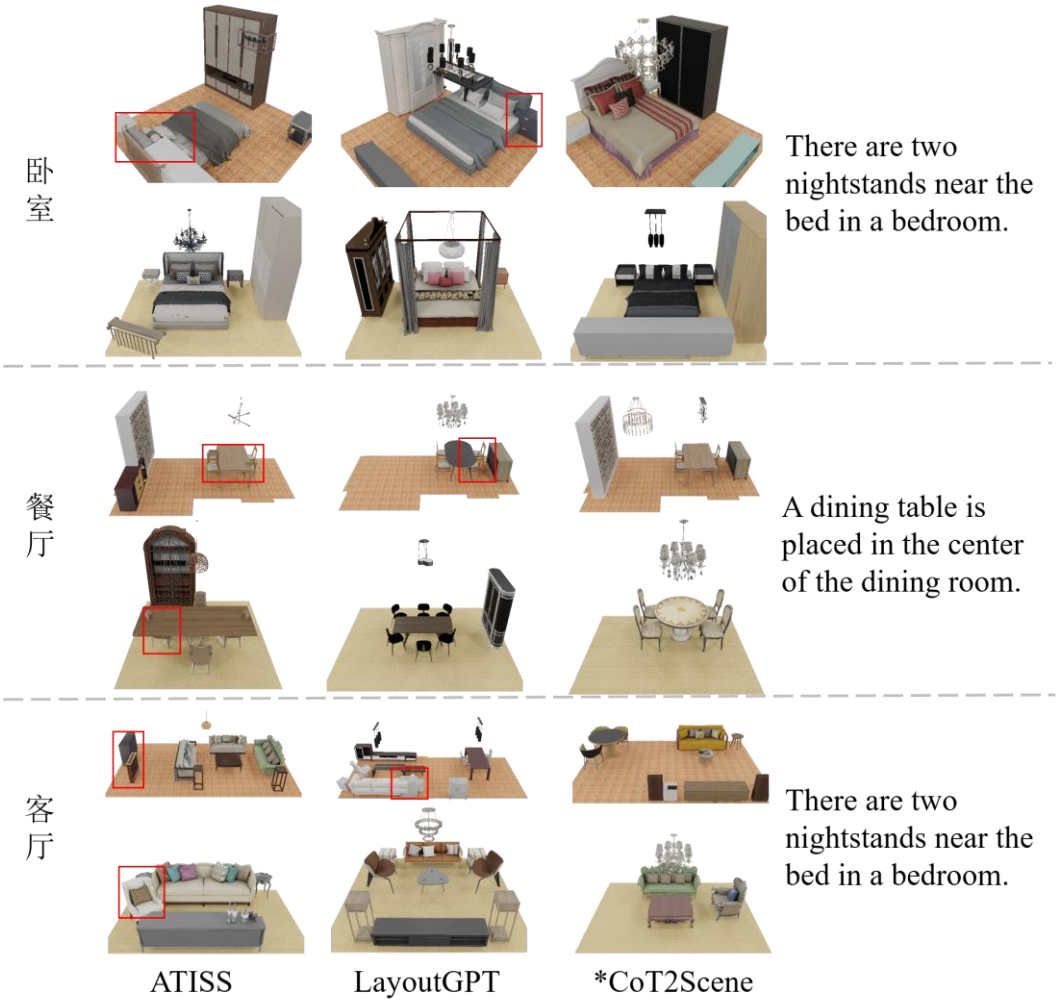

图 3.3　不同类型房间和不同平面图的可视化对比结果

3.3.2　消融实验

针对本章节提出的方法，本小节进行了严谨且详细的消融实验研究。本节首

先对比加入思维链技术前后，在卧室和客厅数据集上的评估指标的数值变化，以验证本章节提出方法的有效性，如表 3.2 所示。第一行的数据为去掉 CoT 这种利用分步提示的方式，改为使用通用的提示模版进行一次性生成的评估结果，第二行则为加入 CoT 后的实验结果。通过观察表格中的数值变化可知，在去掉 CoT 后，LLM 生成场景的能力确实有所下降。具体而言，在越界率方面，加入 CoT 后家具的越界率得到了显著下降，下降了接近 3%。在 CKL 和 FID 方面也有明显的下降趋势，这说明了本方法对提升室内场景布局的合理性和控制家具越界方面有着明显效果。

表 3.2　关于 CoT 的消融实验结果

房间类型	方法	越界率	CKL ↓	FID ↓
卧室	去掉 CoT	26.77	0.1623	30.06
	CoT2Scene	**23.28**	**0.0168**	**29.07**
客厅	去掉 CoT	42.84	0.1006	76.85
	CoT2Scene	**40.07**	**0.0779**	**74.16**

本章节基于空间布局的物理合理性与美学需求，系统性构建了包含全局定位、几何形态、碰撞规避、功能适配及人体工程学等五大类别的十种刚性约束条件。为深入探究不同约束层级对生成效果的影响机制，本章节设计了层次化对比实验：首层以完全自由的无约束空集为基准，第二层级引入对空间结构起基础支撑作用的全局定位约束，并叠加确保物体间交互合理性的相对距离与方位关系约束，第三层级则完整集成全部五类约束条件。消融实验结果显示（表 3.3），初始无约束状态下生成场景存在显著的物体穿透、功能区域错位等缺陷；当融入基础空间关系约束后，场景的拓扑合理性有所提升；而完整约束体系的引入不仅使评估指标全面优化，更在隐性维度上增强了场景的语义连贯性，充分印证了多维度刚性约束在抑制生成偏差、保障空间逻辑自洽方面的核心作用。这种渐进式的约束增强策略，为三维场景生成领域提供了可解释性强的技术路径。

表 3.3　关于约束条件的消融实验结果

约束条件	卧室		
	越界率	CKL↓	FID↓
不使用刚性约束	59.74	0.2441	70.85
使用三种约束	23.36	0.0174	29.16

续表 3.3

约束条件	卧室		
	越界率	CKL↓	FID↓
使用五种约束	**23.28**	**0.0168**	**29.07**

3.3.3　局限性分析

首先，尽管本章节提出的方法在确保生成逻辑一致，符合用户输入的场景上取得了不错的进展，但考虑到目前大语言模型的飞速发展，本研究使用的 GPT-4 并非最优的模型选择。如异军突起的国内自主研发的大语言模型 DeepSeek[56]或许在文本理解和处理上表现由于当前使用的 GPT-4。其次，本章节采用的实验流程涉及到 LLM 对外部知识库的处理，目前是在规模有限的关系库采用深度优先遍历的方式完成，这部分工作可以进一步完善，扩展知识库的规模，并引导 LLM 自动化的完成上述流程，最终实现真正端到端的生成。最后，本研究采用数据 3D-FRONT 尽管建筑元素丰富，但不同房间类型的数目差异显著，除卧室类型外，其他类型的房间样本数量较少，这制约了本研究的方法在其他类型房间上生成的布局质量。

3.4　本章小结

在探索和实验验证了本章节中的方法后，本章节发现基于大语言模型的引导在室内场景生成任务中具有巨大的潜力。本文提出的 CoT2Scene 方法利用 LLM 强大的表示学习和生成能力，有效克服了传统方法在与用户指令对齐和空间一致性方面的局限性。通过思维链的逐步引导，加上强约束，本研究得到了更符合室内设计规律的家具布局方案。3D-FRONT 数据集上的其他实验进一步证明了该方法的通用性和鲁棒性。实验结果表明，CoT2Scene 不仅能够生成视觉逼真的 3D 场景，而且能够更好地捕捉实际细节，从而避免了传统方法在布局设计中常见的错误。

总体而言，本文提出的工作为 LLM 引导的室内场景生成提供了新的视角和方法。它不仅提高了生成效率和场景多样性，而且为未来在虚拟现实和智能家居中的实际应用提供了新的途径。同时，除了将文本作为条件输入之外，在未来将

基于 CoT 的 LLM 扩展为多模态输入（如图像、草图），以实现更灵活的场景生成将是一件值得探索的事情。

第 4 章　基于机器人路径规划约束的室内场景生成优化

将具身智能集成到室内场景合成中，对于未来室内设计应用而言具有巨大潜力。然而，室内场景合成的主流方法[57][58][59]主要坚持数据驱动的学习范式。尽管通过这些方法实现了逼真的 3D 渲染，但目前的场景生成框架忽略了以机器人为中心的功能指标，这些指标对于优化服务类型机器人平台或智慧家庭助手等嵌入式人工智能系统中的导航拓扑和面向任务的交互性至关重要。例如，摆放不当的家具可能会阻碍机器人与环境的有效交互，而这个问题不能仅仅通过引入先验约束来完全解决。为了解决这一难点，本章节提出了 Nav2Scene，这是一种新型的即插即用微调机制，可以部署在现有的场景生成器上进行场景的优化，以增强生成的场景对高效机器人导航的适用性。

4.1　引言

室内场景生成指在三维空间内布置符合人体工程学的家具摆放任务[60]。从传统角度来看，场景生成的目标是满足人类的审美和功能需求，即创造在视觉上合理和有吸引力的房间，以及在物理上可供人们导航的房间。由于在虚拟现实，增强现实，开放世界游戏和机器人等领域的广泛应用，该任务引起了人们的广泛关注。

