LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding, Reasoning, and Planning

1.解决了什么问题？创新点：

1.解决的问题：在复杂3D环境中实现对文本指令和视觉交互的理解和推理，通过多模态指令调整，实现了对不同任务的区分和响应，以及根据文本指令和视觉提示生成有效的3D场景嵌入。模型最后输出了一个自由形式的自然语言，其中部分内容可以被解释为3D坐标。

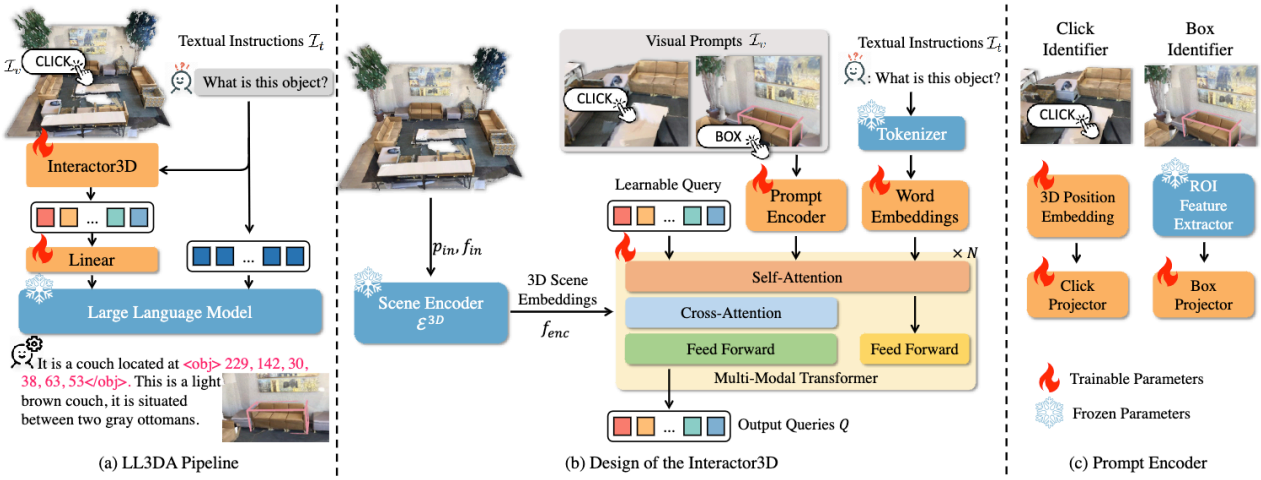
2.创新点：

1）提出了LL3DA大模型，利用注意力机制的思想，用于在复杂3D环境中的理解、推理和规划。

2）模型同时考虑了文本指令和视觉交互作为输入，并提取了交互感知特征以实现有效的指令跟随。

3）通过引入额外的视觉交互（用户点击+3D框注释），进一步消除模糊的文本指令中的歧义。

2.用了什么方法？method：



2.1 模型的输入和输出

Input：1.一组点云表示3D场景；2.文本指令 I_t ；3. 作为补充空间标识符的潜在视觉交互 I_v

Output：自由形式的自然语言，其中一部分可以被描述为3D坐标。

其中，1）指令格式以： ### human 开始， ### assistant 作为生成响应标识符。

2）坐标表示中，点和边界框的表示如下：点（Point）用 x, y, z 表示。边界框（Bounding Box）用其中心点和大小表示，即 c_x, c_y, c_z, w, h, l。这些数值都被离散化为[0, 255]范围内的无符号整数，以适应输入3D场景的边界。这种设计使得大型语言模型能够理解和处理3D坐标，而无需引入任何额外的可学习标记。

2.2 模型的设计

step1：模型先通过Interactor3D 聚合了固定长度的场景嵌入。【将上面提到的3D场景、文本指令和视觉交互作为输入】

step2：将聚合的**场景嵌入**投影到文本指令的前缀，作为冻结LLM的输入。

核心设计在于**Interactor3D**，它通过注意力机制聚合信息，生成固定长度的交互感知场景嵌入，作为 LLM的输入前缀。其详细的结构示意图为上图的(b), 组成内容结构如下：

- 一个冻结的3D场景编码器 ε^{3D} ： 采用在ScanNet检测上预训练的掩码变压器编码器作为场景编码器，输出为 f_{enc} : d维度特征，使用最远点采样(FPS)算法从输入点中下采样得到；
- 一个视觉提示编码器：主要有两种用户交互，用户点击和3D框注释；
  - 用户点击 (P_{click})：归一化然后进行傅里叶位置嵌入
  - 3D框注释：使用 ROI特征提取

