Multimodality Helps Few-shot 3D Point Cloud Semantic Segmentation

Zhaochong An, Guolei Sun*, Yun Liu*, Runjia Li, Min Wu,
Ming-Ming Cheng, Ender Konukoglu, Serge Belongie

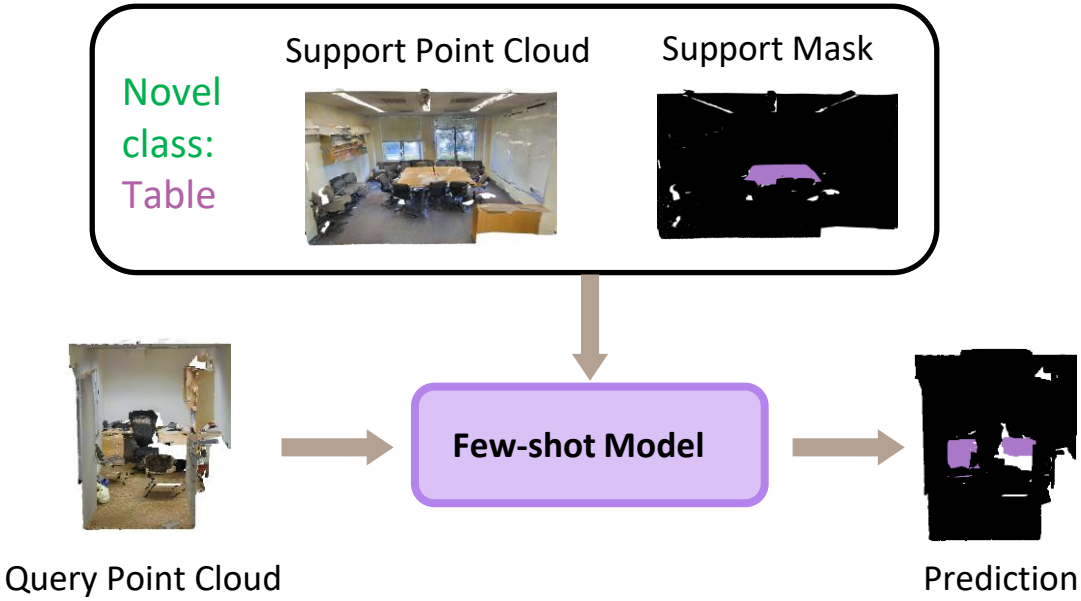
Department of Computer Science, University of Copenhagen
College of Computer Science, Nankai University
Institute for Infocomm Research, A*STAR

Computer Vision Laboratory, ETH Zurich
Department of Engineering Science, University of Oxford

Spotlight Paper [Top 5%]
ICLR 2025

Setting of interest

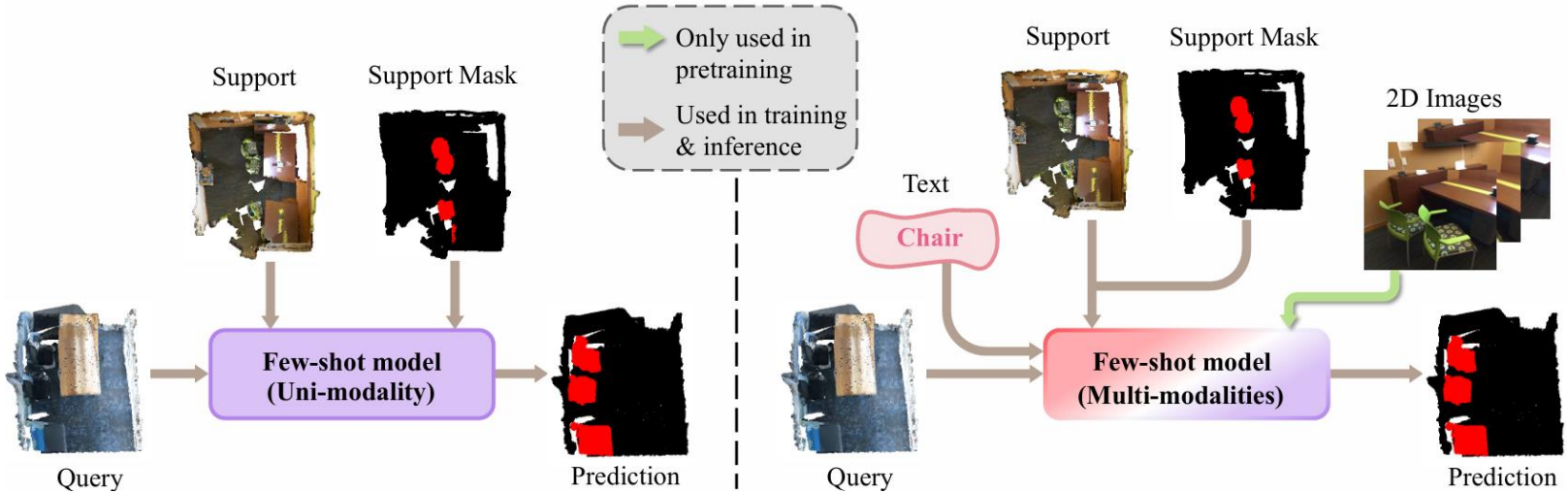
- Few-shot 3D Point Cloud Semantic Segmentation (FS-PCS)