随着具身智能的发展，机器人越来越多地被集成到家庭和医疗保健环境中。这一趋势要求房间布局的设计不仅适合人类，而且对机器人友好，专门针对机器人导航和操作进行优化。然而，本文认为现有的场景生成方法存在普遍的问题。例如，虽然狭窄的家具间距可能在视觉上看起来很吸引人，但它可能会严重阻碍机器人的运动。目前的场景生成方法主要强调人的视角，往往忽略了高效可行的机器人导航的功能要求，比如路径规划和避障[61]。这一限制对推进家用机器人的应用提出了挑战，在家用机器人中，机器人与环境之间的有效导航和交互至关重要。

为了解决这些挑战，本章节提出了 Nav2Scene，如图 4.1 所示。这是一种新型的即插即用微调机制，可以部署在现有的场景生成器（如 ATISS[48]和 DiffuScene[62]）上，以增强生成的场景对高效机器人导航的适用性。本文首先介绍

了路径规划得分（Path Planning Score，PPS），这是一个从机器人的角度评估室内场景可导航性的指标。给定一个场景，基于常规路径规划算法评估的结果计算 PPS。随后，为了加快路径规划评分的计算速度并保证基于 PPS 的优化过程的可微分性，本章节引入了一种轻量级神经网络 ScoreNet。ScoreNet 使用预先计算的 PPS 值作为预测标签，对现有场景数据集进行训练，旨在有效地计算新场景的 PPS，从而显着降低计算开销。最后，本章节将 ScoreNet 集成到现有的场景生成框架中，使用预训练的网络对生成的场景的 PPS 进行实时评估。为了进一步增强机器人导航生成的场景，本章节通过引入组合损失函数对场景生成网络进行微调。该损失函数综合了原始场景生成损失和机器人路径规划损失，后者通过计算 PPS 得到。

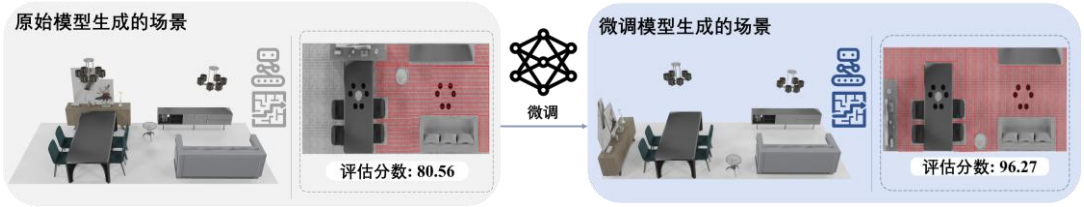

图 4.1　Nav2Scene 框架总览

大量的实验证明，经过微调的模型不仅能够生成更适合机器人导航的场景，而且还能遵循场景生成的原始约束，维持功能性和视觉真实感之间的平衡。

4.2　方法介绍

本节首先介绍本章节的 Nav2Scene 框架，其架构如图 4.2 所示，本章节介绍了 Nav2Scene，一个路径规划微调模块，旨在增强现有的 3D 室内场景生成模型。该框架分两个阶段运行。在第一阶段，本章节使用预先计算的路径规划分数在室内场景数据集上训练 ScoreNet。这使得 ScoreNet 能够快速预测新场景的路径规划分数，并确保后续微调过程的可区分性。在第二阶段，本章节将训练好的 ScoreNet 整合到现有的室内场景生成模型（例如 ATISS[48]和 DiffuScene[62]）中，对其生成的结果进行评估并提供反馈。该方法对生成的场景布局进行了细化，使其更加合理，并针对改进的路径规划和空间功能进行了优化。

下面本节分别从问题陈述、路径规划算法、导航驱动的得分网络以及损失函数四个方面对本章节的核心方法进行阐述。

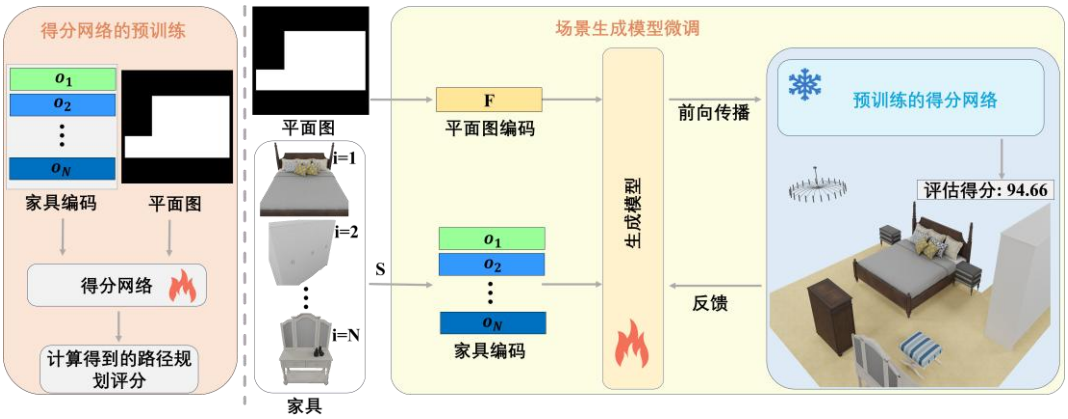

图 4.2　Nav2Scene 框架结构示意图

4.2.1　问题陈述

本章节定义$\mathcal{S}:=\{\mathcal{S}_1,\ldots,\mathcal{S}_M\}$作为室内场景的集合。每个场景$\mathcal{S}_i:=(\mathcal{O}_i,F^i)$由一个无序的对象集合$\mathcal{O}_i:=\{o_j^i\}_{j=1}^N$及其对应的平面图$F^i\in\mathbb{R}^d$组成，其中$d$为特征向量的维数。每个对象$o_i$包括类别$c_j^i\in\{1,\ldots,k_c\}$，其中$k_c$为$\mathcal{S}$中对象类的个数，位置$t_j^i\in\mathbb{R}^3$，轴向尺寸$s_j^i\in\mathbb{R}^3$，方向$r_j^i\in\mathbb{R}$。

为了生成室内场景，本章节根据c和F从高质量的 3D-FUTURE 数据集中检索 3D 物体，并分别根据s、t和r的值依次放置它们。在此之前，本章节需要一个生成模型来预测这些家具属性。该模型将指定类型的空房间或部分装修的房间(例如卧室)及其平面图作为输入。通过训练，模型学会用家具物品填充这些空间，确保生成布局中满足功能一致性和达到空间优化的目的。

4.2.2　路径规划算法

在 Choset 和 Galceran 进行的调查[63][64]中，已经对覆盖路径规划中的大部分相关工作进行了深入的总结和讨论。其中，有四种主要的房间覆盖路径规划方法:经典的精确细胞分解技术[65]，基于莫尔斯的细胞分解方法[66]和基于 LANDMARK 的细胞分解算法[67]，以及广泛使用的 Boustrophedon 细胞分解方法[68][69]，该方法也称牛耕式细胞分解方法，属于在生成的细胞内进行简单的来回运动的细胞分解方法。在本文中，本章节采用了 Boustrophedon 覆盖路径规划算法，这是国内外扫地机器人中使用最广泛的有效区域覆盖方法之一。该方法通过系统分解实现高效的环境覆盖，为 Nav2Scene 提供有效的布局优化反馈。

如图 4.3 所示，该图展示了本章节的路径规划实现原理，左子图为机器人的覆盖范围计算，右子图为路径规划算法的大致工作机理。在运行算法之前，需要对房间朝向进行归一化处理。本章节的牛耕式方法应用于覆盖在房间区域上的格式塔网格或细胞分解。网格方向的选择可以显著影响覆盖路径规划算法的性能，特别是在单元分解的数量、平行轨迹的数量和点转弯的密度方面。这里，本章节将采用一种常用的启发式方法，即将网格与物体主方向对齐。具体来说，本章节将房间布局平面图（带有家具）通过二值化处理转换为图像，其中白色区域表示房间的可访问区域，深色区域表示墙壁或障碍物，然后使用 OpenCV[70]中的 Hough 线检测器提取房间轮廓。

本章节假设房间方向归一化使房间的最长维度沿 x 轴对齐。这样，当扫描线到达一个临界点时，水平扫描线就会从房间地图的顶部移动到底部，即当一个空间段被障碍物平分时，如图 4.3 所示。带有投影传感器的最大网格尺寸(左图)。当两个独立的片段在障碍物后面合并时，空间被相应地分割和合并，从而形成一组可以用简单的来回运动模式覆盖的单元格。然后分析每个单元格的主要方向，以建立一个来回运动模式，使每次通过的长度最大化，并使转弯的持续时间最小化。最终，本章节实现了对整个可达区域的覆盖扫描。

