Vote2Cap-DETR: A Set-to-Set Perspective Towards 3D Dense Captioning

Winner Presentation of the Scan2Cap Challenge

Sijin Chen[1] Hongyuan Zhu[2] Xin Chen[3] Yinjie Lei[4] Gang YU[3] Tao Chen[1†] Taihao Li[5]

[1] Fudan University [2] Institute for Infocomm Research (I2R) & Centre for Frontier AI Research (CFAR), A*STAR, Singapore
[3] Tencent PCG [4] Sichuan University [5] Zhejiang Lab

[†] Corresponding author.

Paper

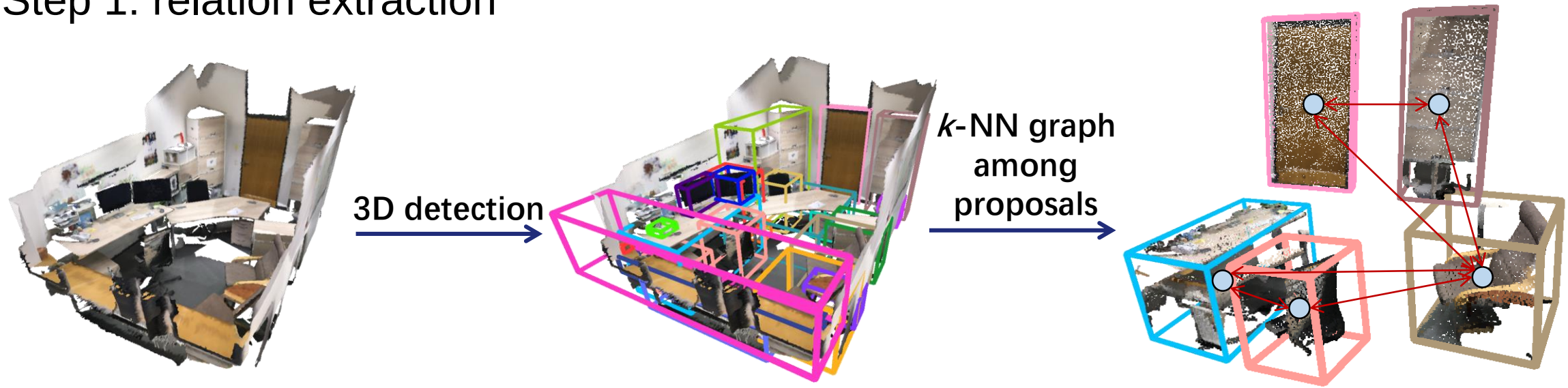
Project

Task Definition: 3D Dense Captioning

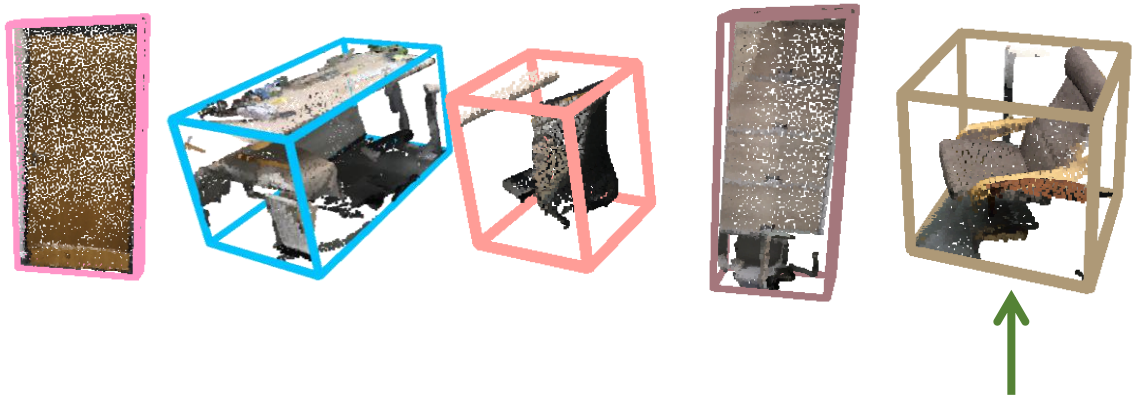
1. **accurate** localization of objects from a cluttered 3D scene;
2. **informative** and **object-centric** descriptions for each instance.

