表 2: 闭集三维实例分割设置比较
ScanNet200 [52]。在整体分割性能和每个子集上， SOLE 与 OpenMask3D [58]进行了比较。SOLE 明显优于 OpenMask3D
在六项评估指标中的五项上。

方法	AP	AP50	AP25	APhead	APcom	APtail
OpenMask3D [58]	15.4	19.9	23.1	17.1	14.1	14.9
SOLE（我们的）	20.1 (+4.7)	28.1 (+8.2)	33.6 (+10.5)	27.5 (+10.4)	17.6 (+3.5)	14.1 (-0.8)
面具3D [53]	26.9	36.2	41.4	39.8	21.7	17.9

评估指标。采用不同 IoU 阈值的平均精度 (AP)
作为评估指标，包括 25% IoU 以下的 AP、50% IoU 和平均 AP
从 50% 增加到 95% IoU。

4.2 与以前方法的比较

闭集三维实例分割。我们将 SOLE 与
在闭集 3D 实例分割设置上，类分割方法[10,11,66]和掩码训练方法[26,58] 取得了不错的效果。与
类分割方法相比，
我们对新的类别进行评估。从表1 中的比较结果来看，我们
可以得出以下观察结果。首先，SOLE 的表现明显优于
即使不使用基类标签，类拆分方法也能实现很大的改进。
其次，尽管 OpenIns3D [26]利用了与
我们的 SOLE 在AP50和AP25 上分别大幅超越了 33.5% 和 32.5%，
第三，我们的 SOLE 甚至可以达到与
尽管没有使用
类标签。最后，我们提供了 SOLE 的两个变体来进一步验证我们的有效性。体素大小为
4cm 的 SOLE 利用 4cm 体素大小而不是 2cm 作为
在以前的工作中。较小的体素尺寸可以节省内存需求，并且
以牺牲精度为代价来加快模型速度。尽管使用了较小的体素尺寸，
体素大小为 4cm 的 SOLE 仍能大幅超越之前的作品。
此外，我们验证了我们的框架的有效性不仅限于
通过在没有任何额外文本信息的情况下进行实验，即 SOLE w/o text sup，来测试字幕模型和 NLP 工
具。在这个实验中，由于字幕是

不可用。尽管只使用掩码视觉关联进行训练，该模型仍然可以实现最先进的性能。此外，我们比较了
SOLE
表 2 中，我们评估了两种方法在 ScanNet200 [52]上的整体分割性能以及在单个 CNN 上的性能。

三个子集。 SOLE在六个指标中的五个上优于 OpenMask3D [58] ，并在尾部类别上实现了相当的性
能。
ScanNet200 [52]进一步证明了我们框架的有效性。

分层和跨域开放集 3DIS。为了评估我们工作的泛化能力，我们将 SOLE 与 OpenMask3D [58]进行了比
较

表 3:ScanNetv2 [8]→ScanNet200 [52]上分层开放集 3D 实例分割设置的比较。 SOLE 在基础类和新类上与 Open-Mask3D [58]进行了比较,并取得了最佳效果。

方法	小说课程			基类			所有职业	
	AP	AP50	AP25	AP	AP50	AP25	AP	APtail
OpenMask3D [58]	11.9	15.2	17.8	14.3	18.3	21.2	12.6	11.5
SOLE（我们的）	19.1 (+8.8)	26.2 (+11.0)	30.7 (+12.9)	17.4 (+3.1)	26.2 (+7.9)	32.1 (+10.9)	18.7 (+6.1)	12.5 (+1.0)

表 4: 开放集三维实例分割设置比较
ScanNet200 [52]→Replica [56]。 SOLE在所有方面都优于 OpenMask3D [58]
评估指标。

方法	面具训练 AP		AP50	AP25
OpenMask3D [58]	ScanNet200 [52]	13.1	18.4	24.2
SOLE（我们的）	ScanNet200 [52]	24.7 (+11.6)	31.8 (+13.4)	40.3 (+16.1)

在开放集设置中,使用 Scannet200 [52]和 Replica [56]数据集。对于 Scan-Net200,两个模型都使用
ScanNetv2 [8] 中的掩码注释进行训练。以下[58]， 53 个语义上接近 ScanNet 的类被分组为

“Base” 类。其余 147 个类被归类为“Novel”类。
表 3 报告了分布外的（“基础”)类别和分布外的（“新”)类别。我们的
SOLE在基础和新颖性方面都远远优于 OpenMask3D [58]
类。此外,为了验证 SOLE 在两种情况下的泛化能力
由于存在域偏移和类别偏移,我们在合成 Replica 基准[56] 上将我们的框架与 Open-Mask3D 进行了比较。
模型在
ScanNet200 上的带注释的掩码。如表 4 所示,我们的方法进一步显示
对来自副本的更多分布外数据具有卓越的稳健性,实现
与 OpenMask3D 相比,AP 分数提高了 11.6%。