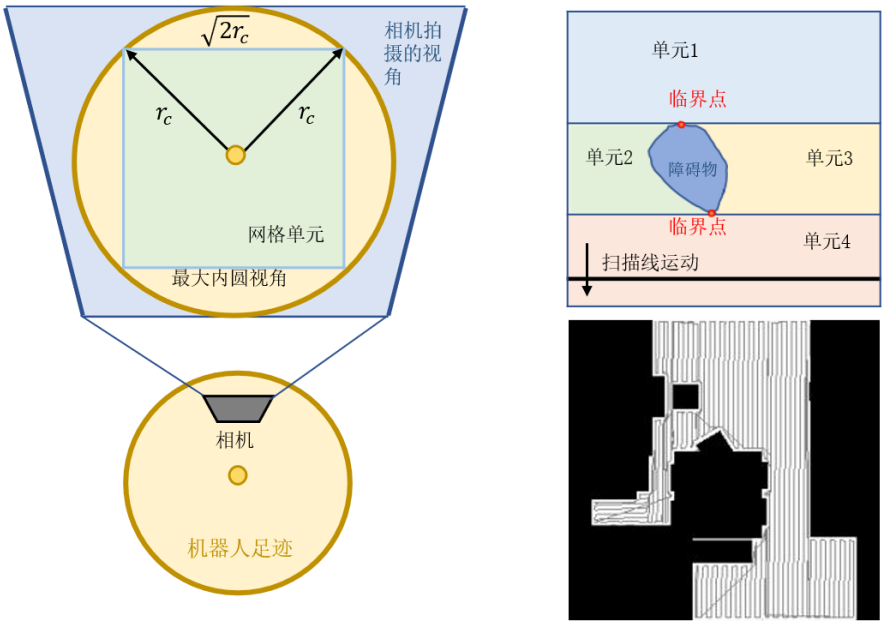

图 4.3　路径规划过程。使用投影传感器的最大网格尺寸（左），Boustrophedon 细胞分解（右上）和细胞内简单运动模式的生成（右下）

4.2.3　机器人路径规划评估数据集构建

考虑到现有的室内场景数据集大多来自艺术家们的室内设计或利用仪器扫描得到，缺少关于机器人路径规划的相关标注。为此，本节介绍了针对室内场景的机器人路径规划评估数据集的来源以及构建过程。

3D-FRONT 自 2020 年发布以来已成为计算机图形学与室内设计领域的重要基准。该数据集通过专业设计师构建了超过 10,000 个高精度三维场景，涵盖现代简约、古典奢华等多种风格，其核心价值在于提供带有材质纹理的家具模型（如沙发、餐桌等）及完整的空间语义标注（如墙面、门窗结构）。在现有研究中，该数据集主要服务于虚拟现实渲染、室内布局生成等任务，其结构化数据特征（如场景边界框、物体空间坐标）为算法提供了丰富的几何与语义信息。

尽管 3D-FRONT 在视觉保真度与场景多样性上具有优势，但其原始设计目标与机器人导航需求存在显著偏差。具体表现为：

1.布局合理性偏差：设计师主导的场景构造倾向于美学表达，导致部分样本出现装饰物阻塞通道、可行走区域碎片化等问题（如图 1.2 所示，左侧的沙发挡住了后面的过道），这与真实环境中机器人运动的空间连续性需求相悖。

2.标注体系缺失：数据集未包含导航算法评估所需的关键标注，包括可行区域拓扑图、动态障碍物运动轨迹、多目标路径规划起点终点对等。例如，在厨房场景中，操作台与橱柜之间的通行宽度标注缺失，使得算法无法验证狭窄通道的避障能力。

3.物理属性不足：物体仅标注几何尺寸与材质类型，缺乏摩擦系数、重量等影响机器人运动规划的物理参数，导致仿真环境与真实物理世界存在建模误差。未来具身智能的研究需要同室内场景的物体进行交互，这些物理属性的缺失限制了将该数据集应用到更广泛研究的可能。

为提升 3D-FRONT 在机器人导航领域的适用性，本研究对数据集进行重构。重构后的 3D-FRONT 数据集拥有与原始版本场景相同的规模，较原始版本新增每个房间的正交二值化投影、对应的路径规划结果图（如图 4.4 所示），以及相应的参数数据标注（如表 4.1 所示）。该数据集填补了设计导向场景与机器人导航需求之间的鸿沟，为 SLAM 算法验证、多机器人协同等研究提供了高保真测试平台。未来可通过引入 LiDAR 点云融合、人类活动轨迹模拟等扩展数据维度，

进一步提升其在复杂动态场景中的实用性。

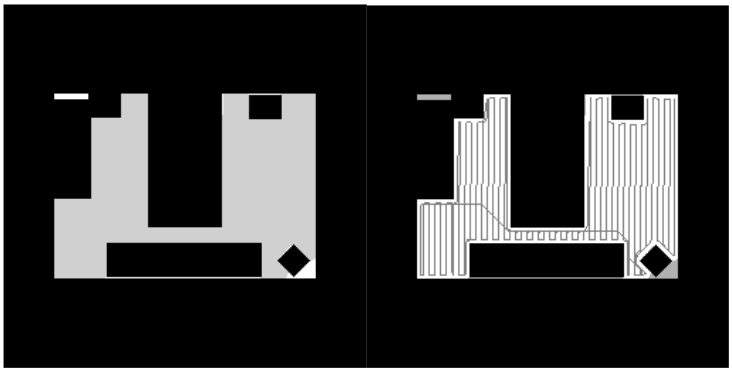

图 4.4　3D-FRONT 某房间的正交投影（左）及对应的路径规划结果图（右）

下面本研究介绍如何将机器人路径规划集成到 3D-FRONT 数据集中。经过调研发现，需要先将 3D-FRONT 数据集中每间屋子的房间进行渲染，得到其正交投影图，再根据投影得到的图片进行二值化处理。这里本研究为便于二值化处理，将渲染用地板颜色置为黑色。完成上述工作后，接下来进行技术细节的阐述。

表 4.1　3D-FRONT 某房间的路径规划参数标注示例

计算时间$[s]$	地图路径长度$[m]$	旋转值$[rad]$
2.53	95.162	206.834
覆盖面积[%]	房间面积$[m^2]$	轨迹角度得分平均值
0.978114	34.79	0.94455
墙壁角度得分平均值	重访时间平均值$[s]$	交叉次数
0.977469	0.0573624	144
旋转次数		
50		

首先是环境建模与代价地图构建，本研究将黑白二值图通过 ROS 的 Map Server 包转换为 Occupancy Grid 格式，其中灰度阈值设定为：像素值<25%判定为自由空间（白色），>65%为障碍物（黑色），中间区间为未知区域。通过激光扫描仿真器（如 Stage 或 Gazebo）注入虚拟激光雷达数据，建立基于贝叶斯更新的代价地图（Cost Map[71]）。特别需处理设计师布局中的非结构化障碍物，如装饰性镂空隔断，需通过形态学闭运算填补细小孔洞，避免路径穿越无效区域。

接着，进行多层路径规划算法集成。本研究在 ROS 导航栈中配置分层规划器，在全局规划层，本研究采用 A*算法[72]，启发函数权重设置为欧式距离的 1.5

倍以平衡最优性与计算效率。针对复杂户型（如环形走廊），引入拓扑地图自动生成模块，通过 Voronoi 图[73]提取骨架路径，使用优化的最短路径算法 Dijkstra 算法将规划时间从$O(n^2)$降至$O(nlogn)$。在局部规划层，本研究融合动态窗口法（Dynamic Window Approach[74]，DWA）与时序弹性带（Timed Elastic Band[75]，TEB）算法。对 DWA 的速度采样空间进行约束，TEB 层通过时空优化处理动态障碍物，响应延迟控制在 150 ms内。

本研究利用 ROS 的通信机制模拟扫地机器人的导航过程，采用牛耕法[76]进行空间扫描，同时触发避障机制，采用优化算法降低碰撞次数。运动中最小化转动次数和转弯角度，实现机器人覆盖式导航算法的最优解。

最终，进行系统优化与评估指标。本研究针对路径质量进行评估，具体而言，即建立多维度评价体系，包括路径长度比（实际/理论最优）、最大曲率（$< 0.35\ m^{-1}$）、加速度方差（$< 0.1\ m^2/s^3$）等。采用蒙特卡洛法在 1000 个随机起点对中测试，统计成功率与异常终止原因（如局部极小值陷阱）。关于路径质量评估的最终指标，本研究会在下一章进行详细阐述。

