XMask3D: Cross-modal Mask Reasoning for Open Vocabulary 3D Semantic Segmentation

Ziyi Wang*, Yanbo Wang*, Xumin Yu, Jie Zhou, Jiwen Lu

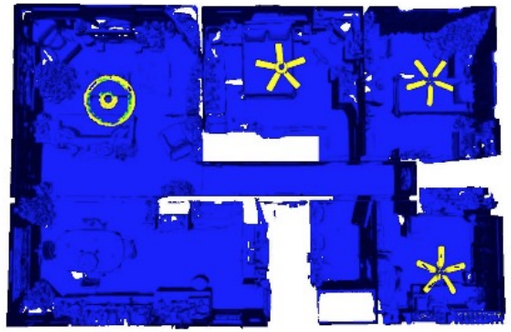
Department of Automation, Tsinghua University

Introduction

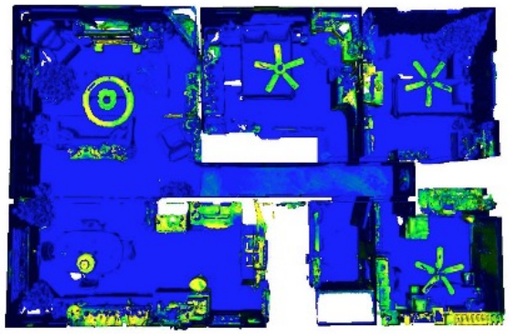
Open Vocabulary Perception

- Taking the point clouds and text instruction as the input, the model predicts the semantic mask