- Segment new, unseen categories
- Step beyond the limited semantic space of fully supervised
- Perform better on target classes compared to zero-shot

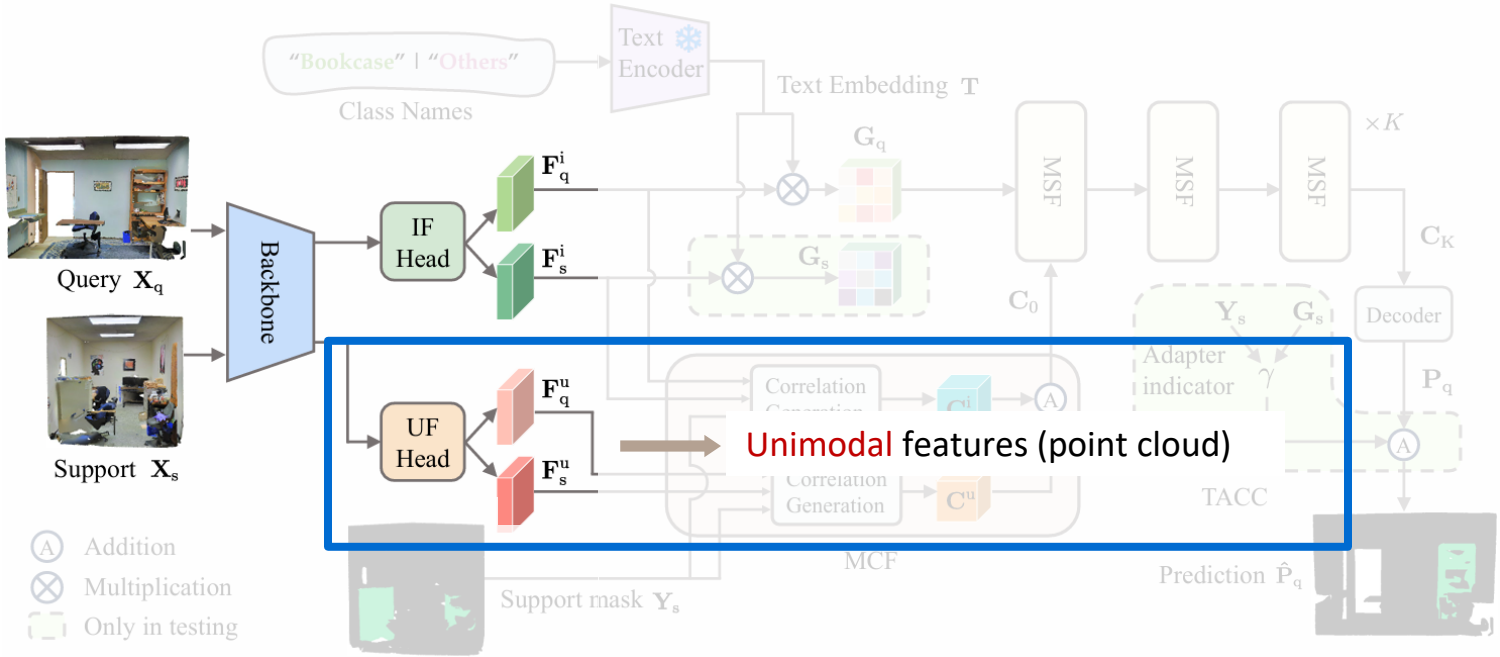
Motivation