4.2.4　导航驱动的得分网络

受使用来自人类反馈的强化学习来对齐 InstructGPT[77]中的人类偏好的启发，本章节利用路径规划指标作为反馈来优化室内场景生成。为了实现这一点，本章节使用机器人路径规划评估数据集来训练一个轻量级的 MLP 网络，称为 ScoreNet。ScoreNet 作为奖励网络，而场景生成模型作为智能代理。然后，本章节利用 ScoreNet 的输出来微调预训练的生成模型，生成满足预期的房间内家具布局。本章节根据路径规划算法计算路径规划得分（Path Planning Score，PPS）。具体而言，首先，记录机器人全部工作路线的运行时间T。然后，将机器人在每个房间的工作轨迹重复时间t除以运行时间T。接下来，本章节计算机器人的轨迹覆盖率，即轨迹覆盖面积 s 除以房间的可清洁面积S（家具占据部分为不可清洁面积）。总而言之，通过结合时间和空间维度，本章节对所得路径对清洁专业人员的主观吸引力进行评估，计算公式在第二章节有所提及，具体如下所示：

$$PPS = \alpha \cdot \frac{s}{S} + \beta \cdot \left(1 - \frac{t}{T}\right) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (4.1)$$

$$\alpha + \beta = 1 \cdots\cdots (4.2)$$

这里α和β值是可调节的计算参数。在后续的消融实验中，本章节将更详细地探讨不同数值对α和β的影响。

由于本章节不能直接使用路径规划分数作为真实值来训练室内场景数据，这源于获得地板布局和生成分数所涉及的过程具有不可微分性质。为此，本章节设计了一个独特的网络模型，即 ScoreNet，以改进构建过程。为了支持这一点，本章节通过使用路径规划分数，通过相应的分数标注原始 3D-FRONT 数据集中的 4041 个房间，从而创建一个全新的数据集，即机器人路径规划评估数据集。这个数据集是训练 ScoreNet 的基础。如图 4.5 所示，本章节方法中的微调模块架构（ScoreNet）以多个对象的属性（3D 包围盒、对象类、平面图编码）和分数作为输入，并使用可训练的路径规划算法模块对其进行评估。具体而言，ScoreNet 基于一个简单的、全连接的神经网络，该神经网络由两个隐藏层和一个输出层组成，其中 ReLU 函数用作隐藏层中的激活函数。

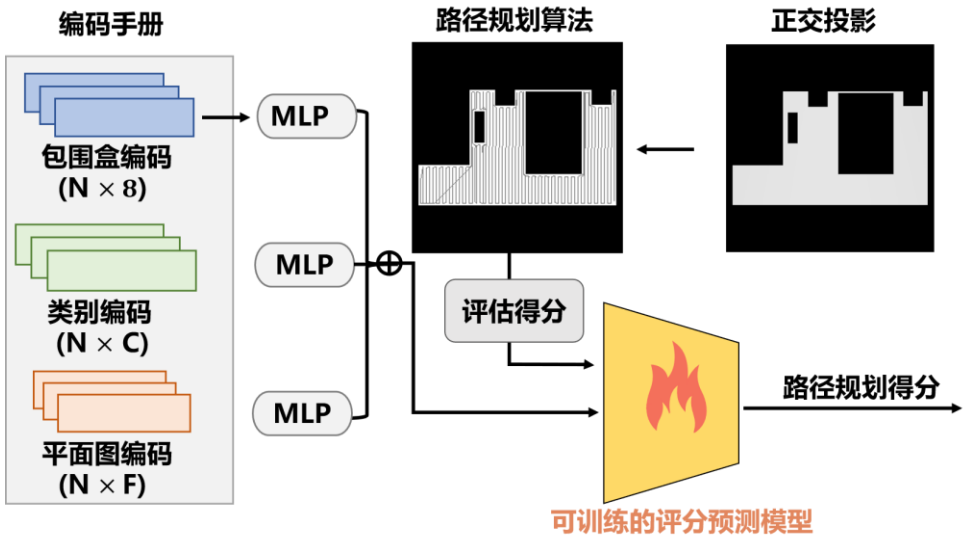

图 4.5　微调模块（ScoreNet）示意图

4.2.5　损失函数

目前用于该领域生成模型的主要方法是自回归变压器网络和扩散模型。本章节的目标是对这些方法进行微调，以利用机器人导航的反馈来优化场景布局。生成场景本质上是一个预测对象（3D 包围盒）的大小、位置、方向和类别的过程。通常，物体的类别和大小受到房间类型的约束，因此位置和方向的变化是影响路

径规划算法的主要因素。本章节的微调模块主要专注于优化这两个属性，并利用上一小节提出的 ScoreNet 来微调整个系统。

训练过程旨在以路径规划得分为指导对现有的场景生成网络进行微调，优化网络参数以生成可信且对机器人友好的场景。本章节的训练损失由两部分组成：i）原始场景生成模型的损失$L_{original}$，该模型一般限制生成的对象集为近似底层数据的分布和底层数据的分布；ii）基于 ScoreNet 预测的 PPS 的路径规划损失L_{PPS}。具体来说，本章节简单地定义$L_{PPS} = -PPS$，即为 PPS 值较低的场景分配较高的损失。最终训练损失L_{total}定义为：

$$L_{total} = L_{original} + \lambda \cdot L_{PPS} \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (4.3)$$

其中λ(经调整后设置为$2e-3$)是一个超参数，用于控制机器人规划得分对场景生成的影响程度。在这里，本章节探讨了路径规划算法对场景生成的内在影响。特别是在考虑参数λ趋近于无穷大的极端情况时，系统会产生不合理的优化结果：所有对象都会以堆叠的形式聚集，这种以牺牲场景合理性为代价追求路径规划优化的极端做法违背了实际应用的要求。基于此，本章节最终通过参数调优选择$\lambda = 2e-3$，在以下两个核心维度上达到最佳平衡：场景可信度，即保证生成的场景符合物理规律和现实逻辑，避免物体堆叠等反直觉现象；人机交互效率，即在保证场景可信性的前提下，使机器人的路径规划效率最大化。在训练过程中，通过最小化L_{total}对场景生成网络进行微调。

4.3　实验结果与分析

4.3.1　场景生成任务的实验分析

本文的预训练阶段，本章节遵循如 ATISS 和 DiffuScene 的默认实验设置。本章节的得分网络模型是在一个批量大小为 128 的 A40 GPU 上进行训练的。该模块最初训练了 1000 次，卧室、客厅和餐厅的 PPS 预测精度分别为 0.98、0.95 和 0.94。然后，将 ScoreNet 集成到生成模型中，并用于微调其预训练的场景生成网络 100 个 epoch。学习率初始化为$lr = 2e-4$，然后以每 15 个 epoch 0.5 的衰减率逐渐降低。在推理过程中，本章节首先使用祖先采样策略获得对象属性，然后根据生成的形状代码在 3D-FUTURE 中为每个对象检索最相似的 CAD 模型。

本章节将本方法与 3D 场景生成任务的两种最先进的方法进行了比较，分别是如 ATISS 和 DiffuScene：（1）ATISS 是一个基于变压器的自回归网络，它将场景视为一组无序的对象，并按顺序生成对象和属性。将场景视为一组无序的对象，并按顺序生成对象及其特征。（2）DiffuScene 是一种扩散模型，它在用固定物体的属性填充场景后将其视为二维矩阵。这两种方法都可以生成合理的室内布局，但本章节的初步实验表明，这两种基线在生成对机器人导航友好的室内场景时面临挑战。为了解决这个问题，本章节用 Nav2Scene 机制增强了它们，提高了它们生成机器人友好的场景的能力。

首先，本章节描述了覆盖路径规划算法的设置。在经典配置中，覆盖装置安装在机器人的中心位置，通常有一个有限的范围，比如机器人的半径。本章节的路径规划算法已应用于所有具有圆形足迹的 3D-FRONT 数据集，覆盖半径为 r_c=0.3 m。在对数据集进行预处理后，本章节分别基于不同类型的室内房间进行训练。

为了进行定量比较，首先对室内场景的路径规划得分进行评估。此外，根据先前的研究，本章节使用弗雷切特起始距离（FID），核起始距离（KID），场景分类精度（SCA）和类别 KL 散度（CKL）来测量 1,000 个合成场景的可行性和多样性。本章节将生成的和真实的场景都呈现为可用于 FID、KID 和 SCA 这三项指标评估的自上而下的正交图像。每个家具对象的纹理由与其语义类别相关联的唯一颜色决定。本章节在所有方法中使用统一的相机和渲染设置，以确保公平的比较。对于 CKL，本章节计算了合成场景和真实场景的语义类别分布之间的 KL 散度。对于 FID、KID 和 CKL，越低的值表示越接近数据分布。FID 和 KID 也捕捉到了结果的多样性。对于 SCA，接近 50%的分数表明生成的场景与真实场景无法区分。

在上面描述的优化模块的基础上，本章节可以以最小的修改支持各种下游任务。为此，本章节进行了一系列相关实验，并比较了不同任务之间的结果。首先是场景生成任务的相关实验。本章节的框架能够直接生成与认知先验一致的多样化和以人为中心的室内布局，从而提高室内场景内机器人路径规划的效率和合理性。

如表 4.2 所示，在无条件场景生成任务中，本章节使用路径规划得分指标将

本章节的方法与其他方法的性能进行了比较。在本章节努力创建无约束场景的过程中，本章节使用路径规划评分对本章节的方法与其他方法的性能进行了彻底的比较。结果清楚地表明，本章节的方法是高度机器人友好的，显著提高了不同房间类型的路径规划分数。此外，与来自 3D-FRONT 数据集的原始场景相比，本章节的方法生成的场景显示出显着的改进。这一令人信服的证据突出了本章节的方法如何有效地克服了数据集的局限性，并展示了提高数据质量的巨大潜力。

