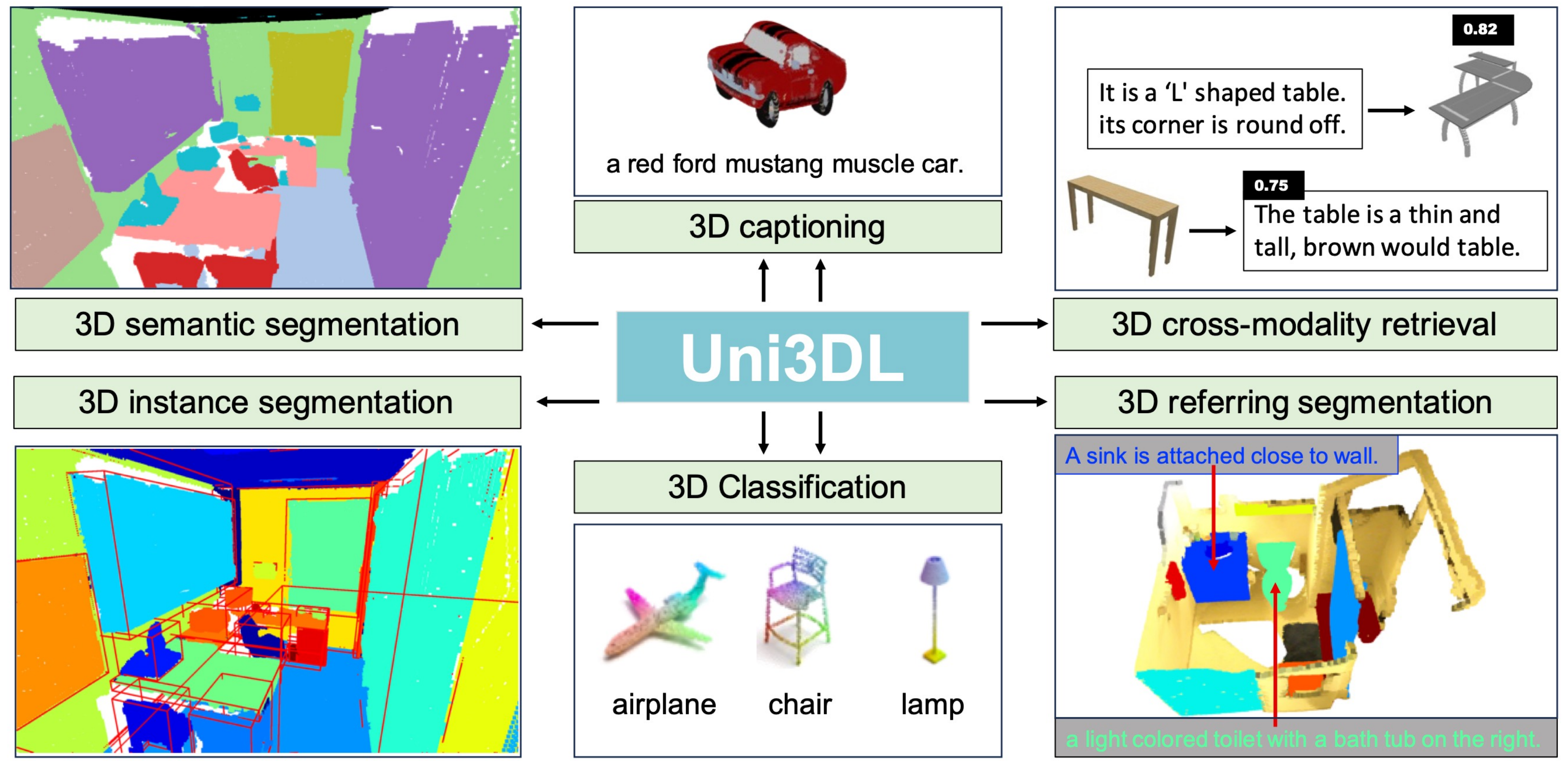
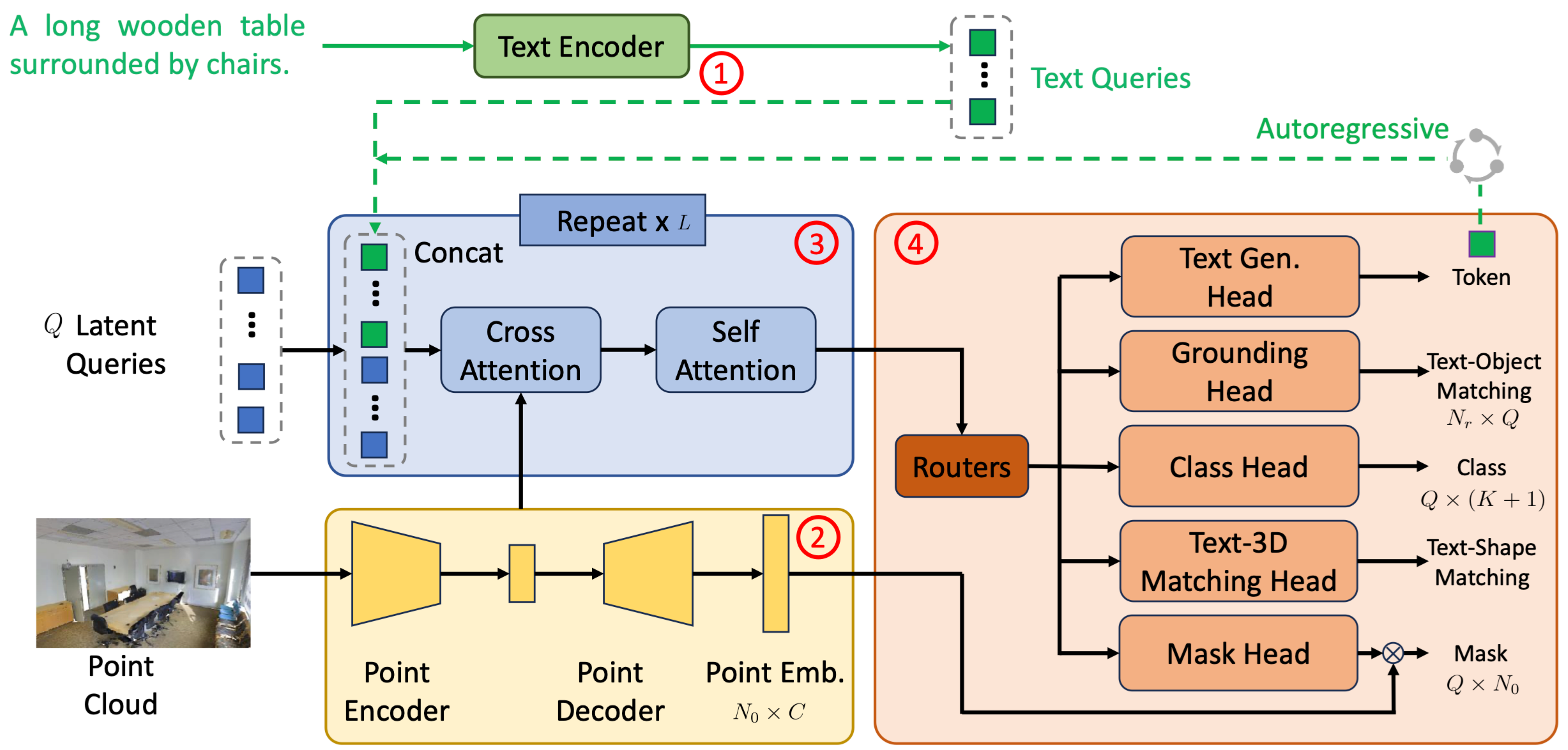
# Uni3DL: Unified Model for 3D and Language Understanding