- Previous model only focused on unimodal point cloud inputs
- Introduce a cost-free multimodal FS-PCS setup
    - Text and 2D images
    - 2D only used during pretraining



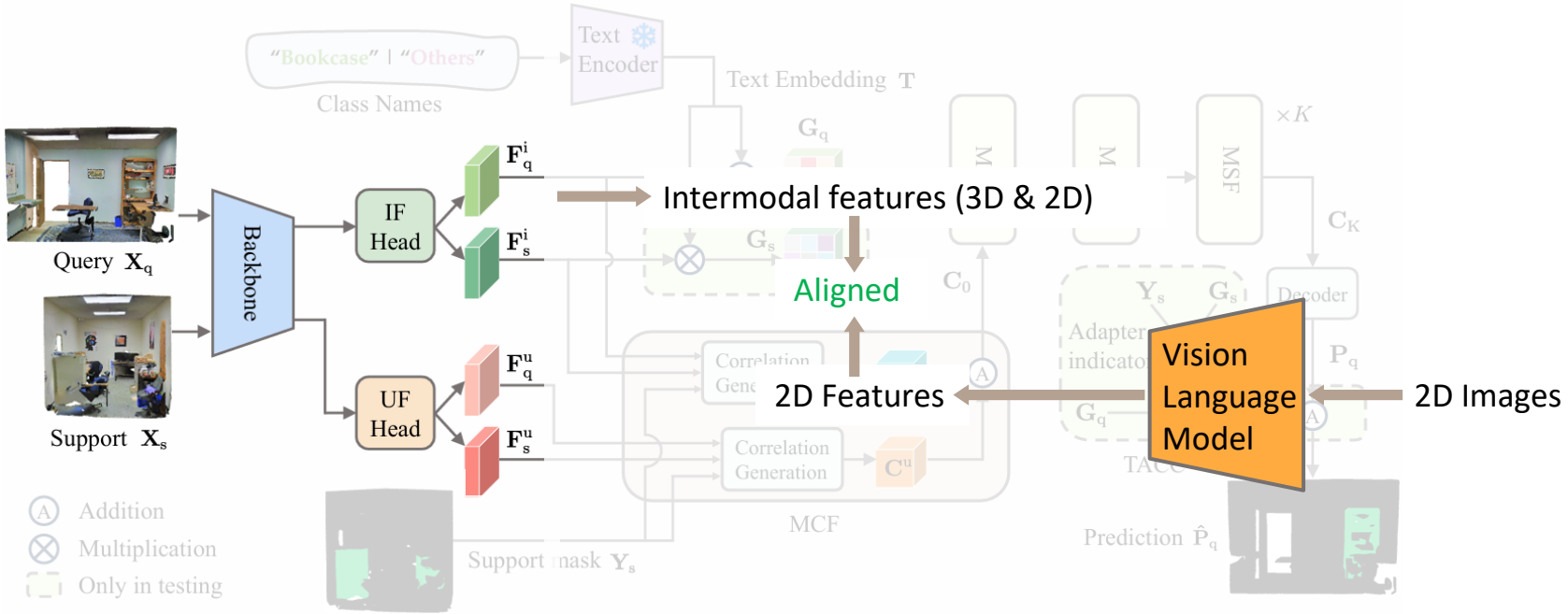
Method: MultiModal Few-Shot SegNet (MM-FSS)