表 4.2　无条件场景生成任务的定量对比结果

房间类型	方法	SCA	FID↓	KID↓	CKL↓
卧室	ATISS	0.56	21.84	1.83	1.08
	ATISS+Ours	**0.45**	**21.47**	**1.66**	**0.77**
	DiffuScene	0.66	25.34	2.98	0.49
	DiffuScene+Ours	**0.64**	**26.95**	**2.79**	**0.45**
客厅	ATISS	0.52	44.06	4.80	0.43
	ATISS+Ours	**0.61**	**43.78**	**4.62**	**0.29**
	DiffuScene	0.72	44.17	5.51	0.48
	DiffuScene+Ours	**0.63**	**44.91**	**8.06**	**0.37**
餐厅	ATISS	0.40	40.61	4.23	0.20
	ATISS+Ours	**0.59**	**40.08**	**3.52**	**0.17**
	DiffuScene	0.62	38.29	1.14	0.39
	DiffuScene+Ours	**0.61**	**39.01**	**4.64**	**0.39**

表 4.3 呈现了基于多种评估指标的定量对比结果。本章节针对不同方法生成的场景开展路径规划算法实验，进而获取了相应的路径规划性能指标（PPS）。从实验结果来看，本章节所采用的方法在几乎所有评估指标上均展现出显著优势，明显优于其他对比方法。这一结果有力地证实了本章节方法在生成场景方面的有效性。具体而言，其不仅能够为机器人提供更适宜的运行环境，也进一步表明 Nav2Scene 方法对机器人具备良好的友好性。此外，该方法还能生成更多样化且更具可信度的场景，这为机器人在复杂环境下的高效运行提供了有力支撑。值得关注的是，在室内场景的原始训练过程中，由于未充分考虑机器人因素，使得在与机器人相关的指标评估中，数据集里真实场景的得分相对较低。这一现象为后

续研究提供了重要启示，在后续的研究中，应充分考虑机器人因素对场景生成的影响，以进一步优化场景生成方法，提升机器人在实际应用中的性能表现。

表 4.3　无条件场景生成的路径规划评分（PPS）对比结果

方法	路径规划得分（PPS ↑）		
	卧室	客厅	餐厅
ATISS	0.86	0.84	0.91
ATISS+Ours	**0.94**	**0.92**	**0.97**
DiffuScene	0.87	0.83	0.92
DiffuScene+Ours	**0.96**	**0.93**	**0.99**
数据集场景	0.88	0.87	0.91

图 4.6 直观地呈现了在无条件生成任务中，不同场景生成方法的定性比较结果。这一定性实验揭示了现有场景合成方法普遍存在的空间不一致性问题，具体表现为家具间距过小以及重叠安排等不合理现象。在现有的方法中，ATISS 尝试通过自回归的场景先验来解决上述问题，而 DiffuScene 则采用带有去噪机制的全局布局编码方式。然而，这两种方法在实际应用中均难以有效确保生成布局的空间一致性。与之形成鲜明对比的是，本章节所提出的优化模块展现出显著优势。该模块在设计过程中，充分考虑了机器人通行的便利性，能够综合自然且多样化的场景安排。从图 4.6 中可以清晰看到，以每两行图片为一组进行观察，上图部分将本章节的方法与当前最先进的生成方法进行了对比。结果表明，本章节的方法在场景生成方面具有更出色的表现。再看下图，其为路径规划路径的可视化展示。通过对比可以发现，本章节的方法所生成的规划路径更加简洁。这意味着在实际应用中，基于本章节方法生成的场景布局能够为机器人提供更高效、更流畅的通行路径。

此外，经过优化处理后，本章节方法的实验结果还呈现出更多合理化的放置特征。例如，家具摆放能够与墙面实现完美对齐，并且出现了更多对称的家具对。这种合理化的放置不仅提升了场景的美观度，更重要的是进一步优化了空间利用效率，使得整个场景布局更加科学、合理。综上所述，本章节所提出的优化模块在解决现有场景合成方法的空间不一致性问题上具有显著成效，并且在场景生成和路径规划方面展现出明显的优势，为相关领域的研究和应用提供了新的思路和方法。

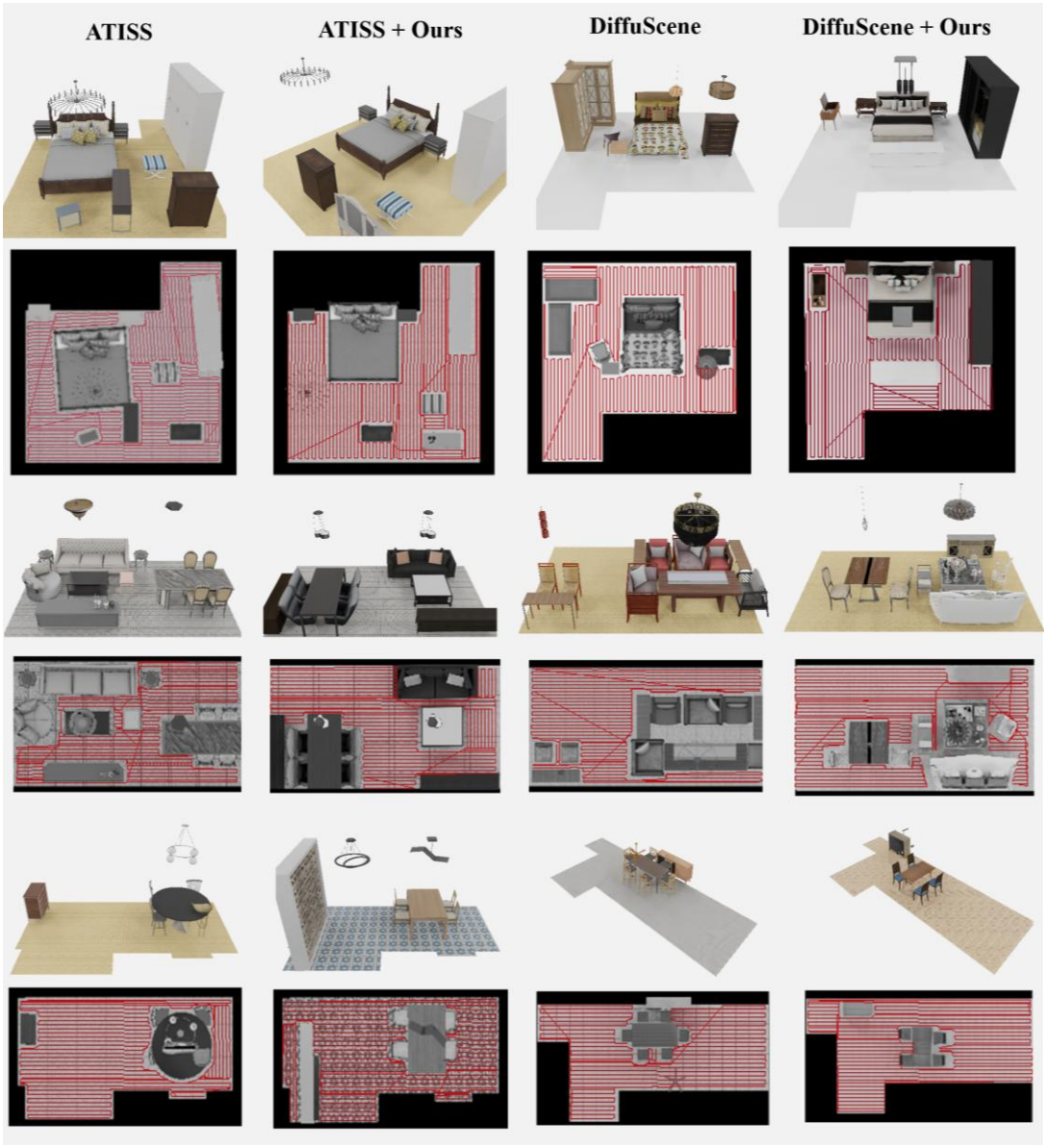

图 4.6　无条件场景生成的可视化对比结果

除了场景生成任务的实验，本章节还分别进行了一些下游任务的实验，具体包括场景补全和场景重排列。下面小节 4.3.2 分别对其进行详细阐述。

4.3.2　下游任务的实验分析

场景完成。假设一个有$M(\leq N)$个对象的部分场景，即$y \in R^{M\times D}$，本章节利用从生成模型中学习到的场景先验将新的$\hat{x}_0$补充到y_0，从而得到一个完整的对象集合$x_0 = (y_0, \hat{x}_0)$。经过 ScoreNet 的微调过程后，机器人在完成场景的同时会考虑对添加对象的放置偏好。因此，本章节的模型成功地生成了包含灯具、衣柜和梳妆台等多个对象的合理补全。

本章节将本方法与 ATISS 和 DiffuScene 进行场景补全任务的比较。如图 4.7 所示，同样，本章节在红框中突出显示了有问题的家具，并同时在本章节的结果中用绿色标记了相对应的家具，以方便区分。很明显，在 ATISS 生成的卧室结果中，灰色的鞋架放置得太靠近床的尽头，它靠近梳妆台的位置挡住了过道。从功能上来说，棕色和灰色的鞋架造成了功能上的冗余。在 DiffuScene 生成的结果中，两件家具之间没有明显的重叠。相比之下，本章节的卧室布局展示了一种最佳安排，有效地保持了家具之间的连通性，同时确保了它们在逻辑和功能放置上的合理性。

