Diffusion base：

不用太关注diffusion部分。

两篇论文：

INSTRUCTSCENE：Lin, Chenguo, and Yadong Mu. INSTRUCTSCENE: INSTRUCTION-DRIVEN 3D INDOOR SCENE SYNTHESIS WITH SEMANTIC GRAPH PRIOR.

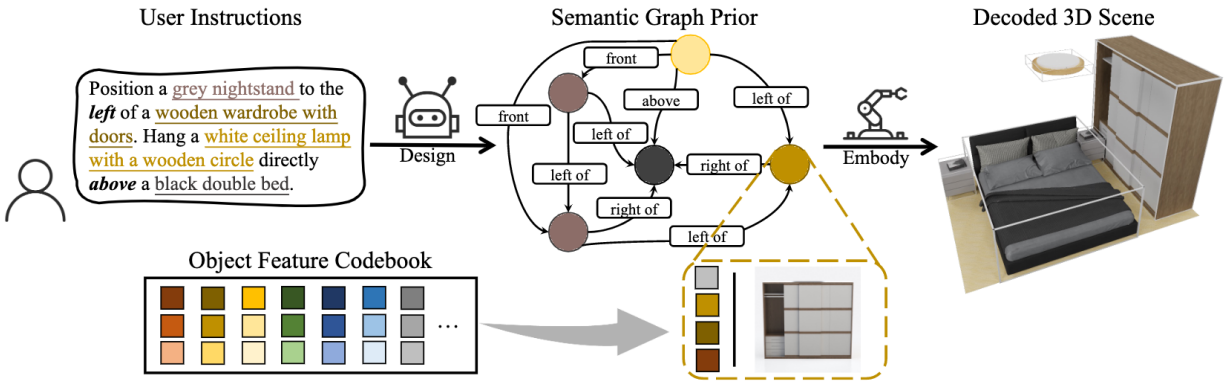
DiffuScene：Tang, Jiapeng, et al. DiffuScene: Scene Graph Denoising Diffusion Probabilistic Model for Generative Indoor Scene Synthesis. Mar. 2023.