- Process 3D inputs by a shared 3D backbone with two heads
  - Unimodal head



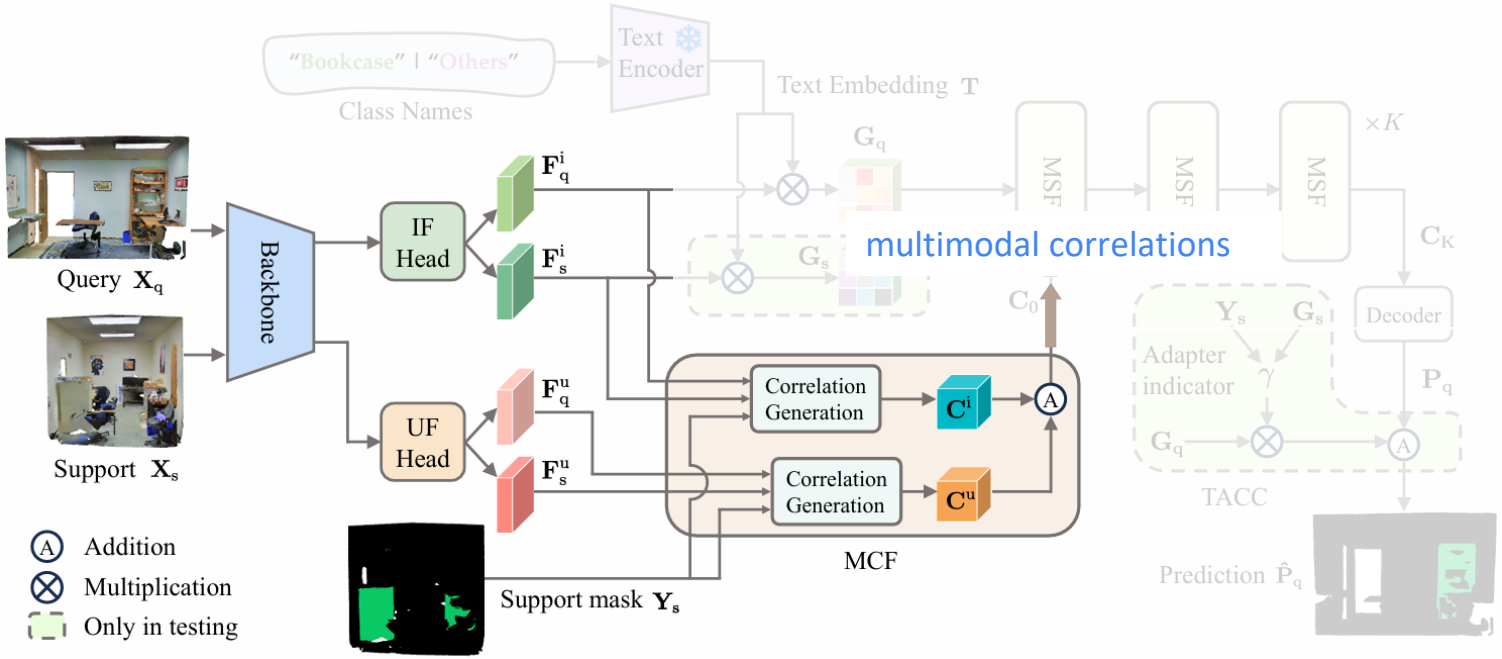
Method: MultiModal Few-Shot SegNet (MM-FSS)

- Process 3D inputs by a shared 3D backbone with two heads
  - Unimodal head
  - Intermodal head