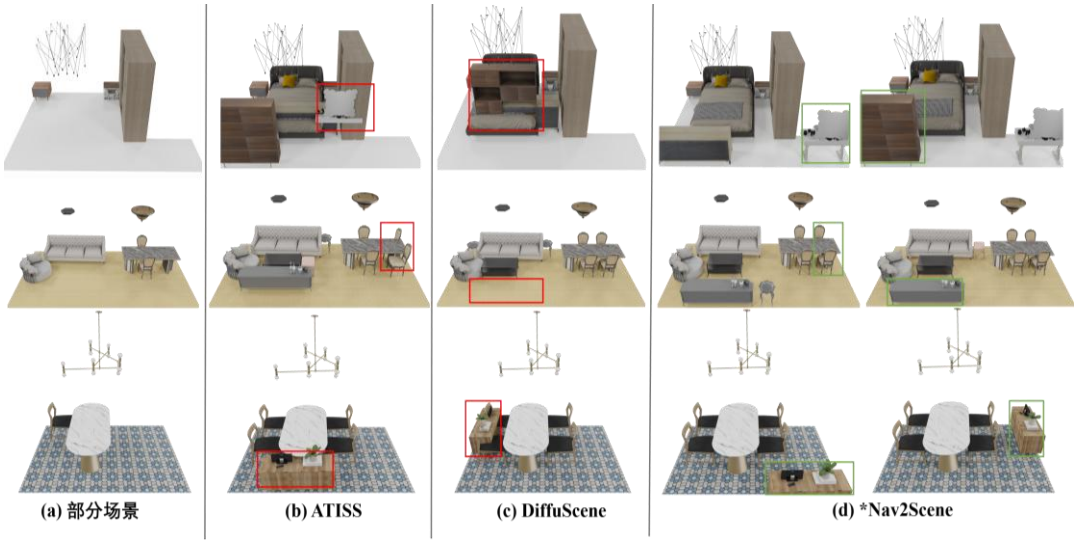

图 4.7　基于部分场景的场景补全的可视化结果

对比实验结果可以看出，该方法经过微调后，客厅和餐厅场景生成结果与 ATISS 和 DiffuScene 相比有明显改善，家具布局呈现出更自然的空间秩序。值得注意的是，在涉及空间配置受限的情况下（例如，紧凑的起居区），ATISS 生成的布局在床头柜和衣柜之间仍旧存在位置冲突。这种限制主要源于空间维度固有的物理约束。通过系统优化，本章节实施了几项空间合理化措施，首先是消除非必要的功能性家具（例如视觉参考中的粉红色和灰色凳子）；其次，策略性地重新定位咖啡桌，以确保走道畅通；最后对齐墙壁的橱柜放置优化，以保持足够的活动间隙。这些空间调整共同解决了布局冲突，同时保持了基本的功能要求。

场景重新安排。给定一组具有随机空间位置的对象，本章节可以使用路径规划模块来优化边界框编码。首先，本章节随机初始化物体的位置。然后，使用本章节的路径规划算法优化框架来重新排列物体，仅仅调整了它们的位置和方向，

以方便机器人的工作。在这个方向上，本研究获得了令人信服的实验结果。

实验中，本节比较了场景补全任务与 DiffuScene。如图 4.8 所示，卧室型房间的生成结果揭示了几个问题。ATISS 的输出显示棕色凳子侵占了过道，白色扶手椅同样阻碍了床和衣柜之间的通道。同样，DiffuScene 展示的实验结果也有两个明显的错误：椅子随意摆放，与墙壁没有对齐，白色扶手椅挡住了过道，产生了穿模现象。相比之下，本章节的结果展示了一个更合乎逻辑的场景布局、更自然的排列，床头柜适当地靠墙放置，显著增强了机器人的导航路径。此外，本章节的方法精确地纠正了客厅里咖啡桌、凳子和无序的餐椅的位置，从而使布局配置更加有序。

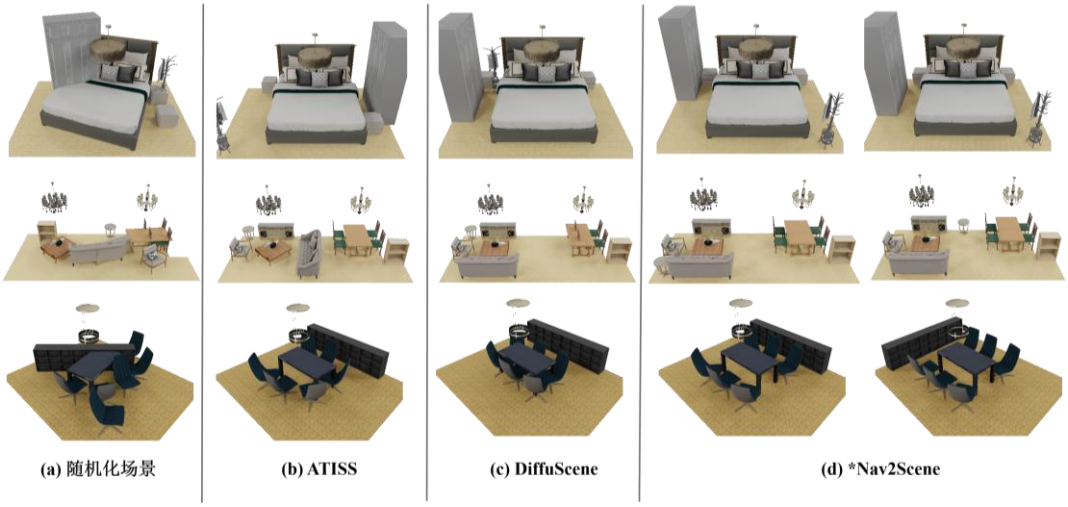

图 4.8　基于随机化场景的重新排列的可视化结果

总的来说，本章节提出的方法在室内空间智能化设计领域展现出显著优势，其创新性体现在两个方面：首先通过引入动态路径规划偏好机制，系统性地优化了家具布局的空间逻辑性。如表 4.4 所示，相较于传统方法在路径规划指标上获得显著性能提升，主要得益于算法对房间布局的系统评估，这种评估不仅考虑家具间的物理间隔，更深入分析典型活动场景中的人体工学需求。其次，在保持与原始方法相当的 FID 和 KID 指标水平下，本方法使重组后的家具配置既维持视觉合理性，又提升功能可达性。值得指出的是，这种提升并非以牺牲空间美感为代价，通过约束条件能在优化布局密度的同时保持设计元素的协调统一。实验还表明，这种增强的空间导航性源自多维度特征融合机制：算法在微调阶段同步考虑家具尺寸分布的统计特性、智能体的路径规划轨迹以及功能分区的关联强度，从而生成更符合实际使用需求的布局方案。这些技术改进共同促成了方法在复杂

室内场景中的优越表现，为智能家居设计系统提供了新的技术实现路径。

表 4.4　卧室和客厅的场景重新排列任务的定量对比结果

房间类型	方法	PPS ↑	FID ↓	KID ↓
卧室	ATISS	0.83	22.32	1.58
	ATISS+Ours	**0.92**	**21.18**	**1.56**
	DiffuScene	0.88	26.25	1.03
	DiffuScene+Ours	**0.96**	**23.16**	**1.03**
客厅	ATISS	0.85	45.23	4.83
	ATISS+Ours	**0.93**	**44.37**	**4.07**
	DiffuScene	0.79	44.29	**6.03**
	DiffuScene+Ours	**0.88**	**44.23**	8.61

4.3.3　消融实验

本章节进行了严谨的消融实验，以验证场景生成模型中路径规划指标公式中覆盖范围和时间因素权重的有效性。定量结果如表 4.5 所示。

表 4.5　卧室的无条件场景生成任务的消融实验结果

方法	PPS ↑	FID ↓	KID ↓	SCA	CKL ↓
$\alpha = 0, \beta = 1.0$	0.90	22.36	1.82	0.42	0.79
$\boldsymbol{\alpha = 0.2, \beta = 0.8}$	**0.94**	**21.47**	**1.66**	**0.45**	**0.77**
$\alpha = 0.5, \beta = 0.5$	0.86	22.07	1.79	0.42	0.83
$\alpha = 1.0, \beta = 0$	0.78	25.13	1.85	0.40	1.03

在这个实验中，本章节的公式包括两个关键因素:覆盖率和花费在非重复路径上的时间比例。覆盖率是覆盖路径规划算法中的主要评价指标，它受到扫地机器人的覆盖半径和旋转角度的影响。为了一致性，本章节在整个实验过程中保持设置统一，覆盖半径设置为$0.2\,m$，旋转角度设置为$\pi/2\,rad$。关于后一个因素，本章节承认智能代理的效率对于未来的智能家居环境至关重要。因此，将花费在非重复路径上的时间比例作为时间维度的评估指标。两个因素的总权重之和为 1.0，用百分比表示。经过多次调整，本章节确定了最优的一组值。