Xiang Li[1,*], Jian Ding[1,*], Zhaoyang Chen[2], Mohamed Elhoseiny[1]
[1]King Abdullah University of Science and Technology, [2]Ecole Polytechnique

## Single Model for Multiple V&L Tasks in 3D


Uni3DL
3D semantic segmentation
3D instance segmentation
3D captioning
a red ford mustang muscle car.
3D Classification
airplane
chair
lamp
3D cross-modality retrieval
0.82
It is a 'L' shaped table. its corner is round off.
0.75
The table is a thin and tall, brown would table.
3D referring segmentation
A sink is attached close to wall.
a light colored toilet with a bath tub on the right.

## Model Architecture



**Unified Formulation**

$$\mathbf{O}^m, \mathbf{O}^s = \mathcal{D}([\mathbf{F}_Q; \mathbf{F}_T], \mathbf{V})$$

## Task Router

| Task | Obj-Cls | Mask | Grounding | Text-Gen | Matching |
|---|---|---|---|---|---|
| Semantic Segmentation | ✓ | ✓ | | | |
| Instance Segmentation | ✓ | ✓ | | | |
| Grounded Segmentation | | ✓ | ✓ | | |
| Captioning | | | | ✓ | |
| Retrieval | | | | | ✓ |
| Shape Classification | | | | | ✓ |

## Visualization Results of Different Tasks

Input Ours Sem. GT Sem. Ours Inst. GT Inst.

Semantic and instance segmentation

Input GT Ours

this black chair is next to the black couch. it appears to be leather. it is black. there is a snack machine on the opposite wall.

Input GT Ours

it is a small pillow located on the coutch. you can notice it directly on your left when walking through the door into the room.

Referring Segmentation

*GT: a small white NASA space shuttle airplane flying in the sky.*
Ours: a small white airplane flying in the air

*GT: an old red and white car with an American flag painted on it.*
Ours: an old red and white race car with its rear paintings featuring stickers

*GT: a white house with a roof.*
Ours: a white house with a roof and stairs

*GT: a small blue toy car with red accents and a helmet on top.*
Ours: a small blue toy vehicle, resembling a car with wheels

3D Captioning Results

| | top1 | top2 | top3 | top4 | top5 |
|---|---|---|---|---|---|
| a round table with differnt type of look and is good | <br>0.91(GT) | <br>0.90 | <br>0.90 | <br>0.90 | <br>0.88 |

Text-to-Shape Retrieval

## Main Results

| Method | Semantic Segmentation | | | Object Detection | | Instance Segmentation | | | | Grounded Segmentation | | | 3D Captioning | | | 3D Retrieval | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | S3DIS (Area 5) | | SN Val | SN Val | | SN Val | | S3DIS (Area 5) | | ScanRefer | | | Cap3D | | | Text2Shape | |
| | mIoU | mAcc | mIoU | $bAP_{50}$ | $bAP_{25}$ | mAP | $mAP_{50}$ | $mAP_{50}$ | $mAP_{25}$ | mIoU | Acc@0.25 | Acc@0.5 | B-1 | R | M | R@1 | R@5 |
| MinkowskiNet42 [16] | 67.1 | 74.4 | 72.2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| FastPointTransformer [45] | 68.5 | 76.5 | 72.1 | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| PointNeXt-XL [49] | 71.1 | 77.2 | 71.5 | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| StratifiedTransformer [30] | 72.0 | 78.1 | 73.1 | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| PointTransformerV2 [60] | 71.6 | 77.9 | 74.4 | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| EQ-Net [68] | 71.3 | * | 75.3 | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Swin3D [67] | 72.5 | * | 75.2 | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Swin3D[†] [67] | 73.0 | * | 75.6 | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| VoteNet [62] | | - | - | 33.5 | 58.6 | - | - | - | - | - | - | - | - | - | - | - | - |
| 3DETR [43] | - | - | - | 47.0 | 65.0 | - | - | - | - | - | - | - | - | - | - | - | - |
| CAGroup3D [56] | - | - | - | 61.3 | 75.1 | - | - | - | - | - | - | - | - | - | - | - | - |
| PointGroup [29] | * | * | * | * | * | 34.8 | 56.7 | 57.8 | * | - | - | - | - | - | - | - | - |
| MaskGroup [75] | * | * | * | * | * | 42.0 | 63.3 | 65.0 | * | - | - | - | - | - | - | - | - |
| SSTNet [35] | * | * | * | * | * | 49.4 | 64.3 | 59.3 | * | - | - | - | - | - | - | - | - |
| SoftGroup [55] | * | * | * | 59.4 | 71.6 | 50.4 | 76.1 | 66.1 | * | - | - | - | - | - | - | - | - |
| Mask3D [52] | * | * | * | 56.2 | 70.2 | 55.2 | 73.7 | 68.4 | 75.2 | - | - | - | - | - | - | - | - |
| Mask-Att-Free[†] [31] | * | * | * | 63.9 | 73.5 | 58.4 | 75.9 | **69.1** | **75.7** | - | - | - | - | - | - | - | - |
| TGNN (GRU) [25] | - | - | - | - | - | - | - | - | - | 26.1 | 35.0 | 29.0 | - | - | - | - | - |
| TGNN (BERT) [25] | - | - | - | - | - | - | - | - | - | 27.8 | 37.5 | 31.4 | - | - | - | - | - |
| InstructBLIP-7B [18] | - | - | - | - | - | - | - | - | - | - | - | - | 11.2 | 13.9 | 14.9 | * | * |
| InstructBLIP-13B [18] | - | - | - | - | - | - | - | - | - | - | - | - | 12.6 | 15.0 | **16.0** | * | * |
| PointLLM-7B [63] | - | - | - | - | - | - | - | - | - | - | - | - | 8.0 | 11.1 | 15.2 | * | * |
| PointLLM-13B [63] | - | - | - | - | - | - | - | - | - | - | - | - | 9.7 | 12.8 | 15.3 | * | * |
| FTST [9] | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 0.2 | 1.6 |
| FMM [9] | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 0.2 | 2.4 |
| Y2S [22] | - | - | - | - | - | - | - | - | - | - | - | - | * | * | * | 2.9 | 9.2 |
| Parts2Words (no parts) [54] | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 5.1 | 17.2 |
| Ours | **72.7** | **79.3** | **76.2** | **67.7** | **77.1** | **60.9** | **80.9** | 65.3 | 74.3 | **32.3** | **39.4** | **36.4** | **31.6** | **33.1** | 14.4 | **5.7** | **19.7** |