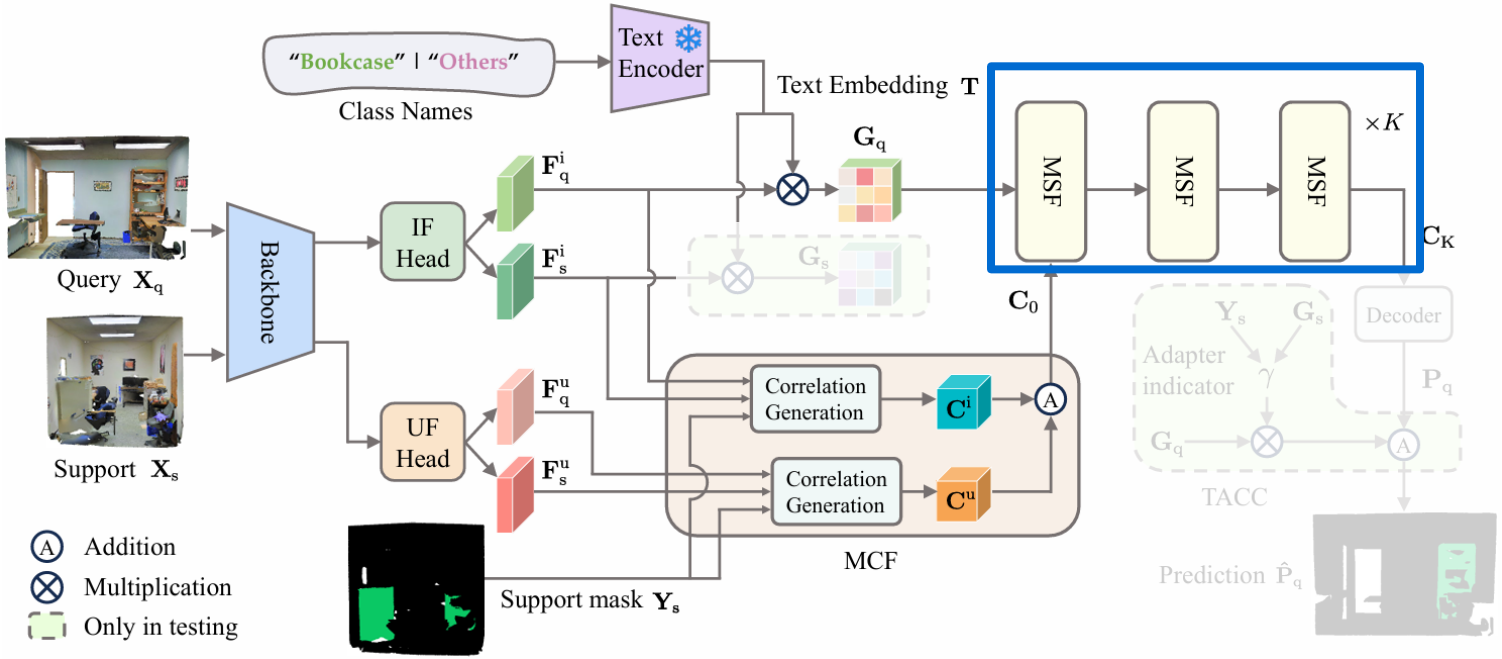
Method: Multimodal Correlation Fusion (MCF)

- Two correlations between query and support using intermodal and unimodal features
- Fused in MCF to generate comprehensive multimodal correlations