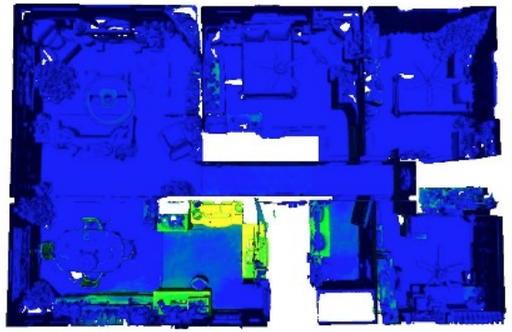
Input 3D Point Cloud



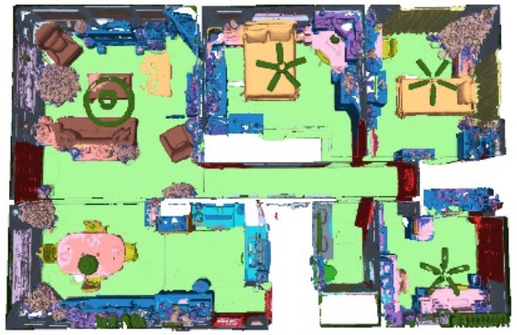
"fan" - Object



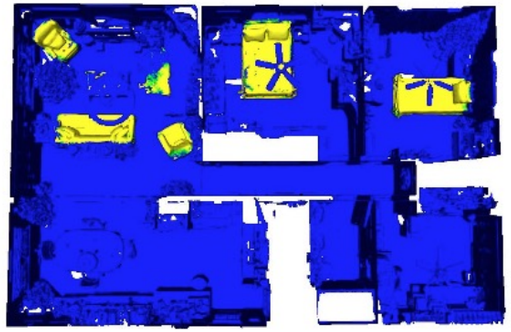
"metal" - Material



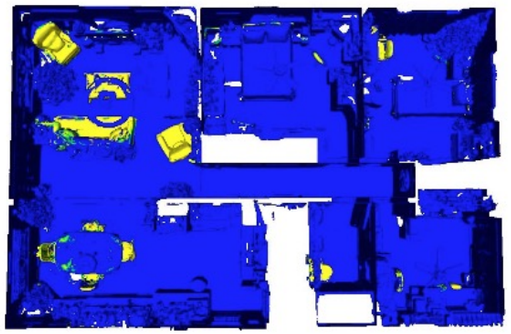
"kitchen" – Room Type



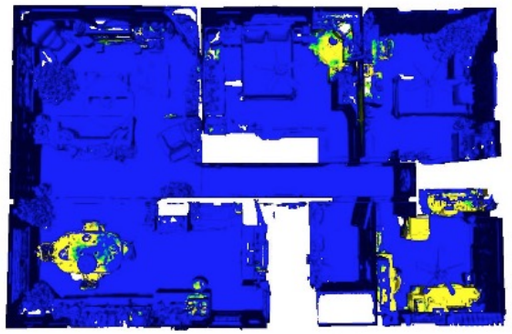
Zero-shot Semantic Segmentation



"soft" - Property



"sit" - Affordance



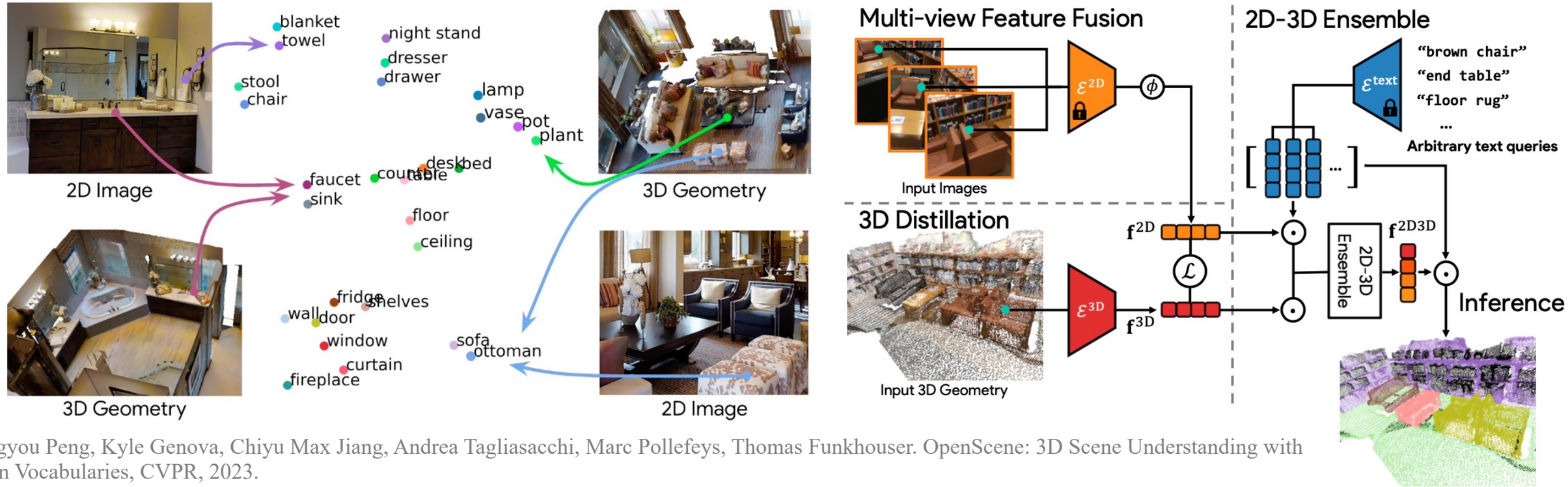
"work" - Activity

Songyou Peng, Kyle Genova, Chiyu Max Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser. OpenScene: 3D Scene Understanding with Open Vocabularies, CVPR, 2023.

Introduction

Literature Review: feature space alignment with VLM

- Global feature alignment: tend to overlook fine-grained 3D geometric details
- Point-wise feature alignment: prone to noise and outliers
- Patch-wise feature alignment: discontinuous correspondence in 3D

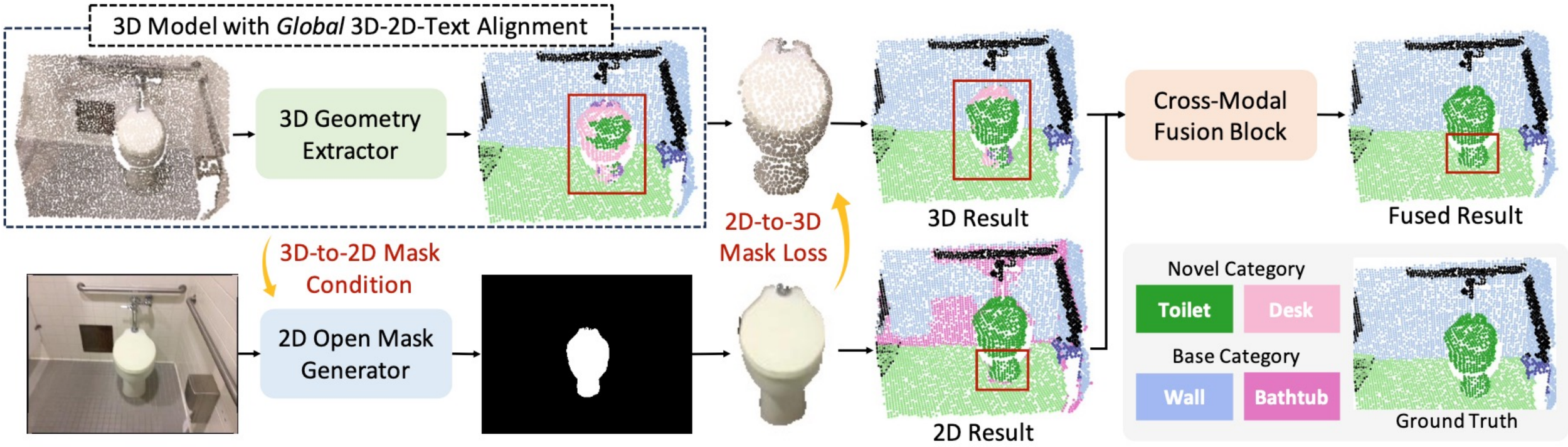


Songyou Peng, Kyle Genova, Chiyu Max Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser. OpenScene: 3D Scene Understanding with Open Vocabularies, CVPR, 2023.