然后分别用相同的前馈网络投影。【前馈网络 FFN：单层的mlp，说白了就是FC层，也称为全连接层，指的是神经网络中每个节点都与上一层的所有节点相连。】

- 一个Q-Former，用于将3D嵌入转换为固定长度的交互感知场景嵌入。

  Q-Former是基于Multi-Modal Transformer(MMT)的架构，MMT的功能和作用主要有以下三点：

  1. 解决3D场景嵌入与位置敏感的因果语言模型之间的矛盾：MMT通过将可学习的查询令牌与编码后的视觉提示和文本指令进行交互，实现了对位置敏感的因果语言模型的融合。【参考原理：BLIP v2】
  2. 连接冻结的单模态模型：MMT通过让可学习的查询令牌与视觉提示进行交互，弥合了冻结的模型之间的差距。
  3. 进行交互感知特征提取：MMT允许可学习的查询令牌与视觉提示与任务无关的3D场景嵌入进行交互，从而满足了交互感知特征提取的需求。

  在实际应用中，MMT的输出是32个查询，最终通过一个简单的线性投影将其映射到LLM的嵌入空间。

3.效果与局限性?

△ 补充：对自然语言生成的评估标准中，指标如下：

1. CiDEr（C）：一种衡量生成文本与人类编写的参考文本相似度的指标，考虑了n-gram重复惩罚以避免重复。
2. BLEU-4（B-4）：衡量生成文本与人类编写的参考文本之间的相似度的指标，基于n-gram匹配计算。
3. METEOR（M）：一种衡量生成文本与人类编写的参考文本之间的相似度的指标，考虑了表面形式的匹配、短语匹配和synonym匹配。
4. Rouge-L（R）：衡量生成文本与人类编写的参考文本之间的相似度的指标，基于最长公共子序列（LCS）计算。

这些指标用于评估生成的自然语言文本与真实的人类编写的文本之间的相似度，以评估模型在理解、推理和规划任务中的性能。

1.效果如下：在相同的数据集下有比较好的效果。

Method	ScanRefer								Nr3D			
	C@0.25↑	B-4@0.25↑	M@0.25↑	R@0.25↑	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
Scan2Cap[11]	56.82	34.18	26.29	55.27	39.08	23.32	21.97	44.78	27.47	17.24	21.80	49.06
MORE[29]	62.91	36.25	26.75	56.33	40.94	22.93	21.66	44.42	-	-	-	-
SpaCap3D[51]	-	-	-	-	44.02	25.26	22.33	45.36	33.71	19.92	22.61	50.50
REMAN[38]	62.01	36.37	26.76	56.25	45.00	26.31	22.67	46.96	34.81	20.37	23.01	50.99
D3Net[7]	-	-	-	-	46.07	30.29	24.35	51.67	33.85	20.70	23.13	53.38
Contextual[62]	-	-	-	-	46.11	25.47	22.64	45.96	35.26	20.42	22.77	50.78
UniT3D[12]	-	-	-	-	46.69	27.22	21.91	45.98	-	-	-	-
3DJCG[4]	64.70	40.17	27.66	59.23	49.48	31.03	24.22	50.80	38.06	22.82	23.77	52.99
3D-VLP[30]	70.73	41.03	28.14	**59.72**	54.94	32.31	24.83	51.51	-	-	-	-
3D-VisTA*[65]	-	-	-	-	61.60	34.10	**26.80**	55.00	-	-	-	-
Vote2Cap-DETR[9]	71.45	39.34	**28.25**	59.33	61.81	34.46	26.22	54.40	43.84	26.68	25.41	54.43
LL3DA (Ours)	**74.17**	**41.41**	27.76	59.53	**65.19**	**36.79**	25.97	**55.06**	**51.18**	**28.75**	**25.91**	**56.61**

【没有开源，暂时跑不了】

补充：BLIPS：v2，提出视觉和语言模型的预训练任务

1.了解 ViT：

ViT的整体思想还是比较简单，主要是将图片分类问题转换成了序列问题。即将图片patch转换成token，以便使用Transformer来处理。听起来很简单，但是ViT需要在海量数据集上预训练，然后在下游数据集上进行微调才能取得较好的效果，否则效果不如ResNet50等基于CNN的模型。

2.了解 BERT模型：

BERT的全称为Bidirectional Encoder Representation from Transformers，是一个预训练的语言表征模型。它强调了不再像以往一样采用传统的单向语言模型或者把两个单向语言模型进行浅层拼接的方法进行预训练，而是采用新的**masked language model（MLM）**，以致能生成**深度的双向**语言表征。

3. BLIP：

   预训练分成2步：固定ViT，进行视觉和语言的表征学习；固定llm模型，学习从图像生成文本。

4.