第一篇：INSTRUCTSCENE

1.解决了什么问题？创新点？

提出了一个指令驱动的生成框架，该框架集成了语义图先验和布局解码器，以提高 3D 场景合成的可控性和保真度。

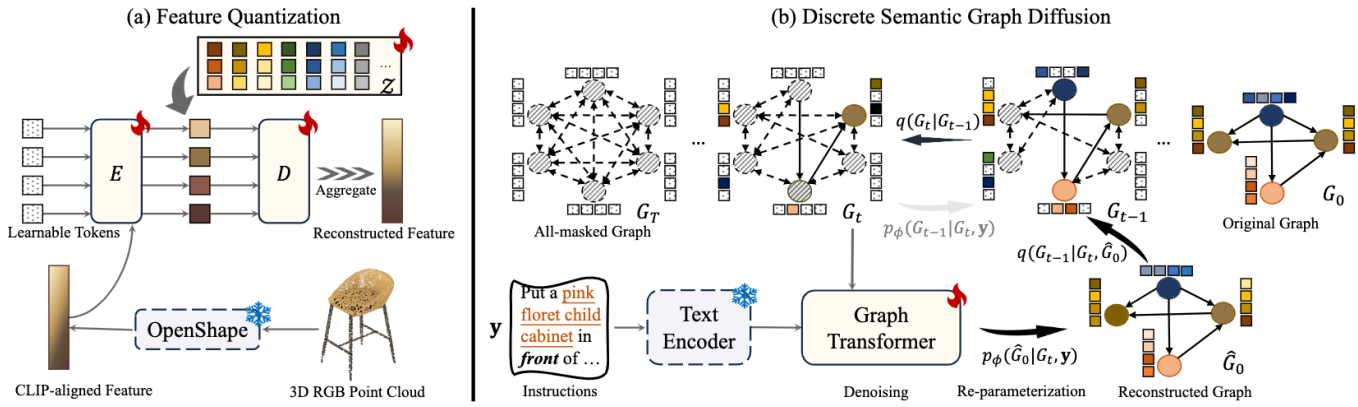
2.模型结构？用了什么method？



INSTRUCTSCENE主要包括两部分：语义图先验（Semantic Graph Prior）和布局解码器（Layout Decoder）

1.Semantic Graph Prior

通过学习场景中高级对象和关系分布的条件分布，生成具有一定结构和语义的图。



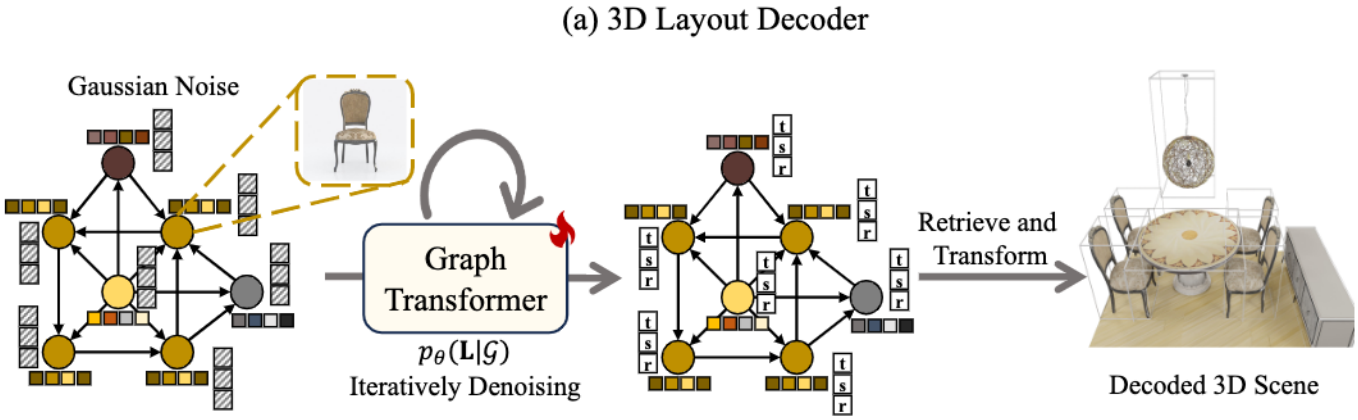
语义图先验的步骤如下：

1.**特征量化**：这一步的目的是把高维特征向量转化为低维特征向量。Openshape 被用于从场景中提取物体的视觉外观和几何形状，提取物体语义特征。

2.**离散语义图扩散**：使用离散扩散模型，根据输入的离散语义图，输出预测后验分布的语义图。【因为扩散模型不是我所需要重点关注的，所以这块的知识没有细看】

2.Layout Decoder

利用生成的语义图生成具有语义一致性的场景布局。



如图所示，以语义图为条件，通过在节点嵌入上附加采样的高斯噪声，然后迭代去噪以产生布局属性。

3.效果如何?

Instruction-driven Synthesis		↑ iRecall%	↓ FID	↓ FID^{CLIP}	↓ KID$_{\times 1e\text{-}3}$	SCA%
Bedroom	ATISS	48.13±2.50	119.73±1.55	6.95±0.06	0.39±0.02	59.17±1.39
	DiffuScene	56.43±2.07	123.09±0.79	7.13±0.16	0.39±0.01	60.49±2.96
	Ours	**73.64**±1.37	**114.78**±1.19	**6.65**±0.18	**0.32**±0.03	**56.02**±1.43
Living room	ATISS	29.50±3.67	117.67±2.32	6.08±0.13	17.60±2.65	69.38±3.38
	DiffuScene	31.15±2.49	122.20±1.09	6.10±0.11	16.49±1.24	72.92±1.29
	Ours	**56.81**±2.85	**110.39**±0.78	**5.37**±0.07	**8.16**±0.56	**65.42**±2.52
Dining room	ATISS	37.58±1.99	137.10±0.34	8.49±0.23	23.60±2.52	67.61±3.23
	DiffuScene	37.87±2.76	145.48±1.36	8.63±0.31	24.08±1.90	70.57±2.14
	Ours	**61.23**±1.67	**129.76**±1.61	**7.67**±0.18	**13.24**±1.79	**64.20**±1.90

4.本文的一些局限性：

1.数据集的准确性不高

2.场景规模比较小，且仅关注了室内的环境生成。

5.展望：将大型语言模型（LLMs）整合到指令驱动的流程中

第二篇：DiffuScene

1.解决了什么问题？创新点？

1）为不同的室内场景合成引入了 3D 场景图去噪扩散概率模型，该模型学习对象一致性、放置和几何的整体场景配置。该模型不止只预测对象的bounding box，而是将语义、定向边界框和几何特征融合在一起，以促进对组成结构和表面几何形状的整体理解。

2）基于该模型，便于从部分场景完成、现有场景中的对象重新排列以及文本条件场景合成。采用场景图表示从数据集中学习复杂的场景组合模式，避免了人为定义的约束和迭代优化过程。