Approaches

XMask3D: Cross-modal Mask Reasoning

- 3D-to-2D Mask Generation: generate geometry-aware 2D masks
- 2D-to-3D Mask Regularization: empower 3D feature with open-vocab capacity
- 3D-2D Mask Feature Fusion: enhance the synergy between multi-modal features

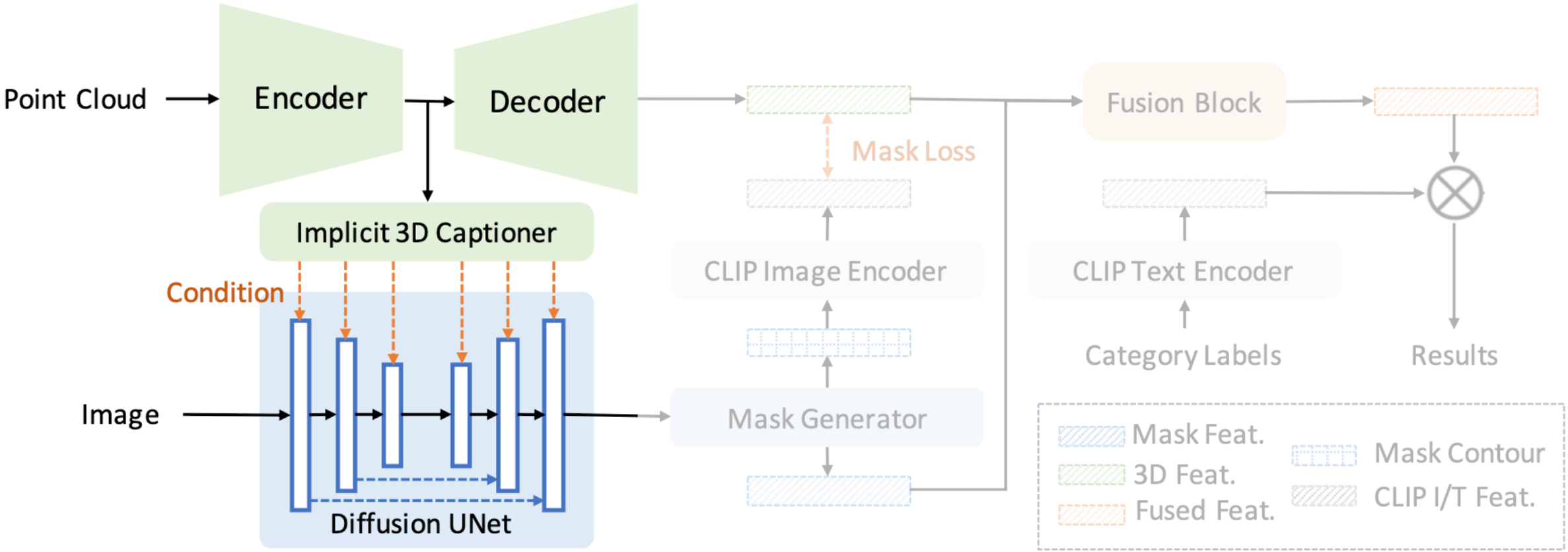


Approaches

XMask3D: Cross-modal Mask Reasoning

- 3D-to-2D Mask Generation: generate geometry-aware 2D masks

$$f = \mathrm{UNet}(x_t, \mathrm{MLP} \circ f_{3\mathrm{d}}), \quad f_{3\mathrm{d}} = \mathcal{E}(P)$$