## Ablation Studies

| Task | Grounded Segmentation | Captioning | Retrieval |
|---|---|---|---|
| | ScanRefer | Cap3D | Cap3D |
| | Acc@0.25/Acc@0.5 | B-1/R | T2S R@1/R@5 |
| Ours ($\beta$=1) | 37.8/34.2 | 16.8/13.7 | 5.5/15.5 |
| - Retrieval | 38.8/35.8 | 13.5/11.2 | N/A |
| - Captioning | 38.3/35.5 | N/A | 5.0/12.8 |
| - Instance Segmentation | 35.8/31.0 | 18.2/14.9 | 4.0/11.0 |
| Ours ($\beta$=0.5) | 38.1/36.5 | 15.7/10.3 | 5.5/10.5 |
| Ours ($\beta$=2) | 36.4/34.0 | 18.3/13.4 | 6.0/16.0 |
| Ours ($\beta$=5) | 35.2/31.3 | 17.7/12.0 | 4.0/15.5 |
| Ours + alt. ($\beta$=1) | 36.8/33.6 | 14.8/14.4 | 5.0/13.0 |

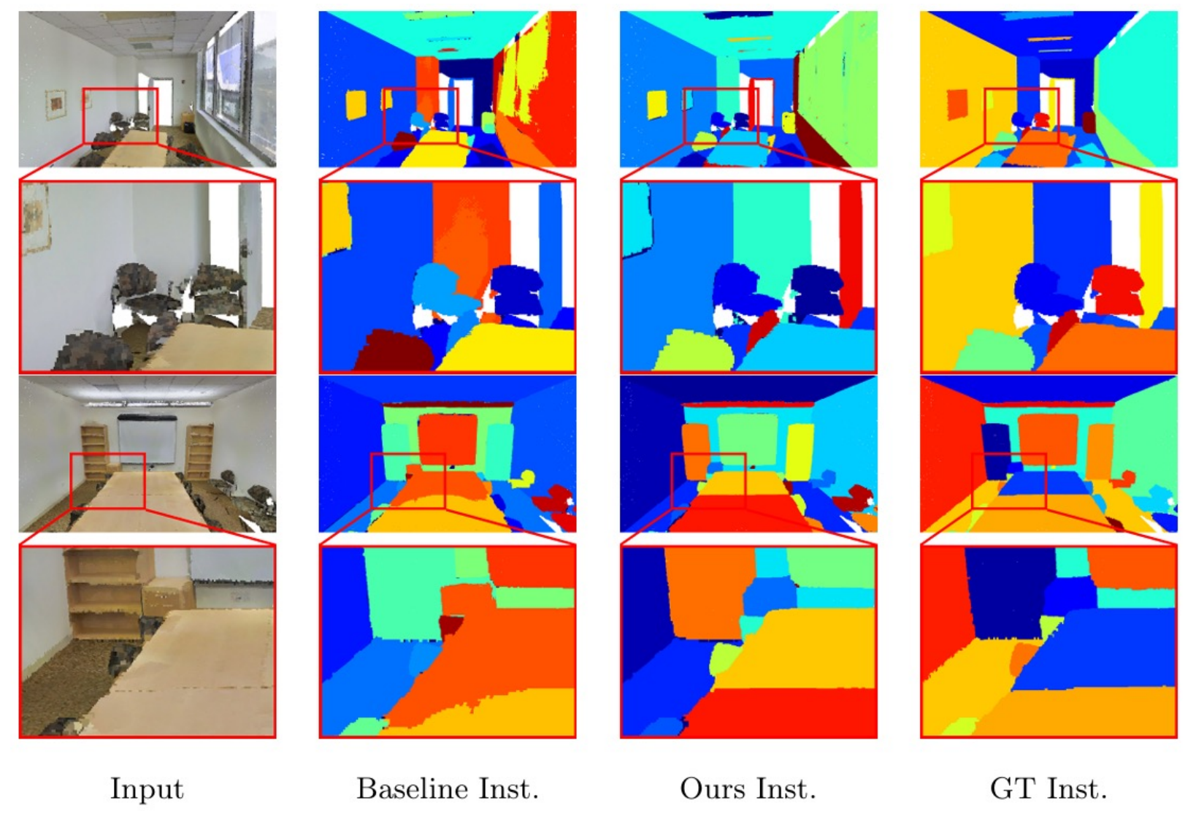
Ablation of scene-object balance

| Task | Semantic Segmentation | Instance Segmentation | Grounded Segmentation | Retrieval |
|---|---|---|---|---|
| | SN Val | S3DIS (Area 5) | ScanRefer | Text2Shape |
| | mIoU/mAcc | $mAP_{50}$ / $mAP_{25}$ | Acc@0.25/Acc@0.5 | R@1/R@5 |
| From scratch | 72.3/81.8 | 61.7/71.7 | 33.8/31.4 | 2.4/7.7 |
| Ours | **76.2/84.8** | **65.3/74.3** | **39.4/36.4** | **5.7/19.7** |

Ablation of pre-training

| Method | Input | Pretraing dataset | Pretrained FM | ModelNet10 | ModelNet40 | |