Previous Explicit Approaches

Step 1: relation extraction



Step 2: IoU based proposal selection



"This is a chair with wooden frame and brown cushion. It is next to a side table and another same chair."

Limitations

1. cumulative error caused by duplicated and inaccurate box proposals

How many monitors indeed?

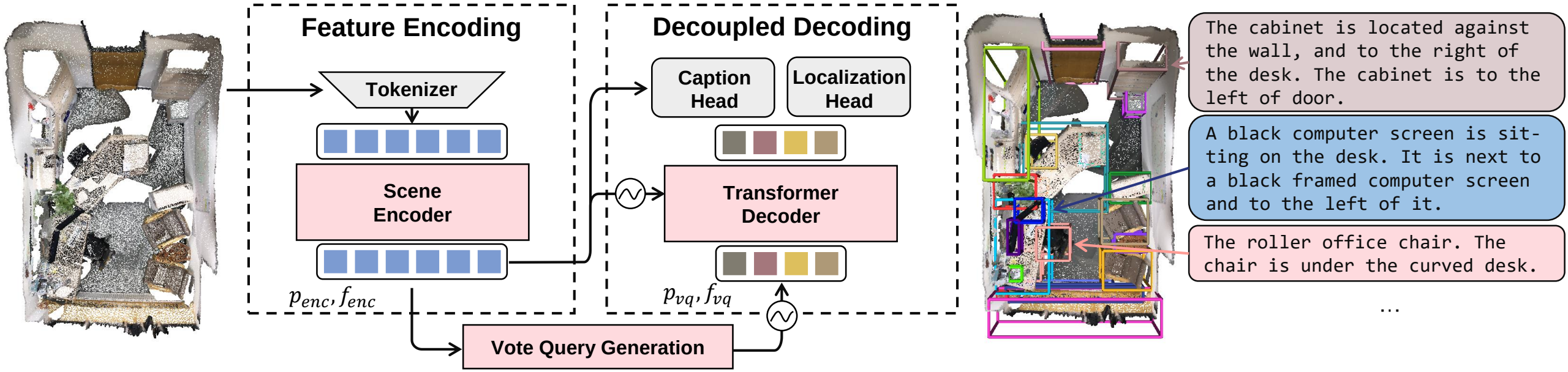
Limitations

2. hyper parameters hard to generalize to <u>diverse</u> 3D scenes

How to construct generalizable *k*-NN Graph?

Vote2Cap-DETR: A Set-to-Set Approach

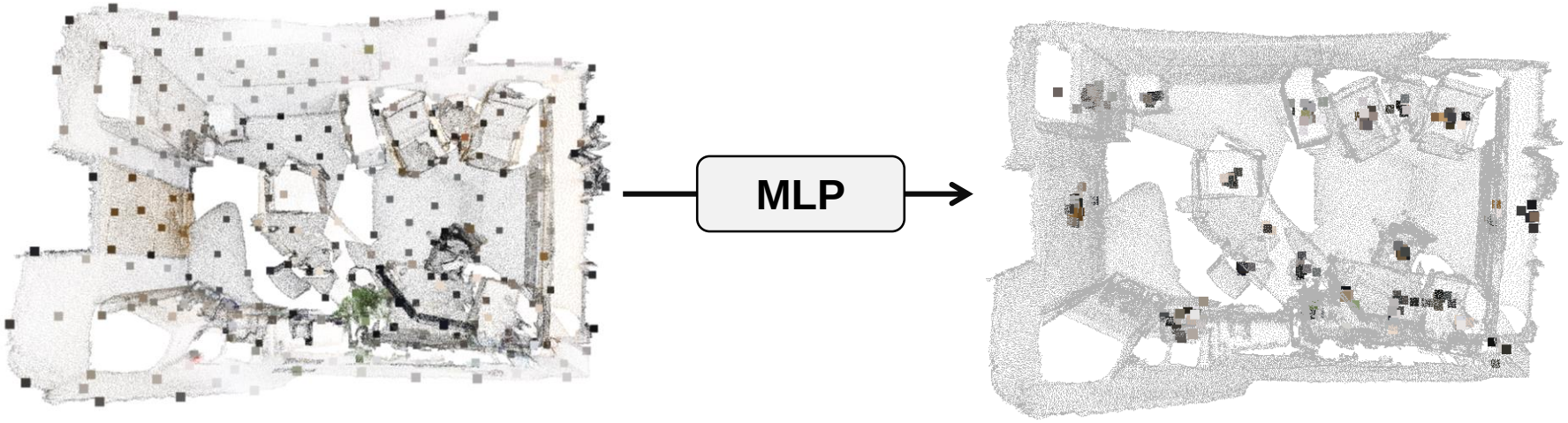
1. translate from a set of points to a set of “box-caption” proposals
2. learn query-to-query, query-to-scene interaction with decoder attention
3. set-to-set training, learning discriminative feature representations