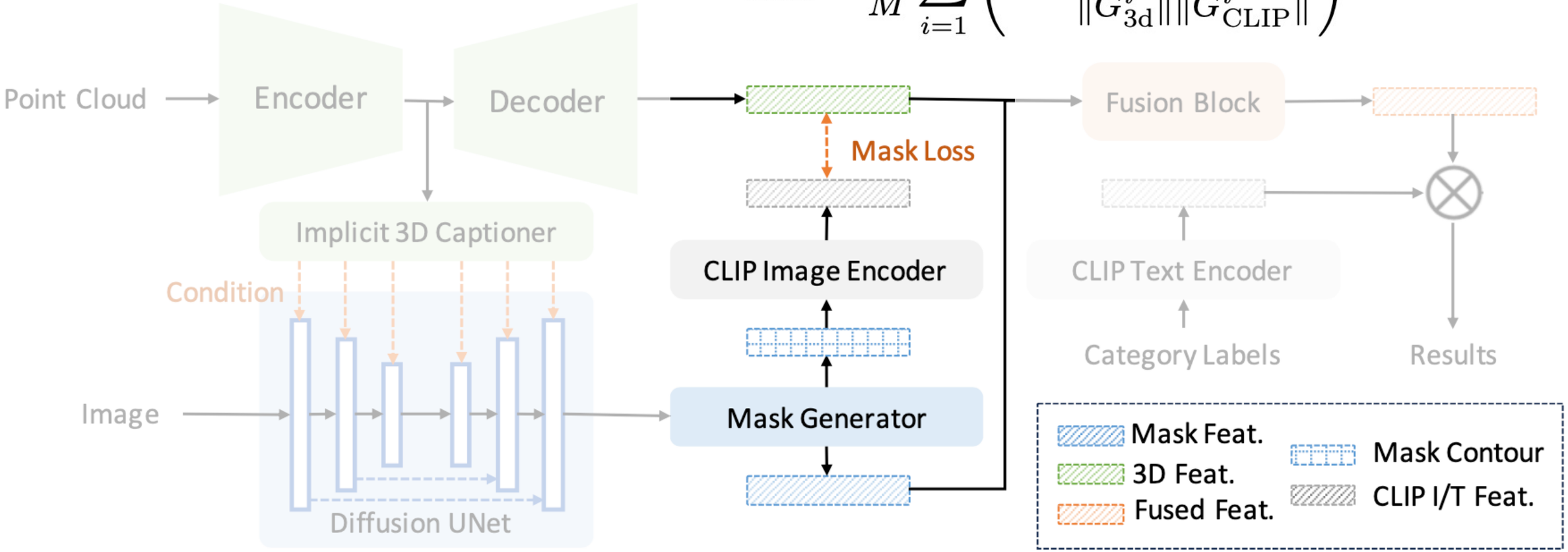
Approaches

XMask3D: Cross-modal Mask Reasoning

- 2D-to-3D Mask Regularization: empower 3D feature with open-vocab capacity

$$G^i_{3d} = \text{avgpool}(\tilde{F}^i_{3d})$$

$$\mathcal{L}_{\text{mask}} = \frac{1}{M}\sum_{i=1}^{M}\left(1 - \frac{G^i_{3d}\cdot(G^i_{\text{CLIP}})^T}{\|G^i_{3d}\|\|G^i_{\text{CLIP}}\|}\right)$$

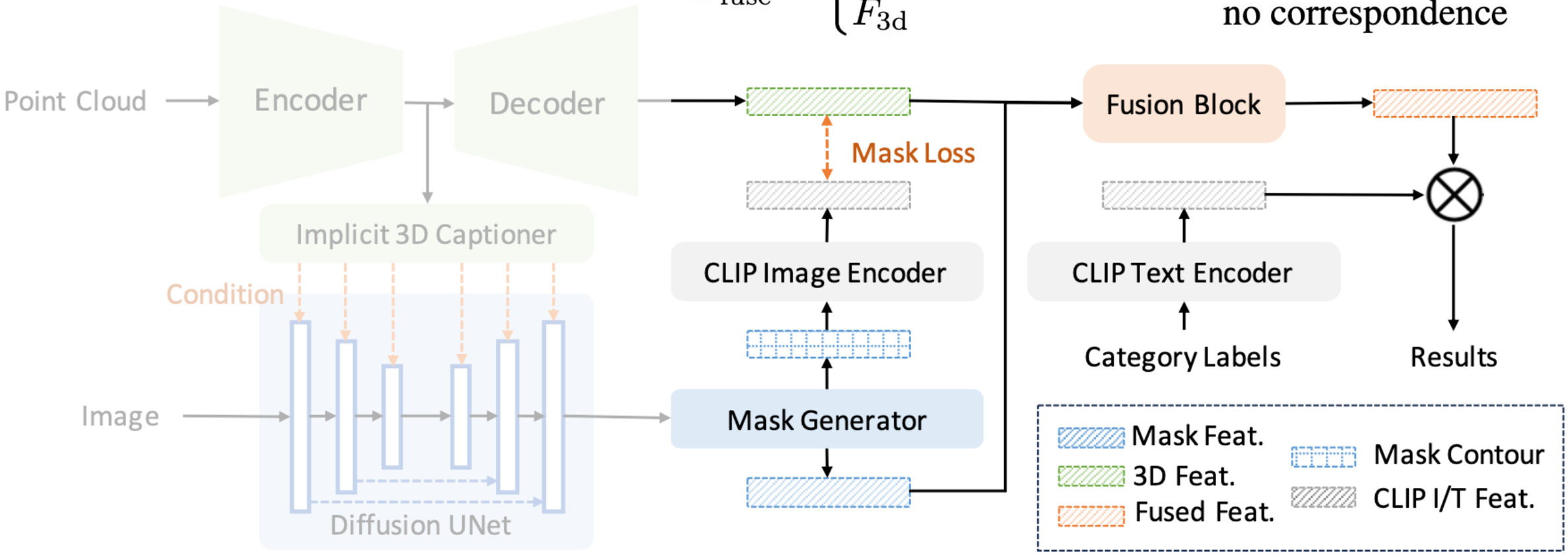


Approaches

XMask3D: Cross-modal Mask Reasoning

- 3D-2D Mask Feature Fusion enhance the synergy between multi-modal features

$$F_{\text{fuse}} = \begin{cases} \text{MLP} \circ \text{cat}(F_{3\text{d}}, F_{2\text{d}}) & \text{have correspondence} \\ F_{3\text{d}} & \text{no correspondence} \end{cases}$$