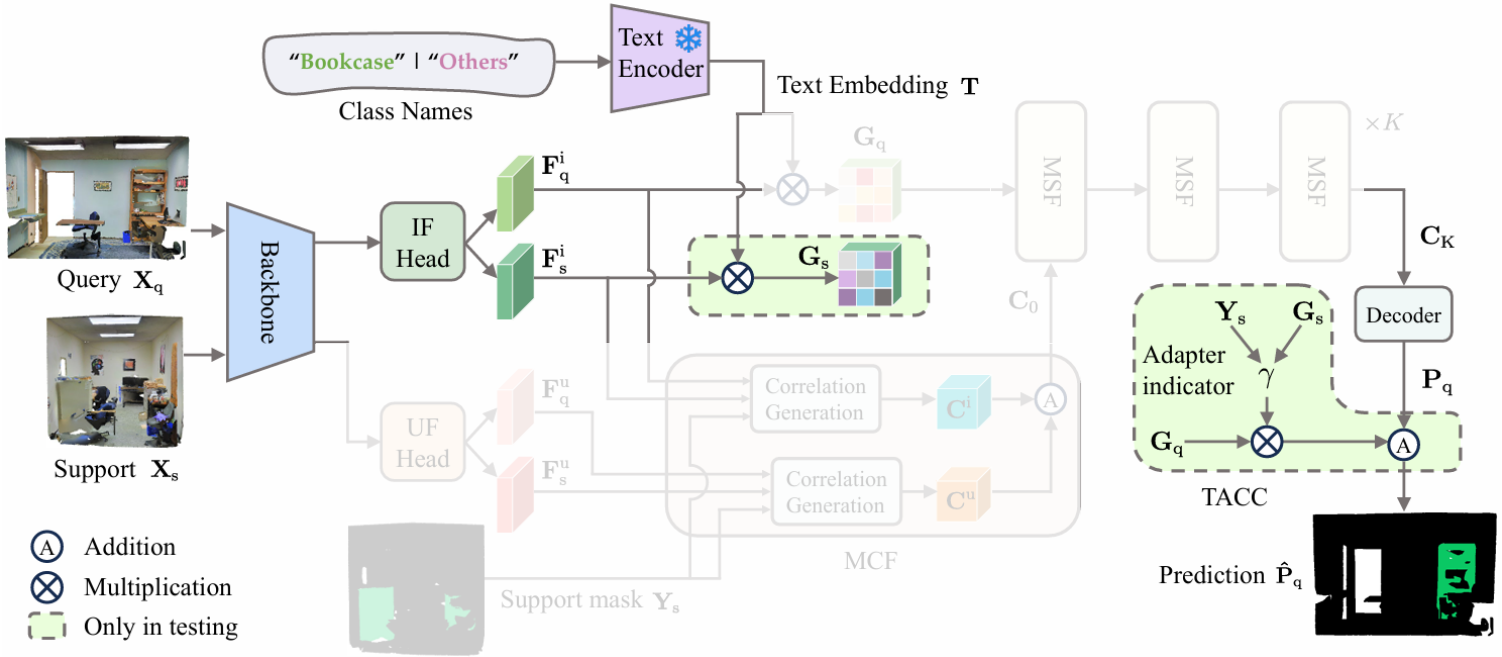
Method: Multimodal Semantic Fusion (MSF)

- Refine the multimodal correlations using text modality
- Adjust the weight of textual guidance considering the varying relative importance of visual and textual modalities



Method: Test-time Adaptive Cross-modal Calibration (TACC)

- Use less-biased textual guidance to calibrate predictions during testing, through an adaptive indicator γ, to mitigate the base bias issue
- γ estimates the quality of textual guidance based on support annotations



Experiments: Main results

Methods	1-way 1-shot			1-way 5-shot			2-way 1-shot			2-way 5-shot		
	S^0	S^1	Mean	S^0	S^1	Mean	S^0	S^1	Mean	S^0	S^1	Mean
AttMPTI (Zhao et al., 2021)	36.32	38.36	37.34	46.71	42.70	44.71	31.09	29.62	30.36	39.53	32.62	36.08
QGE (Ning et al., 2023)	41.69	39.09	40.39	50.59	46.41	48.50	33.45	30.95	32.20	40.53	36.13	38.33
QGPA (He et al., 2023)	35.50	35.83	35.67	38.07	39.70	38.89	25.52	26.26	25.89	30.22	32.41	31.32
COSeg (An et al., 2024)	46.31	48.10	47.21	51.40	48.68	50.04	37.44	36.45	36.95	42.27	38.45	40.36
COSeg† (An et al., 2024)	47.17	48.37	47.77	50.93	49.88	50.41	37.15	38.99	38.07	42.73	40.25	41.49
MM-FSS (ours)	**49.84**	**54.33**	**52.09**(+4.3)	**51.95**	**56.46**	**54.21**(+3.8)	**41.98**	**46.61**	**44.30**(+6.2)	**46.02**	**54.29**	**50.16**(+8.7)

Table 1: **Quantitative comparison with previous methods in mIoU (%) on the S3DIS dataset.** There are four few-shot settings: $1/2$-way $1/5$-shot. S^0/S^1 refers to using the split i for evaluation, and 'Mean' represents the average mIoU on both splits. The best results are highlighted in **bold**.