4.3.4　局限性分析

尽管本章节的方法在一定程度上提升了现有室内场景生成方法的机器人适应性与场景连通性，以及生成结果的整体质量，但该方法仍存在一定的局限性和改进空间。本章节通过将路径规划分数作为反馈信号，优化了现有室内场景生成模型的连通性和机器人适应性。然而，这种优化信号并未解决家具碰撞等问题，仍然可能导致物理上不合理的场景。此外，本章节简单地将模拟机器人假设为具有规则形态的圆柱形扫地机器人，其执行的任务也相对简单。而现实中，未来的家用机器人可能具有复杂的形态结构并执行更加多样化的任务。这些挑战未来可能需要通过更全面的数据集或在逼真仿真环境中的交互来共同解决。

4.4　本章小结

本文提出 Nav2Scene，一种生成适配机器人导航场景的创新框架。具体而言，本章节提出一种即插即用的场景生成微调模块，可无缝集成到 ATISS、DiffuScene 等主流场景生成架构中，为建模机器人友好型室内场景提供新工具。针对现有场景合成数据集缺少机器人评估标注的问题，本章提出了一个机器人路径规划评估数据集，用于支撑本章的核心内容，即利用路径规划算法去优化场景布局。通过基于 ScoreNet 模型预测的路径规划分数对模型进行微调，该框架有效提升了室内场景深度生成模型输出的机器人适应性与场景连通性。通过与现有先进方法进行大量定量与定性对比，实验结果证明 Nav2Scene 在增强生成布局连通性和机器人友好性的同时，对 FID、KID 等生成质量指标亦有一定程度提升。该框架为构建兼顾机器人导航与交互需求的智能场景生成系统提供了新的研究视角。未来的工作可以进一步探索动态机器人友好场景的实时优化机制。

第 5 章　论文工作总结与展望

5.1　总结

随着人工智能与机器人技术的快速发展，将具身智能融入室内场景合成已成为智能家居、服务机器人等领域的核心研究方向。然而，传统场景生成方法在机器人功能适配性和用户意图理解方面存在显著不足。本文针对这两大问题提出了两篇创新性工作，即 CoT2Scene 与 Nav2Scene。为了更好的解析复杂指令或保证空间布局的合理性，CoT2Scene 创新性地将思维链与大语言模型结合，提出分步生成框架：知识增强的意图对齐、逻辑结构化生成流程、高质量 3D 模型检索。为了进一步研究给定场景的优化，本文提出了一个即插即用的场景生成优化模块，即 Nav2Scene 框架，该框架可以无缝集成到主流的场景生成架构中，如 ATISS 和 DiffuScene，并通过基于 ScoreNet 预测的路径规划得分来优化室内场景的空间布局，该框架有效地提高了机器人对生成场景的导航效率。两个工作均是基于 3D-FRONT 完成，并可相互融合，以构建导航功能优化与用户意图对齐的室内场景合成框架。因此，本文从场景生成方法的数据集和方法两方面展开研究，主要完成的工作如下：

（1）本文提出融合具身智能思想的 CoT2Scene 场景生成框架。该框架创新性地结合思维链技术增强大语言模型的空间推理能力。并基于具身认知理论模拟人类空间理解机制，通过层级式 CSS 提示模板实现认知链路的显式表达，将抽象设计指令分解为空间布局、物体定位、关系校验等具象化操作步骤，以此实现分步式场景生成。为强化生成场景的物理可信度，该框架引入环境强耦合约束机制，通过提取 3D-FRONT 先验知识构建外部知识库，在物体碰撞检测、家具空间关系等维度形成闭环反馈系统，最终达成生成场景在空间连贯性、用户意图契合度与设计规范符合性三个维度的协同优化。该工作最终实现了给定用户关于室内场景的描述类文本，能够快速生成与之高度契合的房间布局，并确保生成场景中家具的合理化分布。

（2）本文提出 Nav2Scene 模块化优化框架。该框架通过即插即用的微调机制，将导航兼容性指标融入现有场景生成框架的训练流程，在不影响原始生成质

量的前提下，显著提升输出场景的机器人导航适配性。具体而言，本方法提出以路径规划得分作为机器人导航适配性的核心评估指标。通过整合任务时间与任务完成度两大关键效能维度对室内场景进行量化评估。为此，本方法基于 3D-FRONT 数据集构建路径规划评估数据集，该数据集创新性地将 ROS 系统实现的路径规划算法与 PPS 计算相结合，形成具有导航性能标注的新型训练数据，为场景生成模型提供多维特征支持。本方法通过基于该数据集训练 ScoreNet 模型，实现对任意给定室内场景的路径规划得分自动预测，以此对场景生成框架进行微调优化。实验表明经该框架优化的生成模型可同时保持室内场景的视觉质量、内容多样性和导航可行性。

5.2　展望

虽然本研究的框架在场景合成方面展示了令人信服的结果，但仍存在多维度改进空间。未来工作可从以下方向展开深入研究：

（1）针对机器人形态与任务的复杂性，未来可建立任务驱动型场景生成范式。通过构建包含机械臂运动学模型、移动机器人底盘参数等特征的机器人本体库，结合强化学习在虚拟家庭环境中进行多任务训练，使生成算法能适配不同构型机器人的作业需求。

（2）在技术路径优化层面，未来可尝试将当前基于单一大语言模型的架构升级为多模态大模型协作系统。例如，采用 DeepSeek 等国产大模型处理中文语义理解，同时结合视觉语言模型（Vision-Language Model[78]，VLM）解析空间拓扑关系，形成跨模态的知识表示与推理机制。对于知识库构建问题，可引入图神经网络改进关系抽取过程，通过自动化实体关系发现机制动态扩展知识图谱规模，并设计基于注意力机制的检索增强生成框架提升推理效率。

（3）数据集方面，未来工作中可采用迁移学习策略缓解 3D-FRONT 数据集类型不均衡问题。通过跨房间类型的特征解耦学习，结合风格迁移技术生成稀有类别样本，同时引入真实世界扫描数据构建混合训练集。值得关注的是，具身智能领域的快速发展为场景生成提供了新机遇，未来可探索人-机-环境协同演化框架，使场景生成系统能根据机器人在仿真环境中的交互反馈持续优化布局策略，最终实现动态自适应的智能空间构建。这些技术突破将推动室内场景生成从静态

布局向支持机器人自主作业的认知智能空间演进，为智慧家庭、康复护理等垂直领域提供更强大的数字化支持。

参考文献

[1] 杨淼, 陈宝权. 室内场景生成算法综述[J]. 集成技术, 2022, 11(1): 40-51.

[2] Yao Y, Duan J, Xu K, et al. A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly[J]. High-Confidence Computing, 2024: 100211.

[3] Johannessen C M, Boehm J S, Kim S Y, et al. COT drives resistance to RAF inhibition through MAP kinase pathway reactivation[J]. Nature, 2010, 468(7326): 968-972.

[4] Duan J, Yu S, Tan H L, et al. A survey of embodied ai: From simulators to research tasks[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2022, 6(2): 230-244.

[5] Grisetti G, Kümmerle R, Stachniss C, et al. A tutorial on graph-based SLAM[J]. IEEE Intelligent Transportation Systems Magazine, 2010, 2(4): 31-43.

[6] Gao L, Sun J M, Mo K, et al. Scenehgn: Hierarchical graph networks for 3d indoor scene generation with fine-grained geometry[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(7): 8902-8919.

[7] Wang S, Zeng W, Chen X, et al. Actfloor-gan: Activity-guided adversarial networks for human-centric floorplan design[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 29(3): 1610-1624.

[8] He F, Huang Y, Wang H. iplan: Interactive and procedural layout planning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 7793-7802.

[9] Liang Z, Xu G, Wu H, et al. S-INF: Towards Realistic Indoor Scene Synthesis via Scene Implicit Neural Field[J]. arXiv preprint arXiv:2412.17561, 2024.

[10] Zhang S, Tong H, Xu J, et al. Graph convolutional networks: a comprehensive review[J]. Computational Social Networks, 2019, 6(1): 1-23.

[11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[12] Purkait P, Zach C, Reid I. Sg-vae: Scene grammar variational autoencoder to generate new indoor scenes[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 155-171.

[13] Ma Y, Zhan D, Jin Z. Fastscene: Text-driven fast 3d indoor scene generation via panoramic gaussian splatting[J]. arXiv preprint arXiv:2405.05768, 2024.

[14] Li H, Shi H, Zhang W, et al. Dreamscene: 3d gaussian-based text-to-3d scene generation via formation pattern sampling[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 214-230.

[15] Zhou M, Hou J, Luo C, et al. Scenex: Procedural controllable large-scale scene generation via large-language models[J]. arXiv e-prints, 2024: arXiv: 2403.15698.

[16] Wang K, Lin Y A, Weissmann B, et al. Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks[J]. ACM Transactions on Graphics (TOG), 2019, 38(4): 1-15.

[17] Li M, Patil A G, Xu K, et al. Grains: Generative recursive autoencoders for indoor scenes[J]. ACM Transactions on Graphics (TOG), 2019, 38(2): 1-16.

[18] Song S, Yu F, Zeng A, et al. Semantic scene completion from a single depth image[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1746-1754.

[19] Fu H, Cai B, Gao L, et al. 3D-FRONT: 3d furnished rooms with layouts and semantics[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10933-10942.