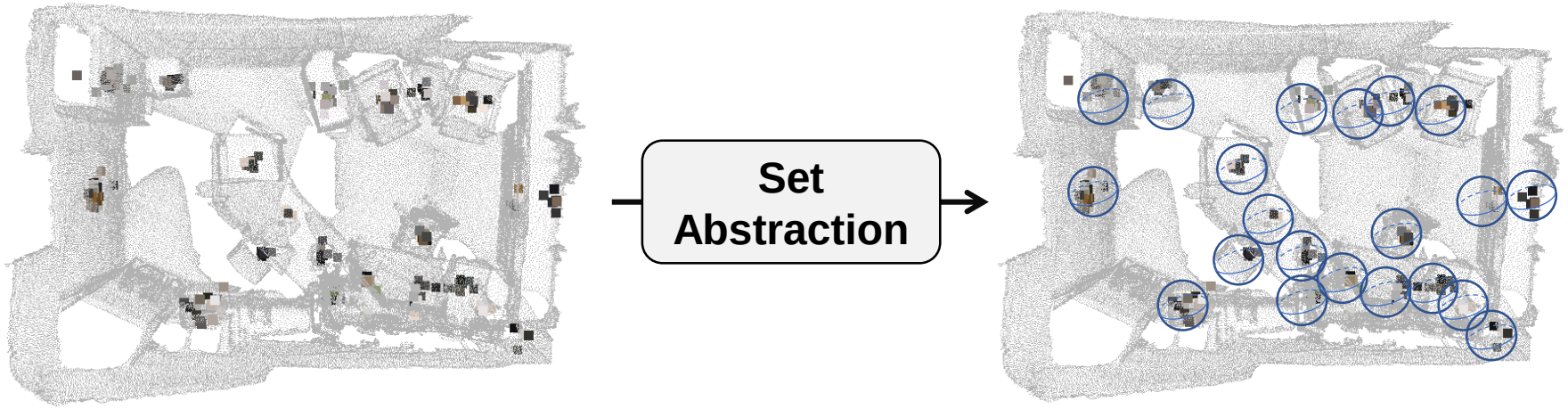
Method: Component Designs

> design of the vote query

Step 1: vote shifting

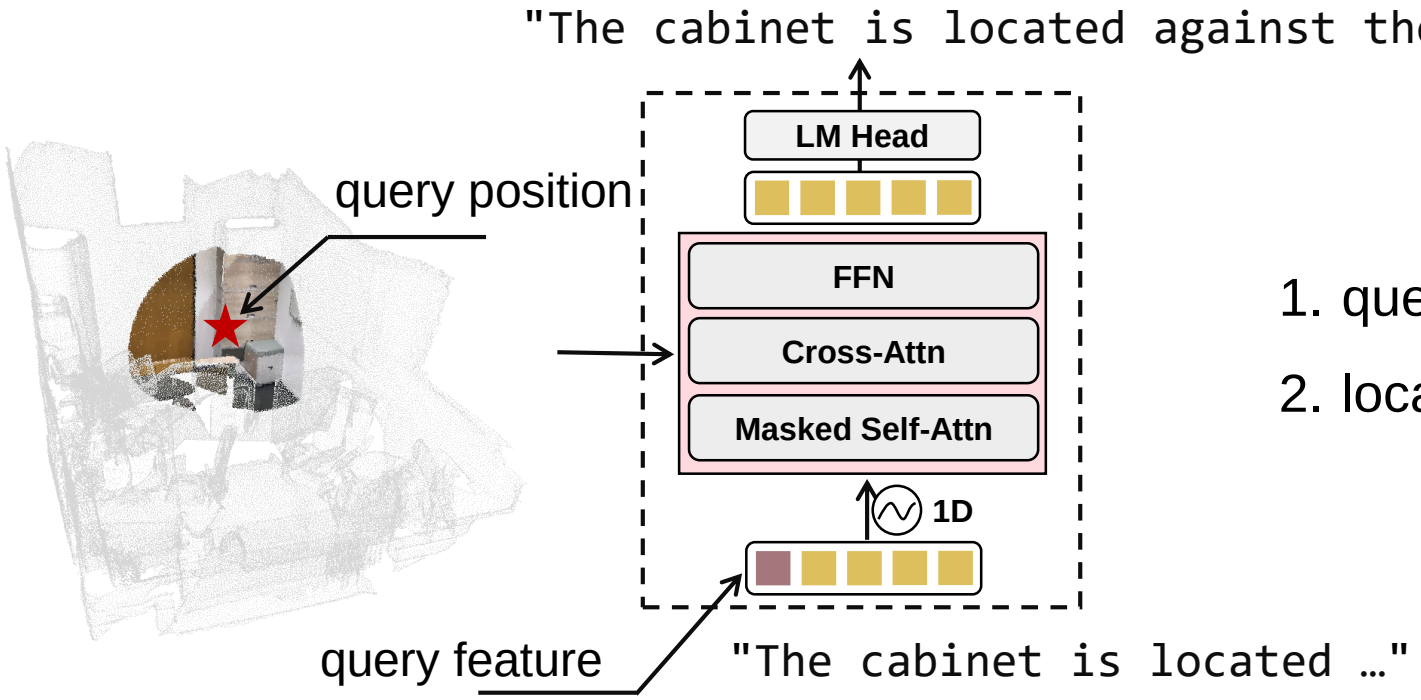


Step 2: local feature aggregation



Method: Component Designs

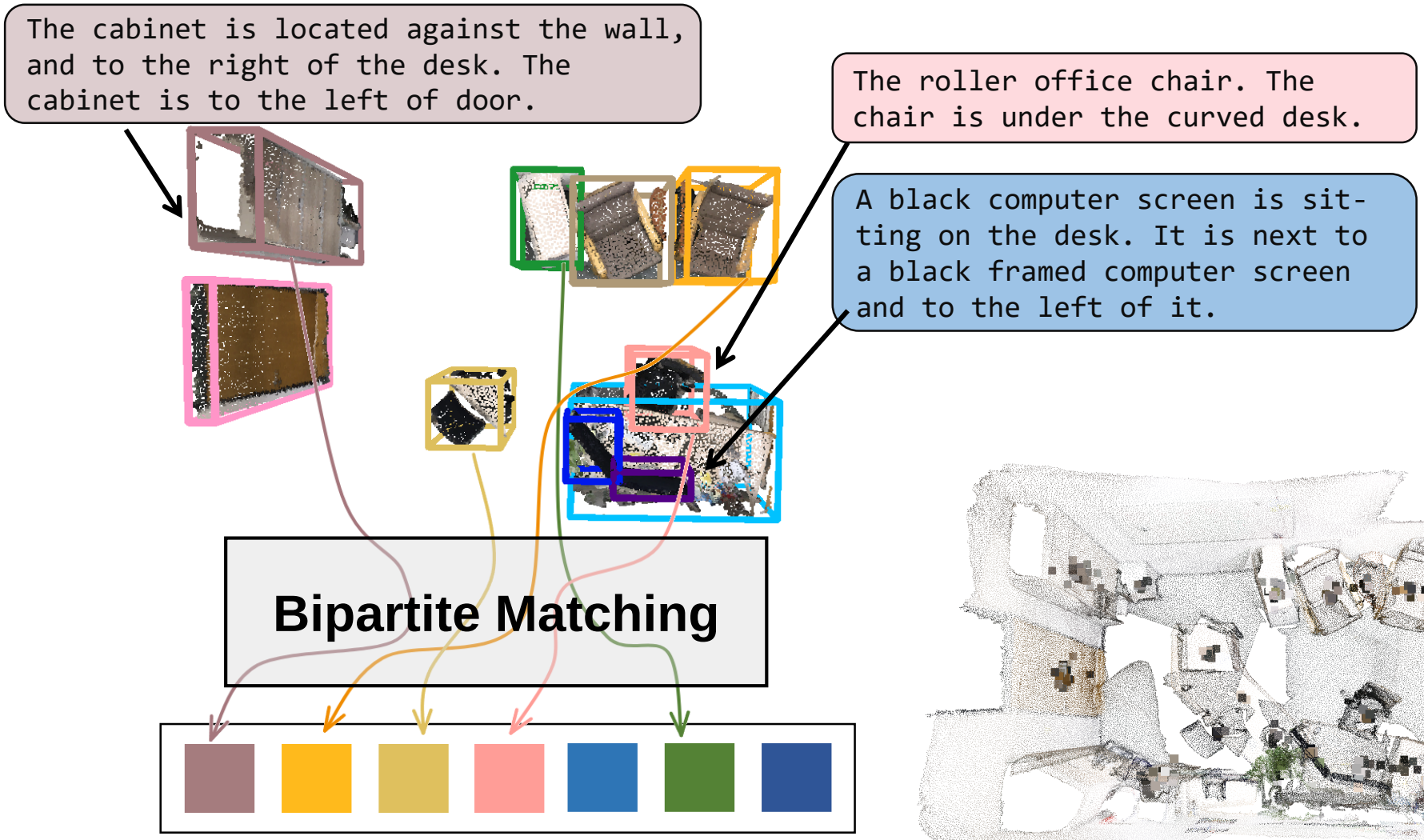
> design of the caption head



1. query feature as the prefix identifier;
2. local context surroundings as guidance.

Method: Set-to-Set Training

> matching queries to a set of “object-caption” pairs



Quantitative Results

> ScanRefer validation set

Method	$\mathcal{L}_{des}$	w/o additional 2D input								w/ additional 2D input							
		IoU = 0.25				IoU = 0.50				IoU = 0.25				IoU = 0.50			
		C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑
Scan2Cap [13]	MLE	53.73	34.25	26.14	54.95	35.20	22.36	21.44	43.57	56.82	34.18	26.29	55.27	39.08	23.32	21.97	44.78