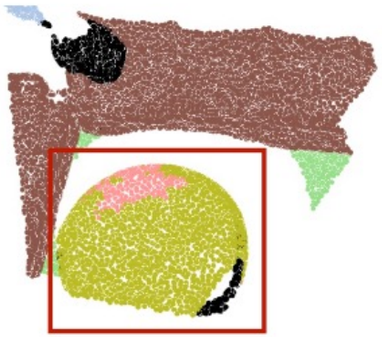
Experiments

Open Vocabulary 3D Semantic Segmentation

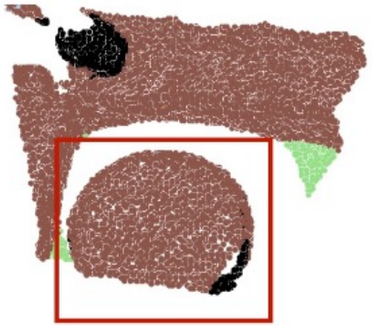
Method	Scannet									ScanNet200					
	B15/N4			B12/N7			B10/N9			B170/N30			B150/N50		
	hIoU	Base	Novel	hIoU	Base	Novel	hIoU	Base	Novel	hIoU	Base	Novel	hIoU	Base	Novel
LSeg-3D [23]	0.0	64.4	0.0	0.9	55.7	0.1	1.8	68.4	0.9	1.5	21.1	0.8	3.0	20.6	1.6
3DGenZ [31]	20.6	56.0	12.6	19.8	35.5	13.3	12.0	63.6	6.6	2.6	15.8	1.4	3.3	14.1	1.9
3DTZSL [5]	10.5	36.7	6.1	3.8	36.6	2.0	7.8	55.5	4.2	0.9	4.0	0.5	0.7	3.8	0.4
PLA [11]	65.3	68.3	62.4	55.3	69.5	45.9	53.1	76.2	40.8	11.4	20.9	7.8	10.1	20.9	6.6
OpenScene [34]	65.7	68.8	62.8	56.8	61.5	51.7	54.3	71.8	43.6	14.2	22.5	10.4	15.2	23.5	11.2
XMask3D	70.0	69.8	70.2	61.7	70.2	55.1	55.7	76.5	43.8	18.0	27.8	13.3	15.5	24.4	11.4

View Image

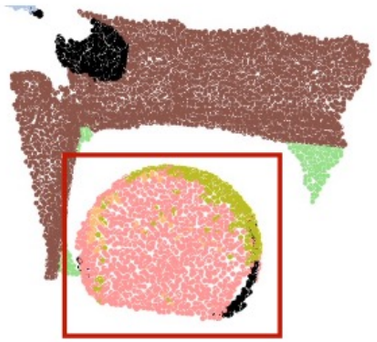
Ground Truth



PLA



OpenScene



XMask3D (Ours)

Experiments

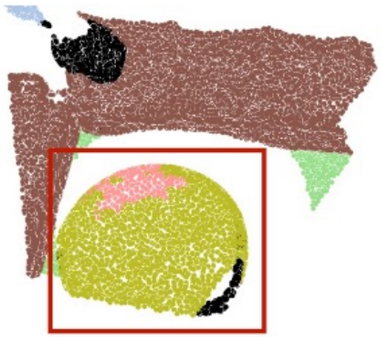
Open Vocabulary 3D Semantic Segmentation

Method	B15	
	hIoU	Ba
LSeg-3D [23]	0.0	64
3DGenZ [31]	20.6	56
3DTZSL [5]	10.5	36
PLA [11]	65.3	68
OpenScene [34]	65.7	68
XMask3D	70.0	69

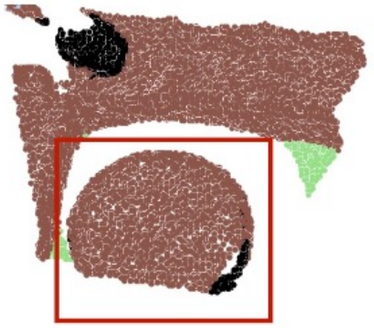
Method	S3DIS						ScanNet200			
	B8/N4			B6/N6			30	B150/N50		
	hIoU	Base	Novel	hIoU	Base	Novel	Novel	hIoU	Base	Novel
LSeg-3D [23]	0.1	49.0	0.1	0.0	30.1	0.0	0.8	3.0	20.6	1.6
3DTZSL [5]	8.4	43.1	4.7	3.5	28.2	1.9	1.4	3.3	14.1	1.9
3DGenZ [31]	8.8	50.3	4.8	9.4	20.3	6.1	0.5	0.7	3.8	0.4
PLA [11]	34.6	59.0	24.5	38.5	55.5	29.4	7.8	10.1	20.9	6.6
OpenScene [34]	42.4	58.6	33.2	44.2	**56.2**	36.4	10.4	15.2	23.5	11.2
XMask3D	**46.8**	**63.1**	**37.2**	**44.9**	52.8	**39.1**	13.3	15.5	24.4	11.4