|---|---|---|---|---|---|---|
| | | | | top-1 | top-1 | top-5 |
| PointCLIP [10] | MV Images | ShapeNet | Yes (CLIP) | 30.2 | 23.8 | - |
| CLIP2Point [6] | MV Images | ShapeNet | Yes (CLIP) | 66.6 | 49.4 | - |
| PointCLIP V2 [13] | MV Images | ShapeNet | Yes (CLIP+GPT3) | **73.1** | 64.2 | - |
| ULIP [8] | MV Images | ShapeNet | Yes (CLIP) | - | 60.4 | 84.0 |
| ULIP [8] | MV Images | Cap3D Objaverse | Yes (CLIP) | - | **67.2** | 83.1 |
| Ours | Point Cloud | Cap3D Objaverse | No | 70.4 | 57.0 | **88.8** |

Zero-shot classification results



Input Baseline Inst. Ours Inst. GT Inst.

Training from scratch *vs.* fine-tuning our model

| Model | Single Stage | Detector | Overall | |
|---|---|---|---|---|
| | | | Acc@0.25 | Acc@0.5 |
| ScanRefer [2] | ✗ | VoteNet | 39.0 | 26.1 |
| InstanceRefer [9] | ✗ | PointGroup | 38.2 | 31.4 |
| 3DVG-Transformer [12] | ✗ | VoteNet | 45.9 | 34.5 |
| 3DJCG [1] | ✗ | VoteNet | **47.6** | 36.1 |
| D3Net [3] | ✗ | PointGroup | - | 35.6 |
| UniT3D [4] | ✗ | PointGroup | - | 36.5 |
| M3DRef [11] | ✗ | PointGroup | - | **40.4** |
| TGNN [5] | ✓ | N/A | 37.4 | 29.7 |
| Uni3DL (Ours) | ✓ | N/A | **37.8** | **33.7** |

Grounded localization performance