MORE [20]		58.89	35.41	26.36	55.41	38.98	23.01	21.65	44.33	62.91	36.25	26.75	56.33	40.94	22.93	21.66	44.42
SpaCap3d [39]		58.06	35.30	26.16	55.03	42.76	25.38	22.84	45.66	63.30	36.46	26.71	55.71	44.02	25.26	22.33	45.36
3DJCG [4]		60.86	**39.67**	27.45	59.02	47.68	31.53	24.28	51.80	64.70	**40.17**	27.66	**59.23**	49.48	31.03	24.22	50.80
D3Net [7]		-	-	-	-	-	-	-	-	-	-	-	-	46.07	30.29	24.35	51.67
Ours		**71.45**	39.34	**28.25**	**59.33**	**61.81**	**34.46**	**26.22**	**54.40**	**72.79**	39.17	**28.06**	**59.23**	**59.32**	**32.42**	**25.28**	**52.53**
χ-Trans2Cap [43]	SCST	58.81	34.17	25.81	54.10	41.52	23.83	21.90	44.97	61.83	35.65	26.61	54.70	43.87	25.05	22.46	45.28
Scan2Cap [13]		-	-	-	-	-	-	-	-	-	-	-	-	48.38	26.09	22.15	44.74
D3Net [7]		-	-	-	-	-	-	-	-	-	-	-	-	62.64	35.68	**25.72**	**53.90**
Ours		**84.15**	**42.51**	**28.47**	**59.26**	**73.77**	**38.21**	**26.64**	**54.71**	**86.28**	**42.64**	**28.27**	**59.07**	**70.63**	**35.69**	25.51	52.28

> Nr3D validation set

Method	$\mathcal{L}_{des}$	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
Scan2Cap [13]	MLE	27.47	17.24	21.80	49.06
SpaCap3d [39]		33.71	19.92	22.61	50.50
D3Net [7]		33.85	20.70	23.13	53.38
3DJCG [4]		38.06	22.82	23.77	52.99
Ours		**43.84**	**26.68**	**25.41**	**54.43**
χ-Tran2Cap [43]	SCST	33.62	19.29	22.27	50.00
D3Net [7]		38.42	22.22	24.74	54.37
Ours		**45.53**	**26.88**	**25.43**	**54.76**

Qualitative Results

scene0011_00

3DJCG: This is a rectangular whiteboard. It is on the wall.

SpaCap3D: The whiteboard is affixed to the wall. It is to the right of the window.