View Image

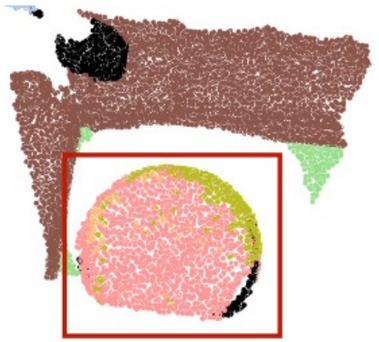
Ground Truth



PLA



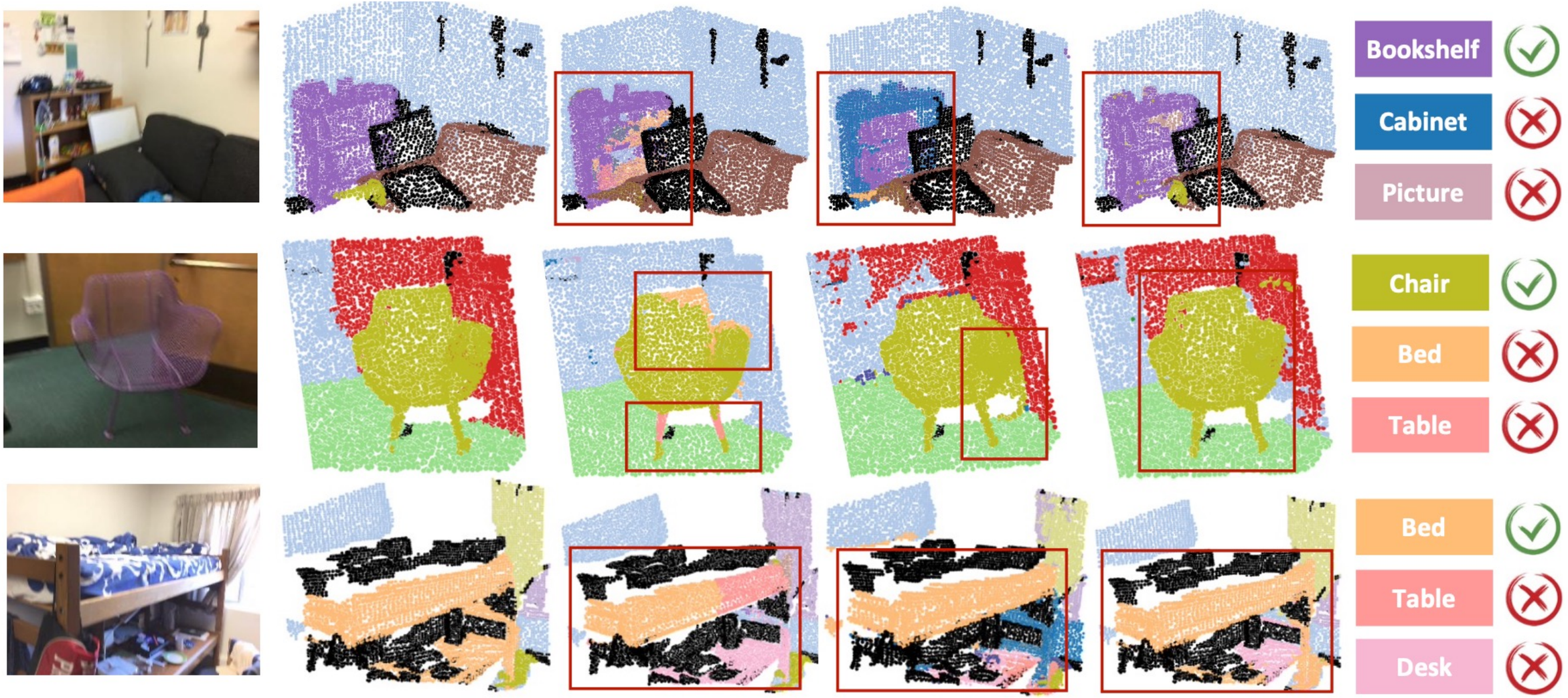
OpenScene



XMask3D (Ours)