4.3 消融研究与分析

在本节中,我们进行了几项消融研究来验证我们的设计选择。
所有研究均在 ScanNetv2 [8]数据集上进行评估。

多模型融合网络。在表5 中,我们进行了成分分析
在多模态融合网络上,验证骨干特征的有效性
集成和跨模态解码器 (CMD)。至于骨干特征集成,利用投影的 2D CLIP 特征 f p （第一行)作为唯一骨干

可以具有更好的语义信息,但缺乏三维几何检测能力，
导致语义识别能力较差。相比之下,单纯使用 3D 实例
骨干特征 f b （第二行)无法继承可泛化的语义信息,导致性能不佳。将这两个特征结合起来

（第三行)在学习过程中可以充分利用广义语义信息

表 5:多模态融合网络的组件分析。3D 实例
骨干特征 f b ,投影二维主干特征 f p ，和跨模态解码器
（CMD） 分别进行调查。

f p编号	乙	命令 AP	AP50	AP25	体素大小
1 ✓		✓ 18.7	36.4	58.1	4厘米
2	✓ ✓ 25.4		47.0	66.0	4厘米
3 ✓ ✓ ✓ 30.8			52.5	70.9	4厘米
4 ✓ ✓		42.8	60.5	68.9	2厘米
5 ✓ ✓ ✓ 44.4			62.2	71.4	2厘米

表 6:多模态关联的成分分析。面具-视觉关联 f MVA,面具-字幕关联 f
马华 和掩码实体关联 f 事物 是
在ScanNetv2上进行了研究[8]。

号 f	MVA编	统 中枢神经系	东西f	AP AP50	AP25	体素大小
1 ✓				24.5	42.0	56.0	4厘米
2		✓		30.4	53.0	68.7	4厘米
			✓ 32.1 29.1 ✓	53.8	70.0	4厘米
3 4 ✓ ✓				✓	50.9	66.8	4厘米
5		30.3 6 ✓ ✓ ✓ 30.8			53.7	70.4	4厘米
					52.5	70.9	4厘米

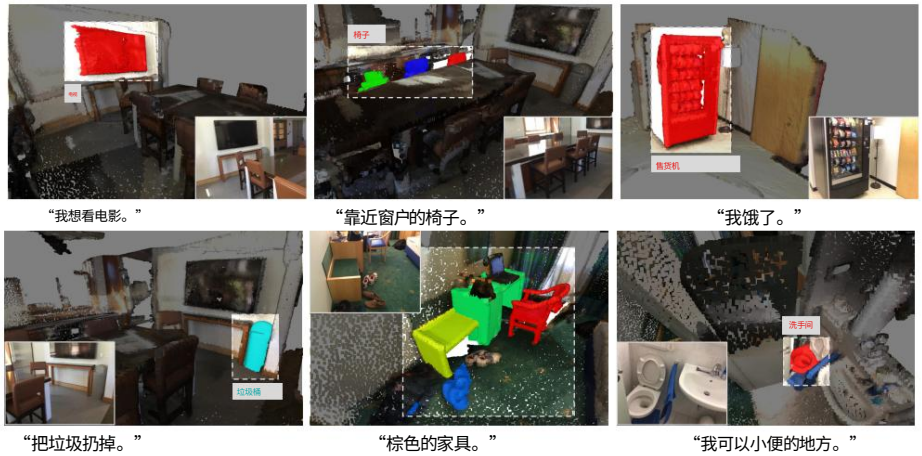
良好的几何检测能力,从而获得最佳结果。此外,跨模态解码器 (CMD) 可以进一步增强

理解语言指令,提高AP1.6%。

多模态关联。我们在表6 中分析了多模态关联的组成部分（f MVA、 f MCA和 f MEA） ,报告了各种组合在 4cm 体素大小的 ScanNetv2 [8]上的得分。我们有以下观察结果。首先,使用任何多模态关联都可以实现显著的