Methods	1-way 1-shot			1-way 5-shot			2-way 1-shot			2-way 5-shot		
	S^0	S^1	mean	S^0	S^1	Mean	S^0	S^1	Mean	S^0	S^1	Mean
AttMPTI (Zhao et al., 2021)	34.03	30.97	32.50	39.09	37.15	38.12	25.99	23.88	24.94	30.41	27.35	28.88
QGE (Ning et al., 2023)	37.38	33.02	35.20	45.08	41.89	43.49	26.85	25.17	26.01	28.35	31.49	29.92
QGPA (He et al., 2023)	34.57	33.37	33.97	41.22	38.65	39.94	21.86	21.47	21.67	30.67	27.69	29.18
COSeg (An et al., 2024)	41.73	41.82	41.78	48.31	44.11	46.21	28.72	28.83	28.78	35.97	33.39	34.68
COSeg† (An et al., 2024)	41.95	42.07	42.01	48.54	44.68	46.61	29.54	28.51	29.03	36.87	34.15	35.51
MM-FSS (ours)	**46.08**	**43.37**	**44.73**(+2.7)	**54.66**	**45.48**	**50.07**(+3.5)	**43.99**	**34.43**	**39.21**(+10.2)	**48.86**	**39.32**	**44.09**(+8.6)

Table 2: **Quantitative comparison with previous methods in mIoU (%) on the ScanNet dataset.**

Experiments: Visualizations

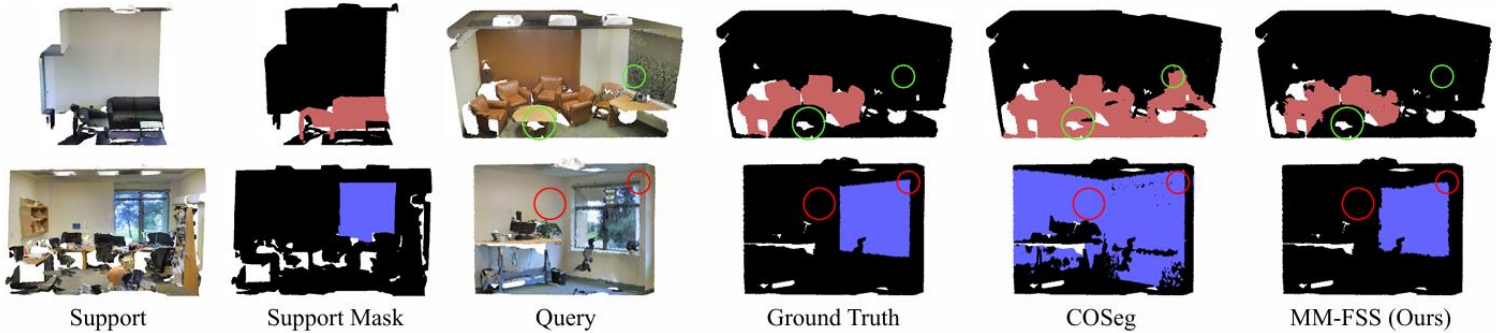


Figure 3: **Qualitative comparison between COSeg and our proposed MM-FSS in the 1-way 1-shot setting on the S3DIS dataset.** The target classes in the first and second rows are sofa and window, respectively. Important areas are marked with circles.