Experiments

Visualization Results Comparison



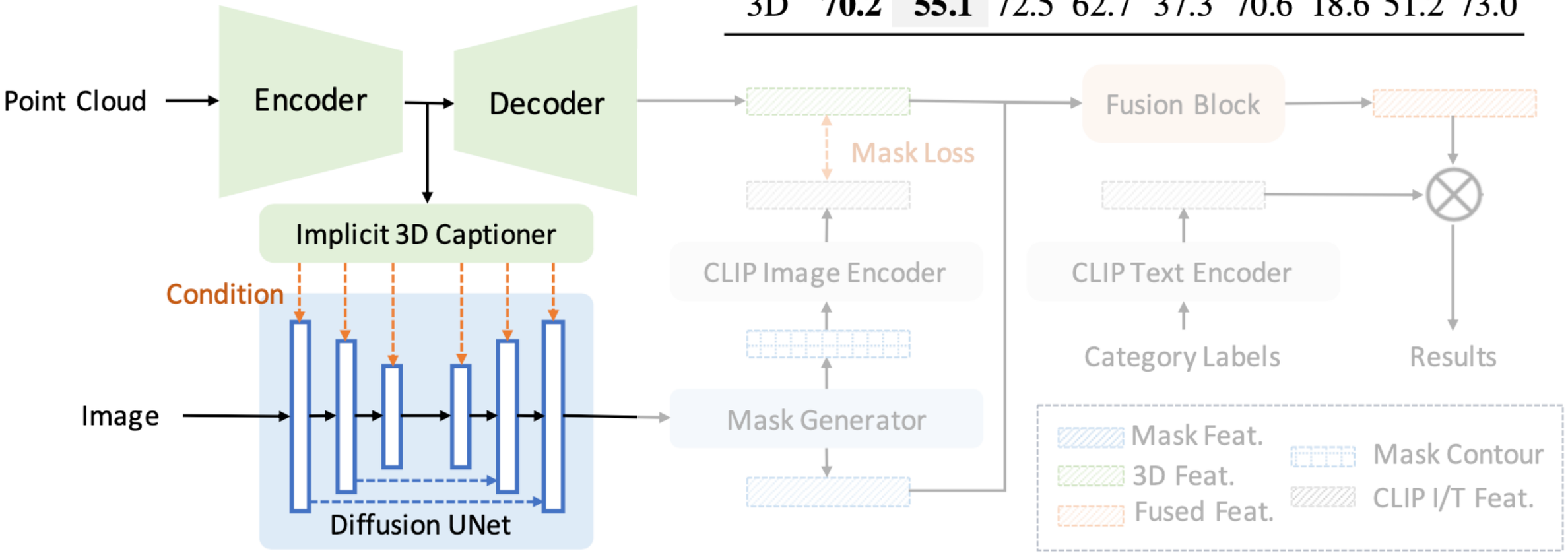
Experiments

Ablation Studies

- 3D-to-2D Mask Generation

(a) Ablation for implicit condition of the diffusion model.

Cond	Base	Novel	bed	chair	table	BKS	pic	sink	BT
Text	69.5	52.7	72.9	60.6	36.7	70.0	14.3	44.6	70.3
2D	69.7	53.6	70.6	63.3	40.9	68.4	12.4	51.6	67.8
3D	**70.2**	**55.1**	72.5	62.7	37.3	70.6	18.6	51.2	73.0