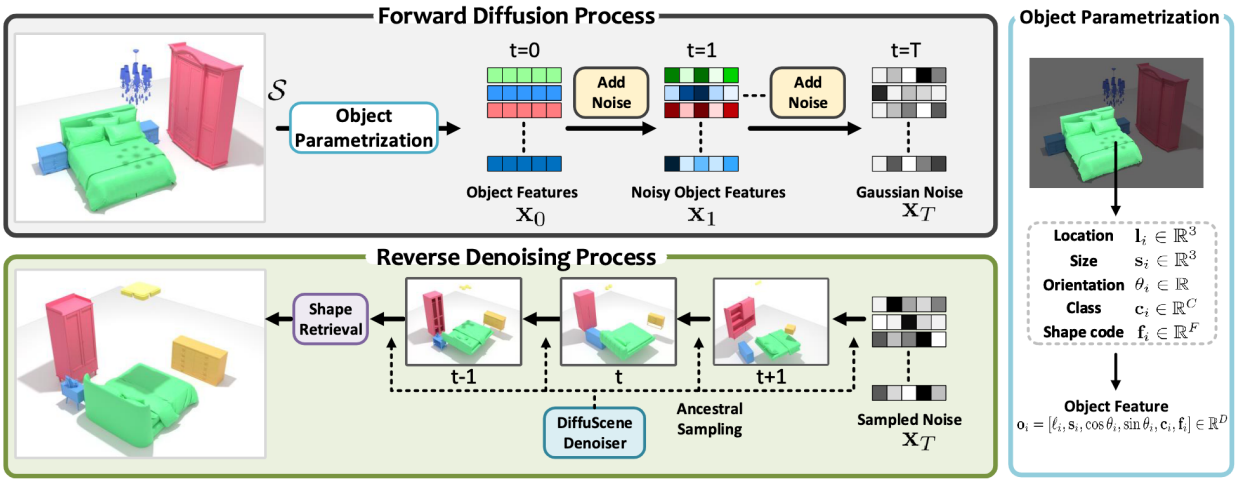
【此创新点来由】：传统的场景建模和合成方法将这个问题表述为一个数据驱动的优化任务，该任务由三个关键模块组成:场景公式、先验表示和优化策略。

1.为了提取对象之间的空间关系，许多方法将场景中的对象表述为图 ，并将它们与人体连接以构建以人为中心的图 。

2.为了合成合理的 3D 场景图，需要合理场景的先验知识（对象频率分布，人体运动的可视性图）来驱动场景优化（迭代方法、非线性优化、手动交互等）。

2.模型结构？用了什么method?

模型：**DiffuScene**，一个场景图去噪扩散概率模型，旨在学习3D室内场景的分布，包括语义类、表面几何形状和多个对象的位置。



过程：

1. 给定一个 N 个对象的 3D 场景 S，我们通过将每个对象参数化为：存储所有对象属性的图节点，来获得其全连接场景图 x_0 , 即位置、大小、方向、类标签和潜在形状代码。

2. 基于一组所有可能的 x_0，使用去噪扩散概率模型：

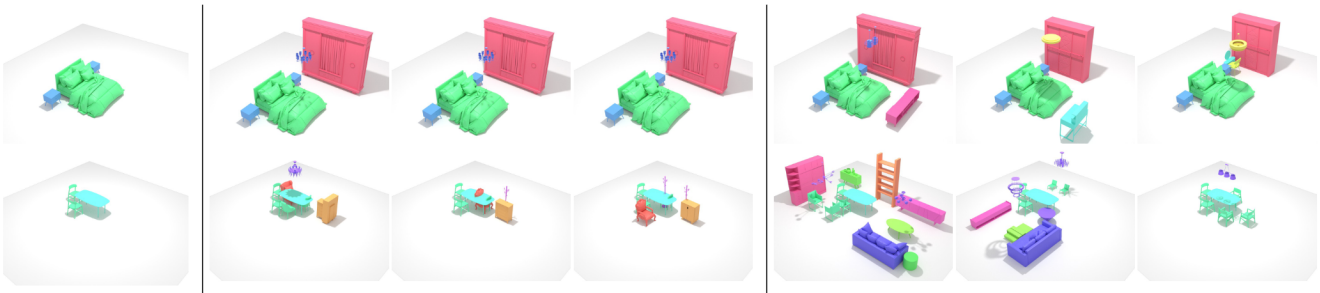
i) 在前向过程中，逐渐将噪声添加到x_0中，直到获得标准的高斯噪声 x_T。

ii) 在反向过程中，即生成过程，去噪网络（有跳跃连接+注意块的MLP）使用原始采样迭代地清理噪声图。

iii) 最后，使用去噪后的对象特征进行形状检索，用于真实场景合成.

3.效果，局限性与展望

效果：更有真实性、丰富性和。



(a) Partial Scenes (b) ATISS [42] (c) Ours

Figure 5: Scene completion from partial scenes with only 3 objects given as inputs. Compared to ATISS, our diffusion-based method produced more diverse completion results with higher fidelity.