Experiments: Visualizations

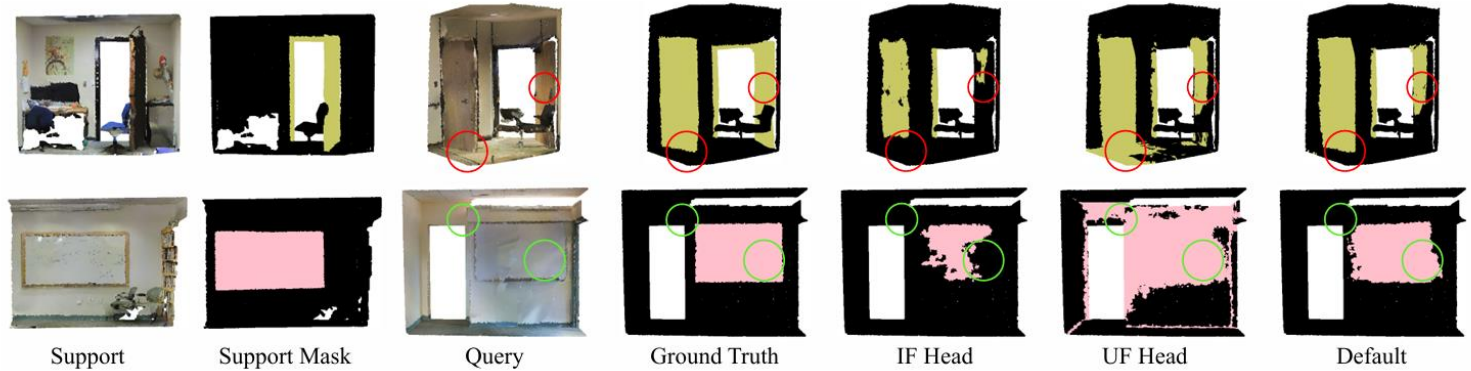


Figure 4: **Qualitative comparison of predictions from each head and our final prediction using TACC (Default) in the 1-way 1-shot setting on the S3DIS dataset.** The target classes in the first and second rows are door and board, respectively.

Experiments: Visualizations

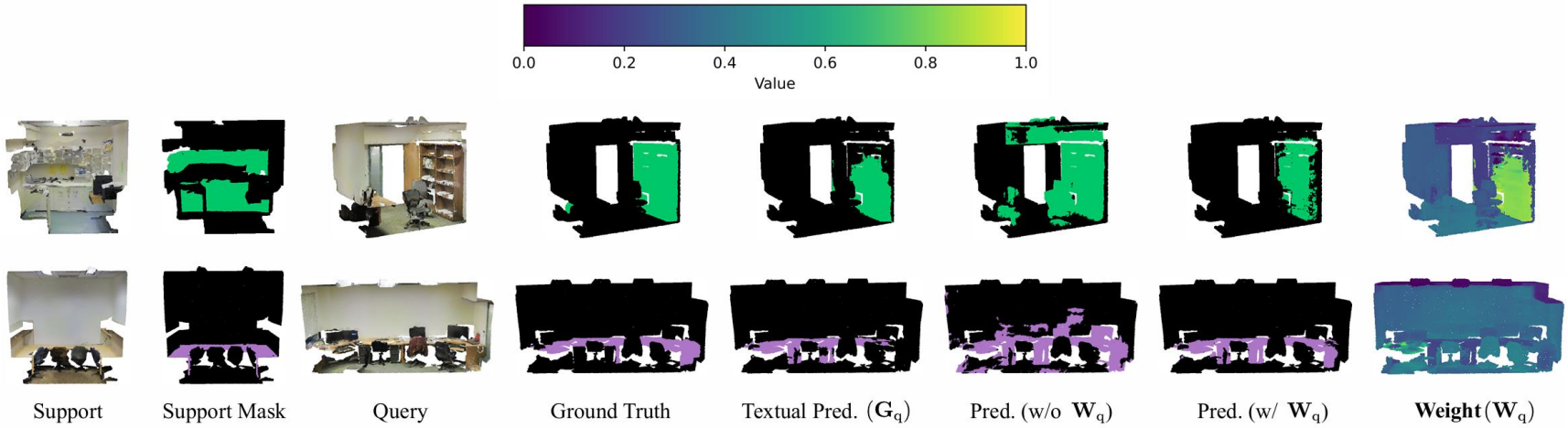


Figure 5: Visualization on the effects of weight $\mathbf{W}_q$ between textual and visual modalities in Eq. (7). The last column displays the heatmap of $\mathbf{W}_q$ with the color bar referenced at the top. Higher values indicate larger weights assigned to textual guidance $\mathbf{G}_q$. Each row represents the 1-way 1-shot setting on the S3DIS dataset targeting bookcase and table, respectively, arranged from top to bottom.

In a nutshell

- The first to propose a novel ***cost-free multimodal few-shot*** 3D segmentation setup

- Under our multimodal setup, we present MM-FSS to effectively exploit information from different modalities, achieving ***state-of-the-art results*** across all FS-PCS settings.

- Offers valuable insights into the ***importance of commonly ignored free modalities*** in few-shot learning and paves the way for future advances.

Thank you!

OUR PAPER

OUR CODE