[20] Hua B S, Pham Q H, Nguyen D T, et al. Scenenn: A scene meshes dataset with annotations[C]//2016 fourth international conference on 3D vision (3DV). Ieee, 2016: 92-101.

[21] Li Z, Yu T W, Sang S, et al. Openrooms: An open framework for photorealistic indoor scene datasets[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 7190-7199.

[22] Sohl-Dickstein J, Weiss E, Maheswara Nathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//International conference on

machine learning. pmlr, 2015: 2256-2265.

[23] Dhiman C. DepthNet: A monocular depth estimation framework[C]//2021 International Conference on Engineering and Emerging Technologies (ICEET). IEEE, 2021: 1-6.

[24] Xiao J, Owens A, Torralba A. Sun3d: A database of big spaces reconstructed using sfm and object labels[C]//Proceedings of the IEEE international conference on computer vision. 2013: 1625-1632.

[25] Dai A, Chang A X, Savva M, et al. Scannet: Richly-annotated 3d reconstructions of indoor scenes[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5828-5839.

[26] Roberts M, Ramapuram J, Ranjan A, et al. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10912-10922.

[27] Zheng J, Zhang J, Li J, et al. Structured3d: A large photo-realistic dataset for structured 3d modeling[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. Springer International Publishing, 2020: 519-535.

[28] Fu H, Jia R, Gao L, et al. 3d-future: 3d furniture shape with texture[J]. International Journal of Computer Vision, 2021, 129: 3313-3337.

[29] Nguyen H T, Chen Y, Voleti V, et al. HouseCrafter: Lifting Floorplans to 3D Scenes with 2D Diffusion Model[J]. arXiv preprint arXiv:2406.20077, 2024.

[30] Chandrasekhar V R, Chen D M, Tsai S S, et al. The stanford mobile visual search dataset[C]//Proceedings of the second annual ACM conference on Multimedia systems. 2011: 117-122.

[31] Handa A, Pătrăucean V, Stent S, et al. Scenenet: An annotated model generator for indoor scene understanding[C]//2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2016: 5737-5743.

[32] Mittler R. ROS are good[J]. Trends in plant science, 2017, 22(1): 11-19.

[33] Salehinejad H, Sankar S, Barfett J, et al. Recent advances in recurrent neural

networks[J]. arXiv preprint arXiv:1801.01078, 2017.

[34] Li Z, Liu F, Yang W, et al. A survey of convolutional neural networks: analysis, applications, and prospects[J]. IEEE transactions on neural networks and learning systems, 2021, 33(12): 6999-7019.

[35] Yuan L, Chen Y, Wang T, et al. Tokens-to-token vit: Training vision transformers from scratch on imagenet[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 558-567.

[36] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.

[37] Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in neural information processing systems, 2017, 30.

[38] Yang L, Zhang Z, Song Y, et al. Diffusion models: A comprehensive survey of methods and applications[J]. ACM Computing Surveys, 2023, 56(4): 1-39.

[39] Brooks S. Markov chain Monte Carlo method and its application[J]. Journal of the royal statistical society: series D (the Statistician), 1998, 47(1): 69-100.

[40] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer international publishing, 2015: 234-241.

[41] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in neural information processing systems, 2022, 35: 24824-24837.

[42] Obukhov A, Krasnyanskiy M. Quality assessment method for GAN based on modified metrics inception score and Fréchet inception distance[C]//Software Engineering Perspectives in Intelligent Systems: Proceedings of 4th Computational Methods in Systems and Software 2020, Vol. 1 4. Springer International Publishing, 2020: 102-114.

[43] Chen R, Huang W, Huang B, et al. Reusing discriminators for encoding: Towards unsupervised image-to-image translation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8168-8177.

[44] Van Erven T, Harremos P. Rényi divergence and Kullback-Leibler divergence[J]. IEEE Transactions on Information Theory, 2014, 60(7): 3797-3820.

[45] 任皎.人体工程学在家具设计中的应用[J]包装工程，2014，35(18)：50-52.

[46] Kim T, Oh J, Kim N Y, et al. Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation[J]. arXiv preprint arXiv:2105.08919, 2021.

[47] Feng W, Zhu W, Fu T, et al. Layoutgpt: Compositional visual planning and generation with large language models[J]. Advances in Neural Information Processing Systems, 2023, 36: 18225-18250.

[48] Paschalidou D, Kar A, Shugrina M, et al. Atiss: Autoregressive transformers for indoor scene synthesis[J]. Advances in Neural Information Processing Systems, 2021, 34: 12013-12026.

[49] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. Advances in neural information processing systems, 2022, 35: 27730-27744.

[50] Fu R, Wen Z, Liu Z, et al. Anyhome: Open-vocabulary generation of structured and textured 3d homes[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 52-70.

[51] Yang Y, Sun F Y, Weihs L, et al. Holodeck: Language guided generation of 3d embodied ai environments[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 16227-16237.

[52] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PmLR, 2021: 8748-8763.

[53] Mildenhall B, Srinivasan P P, Tancik M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1):

99-106.

[54] Kumar V, Rao V N. Parallel depth first search. part ii. analysis[J]. International Journal of Parallel Programming, 1987, 16(6): 501-519.

[55] Mao R, Chen G, Zhang X, et al. GPTEval: A survey on assessments of ChatGPT and GPT-4[J]. arXiv preprint arXiv:2308.12488, 2023.

[56] Guo D, Zhu Q, Yang D, et al. DeepSeek-Coder: When the Large Language Model Meets Programming--The Rise of Code Intelligence[J]. arXiv preprint arXiv:2401.14196, 2024.

[57] Wang X, Yeshwanth C, Nießner M. Sceneformer: Indoor scene generation with transformers[C]//2021 International Conference on 3D Vision (3DV). IEEE, 2021: 106-115.

[58] Armeni I, Sax S, Zamir A R, et al. Joint 2d-3d-semantic data for indoor scene understanding[J]. arXiv preprint arXiv:1702.01105, 2017.

[59] Yeshwanth C, Liu Y C, Nießner M, et al. Scannet++: A high-fidelity dataset of 3d indoor scenes[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 12-22.

[60] 岳亮，谈皓，黄俊凯，等. 数字室内三维场景构建综述[J]. 中国图象图形学报，2024，29(9)：2471-2493.

[61] 霍凤财，迟金，黄梓健，等. 移动机器人路径规划算法综述[J]. 吉林大学学报（信息科学版），2019，36(6)：639-647.

[62] Tang J, Nie Y, Markhasin L, et al. DiffuScene: Denoising diffusion models for generative indoor scene synthesis[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 20507-20518.

[63] Bähnemann R, Lawrance N, Chung J J, et al. Revisiting boustrophedon coverage path planning as a generalized traveling salesman problem[C]//Field and Service Robotics: Results of the 12th International Conference. Springer Singapore, 2021: 277-290.

[64] Choset H. Coverage for robotics–a survey of recent results[J]. Annals of mathematics and artificial intelligence, 2001, 31: 113-126.

[65] Latombe J C, Latombe J C. Exact cell decomposition[J]. Robot motion planning, 1991: 200-247.

[66] Galceran E, Carreras M. A survey on coverage path planning for robotics[J]. Robotics and Autonomous systems, 2013, 61(12): 1258-1276.

[67] Wong S C, MacDonald B A. A topological coverage algorithm for mobile robots[C]//Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453). IEEE, 2003, 2: 1685-1690.

[68] Huang W H. Optimal line-sweep-based decompositions for coverage algorithms[C]//Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164). IEEE, 2001, 1: 27-32.

[69] Bähnemann R, Lawrance N, Chung J J, et al. Revisiting boustrophedon coverage path planning as a generalized traveling salesman problem[C]//Field and Service Robotics: Results of the 12th International Conference. Springer Singapore, 2021: 277-290.

[70] Sharma A, Pathak J, Prakash M, et al. Object detection using OpenCV and python[C]//2021 3rd international conference on advances in computing, communication control and networking (ICAC3N). IEEE, 2021: 501-505.

[71] Zheng K. Ros navigation tuning guide[J]. Robot Operating System (ROS) The Complete Reference (Volume 6), 2021: 197-226.

[72] Tang G, Tang C, Claramunt C, et al. Geometric A-star algorithm: An improved A-star algorithm for AGV path planning in a port environment[J]. IEEE access, 2021, 9: 59196-59210.

[73] Tanemura M, Ogawa T, Ogita N. A new algorithm for three-dimensional Voronoi tessellation[J]. Journal of Computational Physics, 1983, 51(2): 191-207.

[74] Chang L, Shan L, Jiang C, et al. Reinforcement based mobile robot path planning with improved dynamic window approach in unknown environment[J]. Autonomous robots, 2021, 45: 51-76.

[75] Wu J, Ma X, Peng T, et al. An improved timed elastic band (TEB) algorithm of

autonomous ground vehicle (AGV) in complex environment[J]. Sensors, 2021, 21(24): 8312.

[76] Choset H, Pignon P. Coverage path planning: The boustrophedon cellular decomposition[C]//Field and service robotics. London: Springer London, 1998: 203-209.

[77] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. Advances in neural information processing systems, 2022, 35: 27730-27744.

[78] Zhang J, Huang J, Jin S, et al. Vision-language models for vision tasks: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.