性能,优于之前最先进的方法（OpenIns3D [26]）
体素尺寸较大（分辨率较低）。其次,在三类 as-
协会,掩码实体关联 f MEA是最有效的评估方法
指标,因为它可以将掩码与特定类别对齐。第三,当组合 f
事物 与其他两个关联相比,该模型的性能有所下降
AP 和AP50上的性能下降,而AP25上的性能有所提高。这一观察表明,掩码-视觉关联和掩码-字幕关联可以
有助于语义学习,但会损害掩码准确性。为此,我们在图5 中进一步说明了定性结果。给定一个自由形式的语言指令
类别名称,例如“我想看看外面”,模型仅使用 mask-entity
关联不能分割正确的实例（图5a） ,而结合其他关联的模型（图5b和图5c)可以。因此,尽管

稍微损害基准、蒙版视觉联想和
mask-caption 关联对于识别自由形式的语言指令至关重要,有利于实际场景中的应用。



“我想看电影。” “靠近窗户的椅子。” “我饿了。”

“把垃圾扔掉。” “棕色的家具。” “我可以小便的地方。”

图 4:SOLE 的定性结果。我们的 SOLE 通过有效响应自由形式的语言查询（包括视觉问题、属性描述和功能描述）展示了开放词汇能力。

（一） fMEA (b) fMV A,fMCA图 5:多模态关联的 （c） fMV A,fMCA,fMEA

定性分析。给定自由形式的语言指令“我想看看外面。”，仅使用 f 训练的 SOLE 捕获MCA和 f

事物

错误的对象（（a））,而当 f 另外作为监督给出时,它会分割相关对象（（b），（c））。

定性结果。在图1和图4 中,我们展示了定性结果,表明 SOLE 能够处理自由形式的语言查询,包括但不限于视觉问题、属性描述和功能描述。

5 结论

在本文中,我们提出了一种新颖的框架 SOLE,用于使用自由格式语言指令进行开放词汇 3D 实例分割。SOLE 包含一个多模态融合网络,并由三种类型的多模态关联进行监督,旨在使模型与各种自由格式语言指令保持一致。我们的框架在三个基准测试中的表现远远优于以前的方法,同时实现了与全监督方法相媲美的性能。此外,大量的定性结果证明了我们的 SOLE 对语言指令的多功能性。

附录

在此附录中,我们提供了更多的实施细节,对自由形式语言指令进行了额外的评估,并进行了更多的定性和定量分析。

– A节中提供了更多实施细节。
– SOLE 在 3D 视觉基础任务中进行评估,以验证响应
B节中自由形式语言指令的能力– C节中提供了 CLIP 视觉特征的分析。
– 关于掩码标题和分割结果的更多定性结果
如图D节所示。

实施细节

分割网络。遵循 Mask3D [53],我们使用基于 Transformer 的掩码预测范式来获取实例掩码和语义特征。掩码由对象查询初始化,并通过注意层进行回归。对于每个 3D 点云场景,我们使用最远点采样[48]来获取 150 个点作为对象查询。从分割模型获取掩码后,我们使用 DBSCAN [13]将不连续的掩码分解为更小的空间连续簇,以提高掩码质量。最大距离和邻域点数分别设置为 0.95 和 1。

文本信息生成与提取为了有效地为每个掩码生成标题,我们使用 CLIP 空间中的标题模型,即 DeCap [34]。

DeCap 是一个轻量级的 Transformer 模型,用于从 CLIP 图像嵌入生成字幕。它包含一个 4 层 Transformer,其中有 4 个注意力头作为语言模型,而视觉嵌入则从预先训练的 ViT-L/14 CLIP 模型中获得。我们将从投影的 CLIP 视觉特征中平均池化的掩码特征输入到 DeCap 模型中以获得掩码字幕。然后将字幕集成到文本提示“场景中的 {}”中,以更好地与我们的数据对齐,例如“场景中的一把蓝色椅子”。借助掩码字幕,NLP 库 TextBlob [38]和 spaCy [22] 提取名词短语,以获得掩码实体关联。

B 3D 视觉接地

为了进一步验证 SOLE 对各种语言指令的有效性,我们在 3D 视觉基础基准 ScanRefer [3] 上进行了实验。3D 视觉基础旨在通过自由格式的文本描述来定位 3D 物体。

因此,我们使用 ScanRefer 验证集中的每个文本提示查询 SOLE 以获取相应实例,然后从实例掩码中获取 3D 边界框。性能评估基于 IoU 超过 0.25（ACC@25)和 0.5（ACC@50)的匹配精度。

表 7:ScanRefer [3]上3D 视觉接地任务的结果。SOLE 实现了
在弱监督的通用模型上表现最佳。

方法	类型	监管	ACC@25	ACC@50
奥卡兰德[3]	专家	满的	30.0	29.8
投票兰德[3, 47]		满的	10.0	5.3
性病再障[24]		满的	18.7	6.5
单级[67]		满的	20.4	9.0
扫描参考[3]		满的	41.2	27.4
3D-LLM（火烈鸟） [21]	通才	满的	21.2	
唯一		虚弱的	25.2	22.6