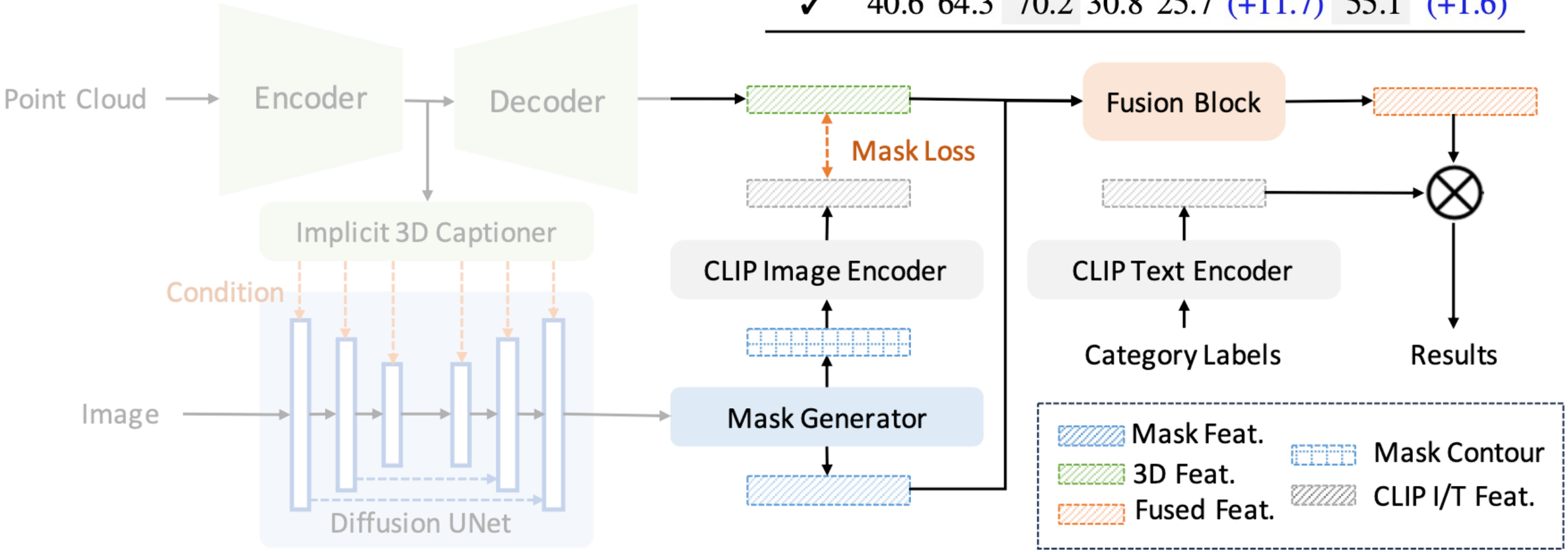
Experiments

Ablation Studies

- 2D-to-3D Mask Regularization
- 3D-2D Mask Feature Fusion

(b) Ablation for mask regularization and fusion block.

$\mathcal{L}_{\text{mask}}$	Base			Novel				
	2D	3D	Fuse	2D	3D	(Δ_{3D})	Fuse	(Δ_{Fuse})
✗	40.1	63.9	70.0	30.9	14.0		53.5	
✓	40.6	64.3	70.2	30.8	25.7	(+11.7)	55.1	(+1.6)



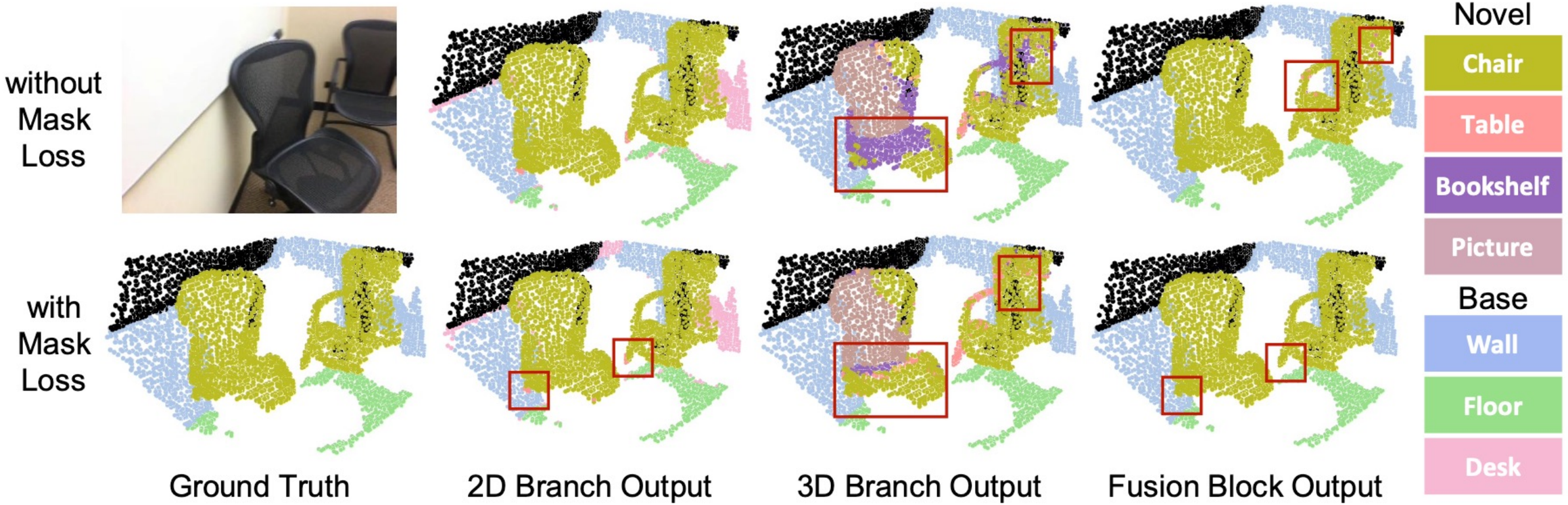
Experiments

Ablation Studies

- 2D-to-3D Mask Regularization
- 3D-2D Mask Feature Fusion

(b) Ablation for mask regularization and fusion block.

$\mathcal{L}_{\text{mask}}$	Base			Novel				
	2D	3D	Fuse	2D	3D	(Δ_{3D})	Fuse	(Δ_{Fuse})
✗	40.1	63.9	70.0	30.9	14.0		53.5	
✓	40.6	64.3	70.2	30.8	25.7	(+11.7)	55.1	(+1.6)



Conclusions

XMask3D

- We propose three cross-modal mask reasoning techniques for open vocabulary 3D semantic segmentation: 3D-to-2D mask generation, 2D-to-3D mask regularization, 3D-2D mask feature fusion
- We achieve competitive results on ScanNet20, ScanNet200, S3DIS benchmarks