Ours: The tv is on the wall. It is to the right of the table.

GT: This is a big black tv. It is above a thin table.

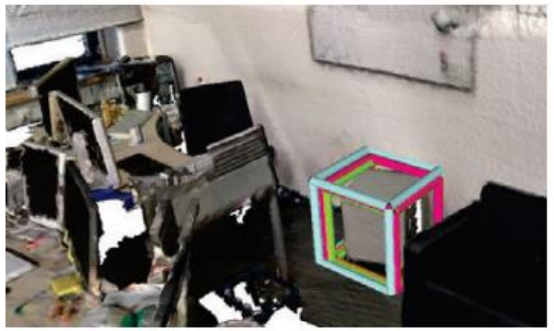
scene0015_00

3DJCG: This is a brown table. It is in the middle of the room.

SpaCap3D: This is a wooden table. It is in the center of the room.

Ours: This is a wooden table. It is in the corner of the room.

GT: This is a small table with a wood look. It is the table closest to the front of the room in the upper left corner.



scene0025_00

3DJCG: The is a small brown cabinet. It is to the right of the desk.

SpaCap3D: The cabinet is below the desk. It is to the left of the chair.

Ours: This is a white cabinet. It is to the right of the table.

GT: A white cabinet is sitting on the floor next to the wall. It is to the left of the couch.

scene0050_00

3DJCG: This is a brown table. It is in front of the couch.

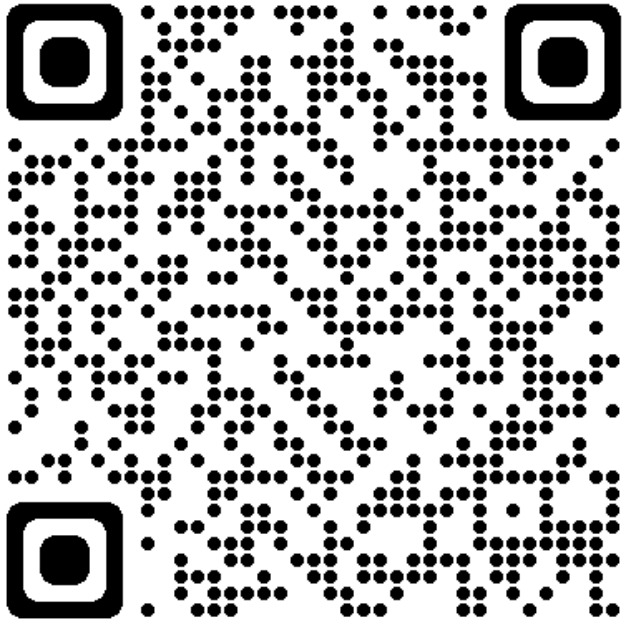
SpaCap3D: This is a wooden coffee table. It is in front of the couch.

Ours: This is a brown ottoman. It is to the right of the chair.

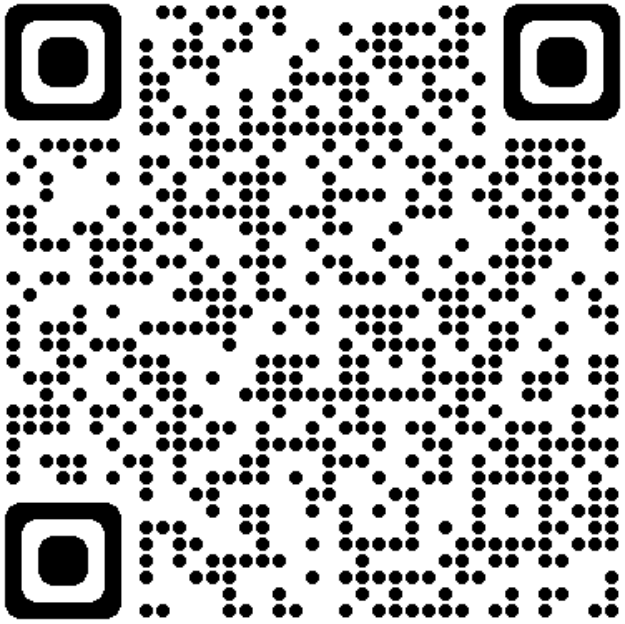
GT: This is a brown ottoman. It is in front of a couch.

One More Thing…

Vote2Cap-DETR++



Github



Weights