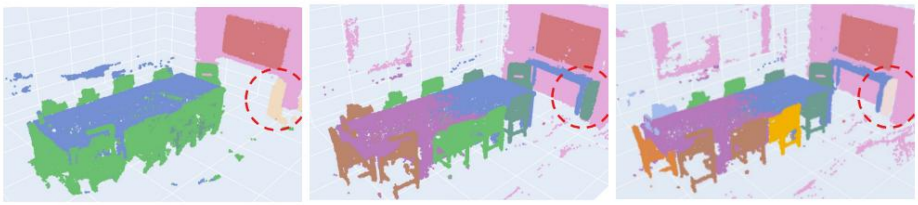
基线。我们将 SOLE 与五个专业基线模型和一个通用模型进行比较。专业模型意味着模型的设计和训练

仅适用于 3D 视觉基础，而通才模型是可以解决
其他任务，如使用类名进行实例分割。对于专家
OCRand [3]使用带有物体真实边界框的 oracle，
并选择一个与对象类别匹配的随机框。VoteRand [3, 47]
利用预先训练的 VoteNet [47]来预测边界框，并随机
选择正确语义类别的框。SCRC [24]和 One-stage [67]是 2D
使用反投影进行 3D 扩展的方法。ScanRefer [3]使用预先训练的 VoteNet [47]和经过训练的 GRU 来选择匹配的边界框。
对于通用模型，3D-LLM 直接预测边界的位置
通过大型语言模型，与文本描述对应的框。注意
除了 OCRand 和 VoteRand 不需要训练外，其余四个
基线模型在 ScanRefer 训练集上进行训练或微调。不同的是，SOLE 仅使用 ScanNetv2 [8]训练集的掩码注释和
ScanRefer 中的文本描述在训练期间被完全忽略。
结果。如表 7 所示，在仅有掩码注释的情况下，SOLE 的表现优于
在ACC@25上，通用模型3D-LLM的成绩提高了4%。此外，SOLE还可以
与接受全面监督的专家同行取得竞争成绩
在 ACC@50 上 (22.6% vs 27.4%)。这样的结果证明了强大的泛化能力和对自由形式语言指令的响应有效性

我们的框架。

C CLIP 视觉特征分析

CLIP视觉特征对SOLE的泛化能力起着重要作用。
在本节中，我们进一步分析了 CLIP 视觉特征在
主干特征集成和推理集成。



（a） f p的 K 均值聚类 （b） f的 K-均值聚类 b （c） f b的 K 均值聚类

图 6:不同骨干特征的 K 均值聚类。不同颜色表示不同的聚类。

表 8:分类概率集成分析。报告结果
在 ScanNetv2 [8]数据集上以 2cm 体素大小进行测试。

成分	AP AP50 AP25体素大小			
合奏团	42.2	58.6	66.9	2厘米
硬几何平均值 软几何平均值	43.7 61.1		70.1	2厘米
（我们的）	44.4 62.2 71.4			2厘米

C.1 主干特征集成

如论文表 5 所示,仅使用 3D 实例骨干特征 f
b 或者投影CLIP视觉特征fp无法达到最佳性能。
3D骨干特征缺乏广义的语义信息,而投影
CLIP 视觉特征缺乏位置和几何信息。为了进一步
验证骨干特征集成的有效性,我们在图 6 中可视化了不同特征的聚类结果。在图6a 中,所有椅子都聚类为

一起（绿色聚类）,表明投影的 CLIP 特征包含
语义信息良好,但无法检测实例。在图6b 中,不同的
可以识别同一类别内的实例,例如,椅子分为三个簇。
然而,语义泛化能力下降。如突出显示的
红色圆圈,垃圾桶由投影的 CLIP 视觉特征检测到（图6a）
但当仅使用 3D 主干时,会错误地将其分类为椅子簇（图6b）。与单独使用这两个特征相比,SOLE 将这两个特征结合起来
从而实现更好的语义泛化能力（分割
垃圾桶)和分割性能（椅子被聚类为 6 个簇）。
可视化结果进一步证明了骨干模型的有效性
特征集成。

C.2 推理集成

在推理过程中,我们将 CLIP 视觉特征与预测的掩码相结合
特征以实现更好的泛化能力。具体而言,在获得
3D 蒙版,每个点的 CLIP 特征在蒙版内进行池化。池化的 CLIP
然后将特征和掩码特征输入到分类器中以获得相应的