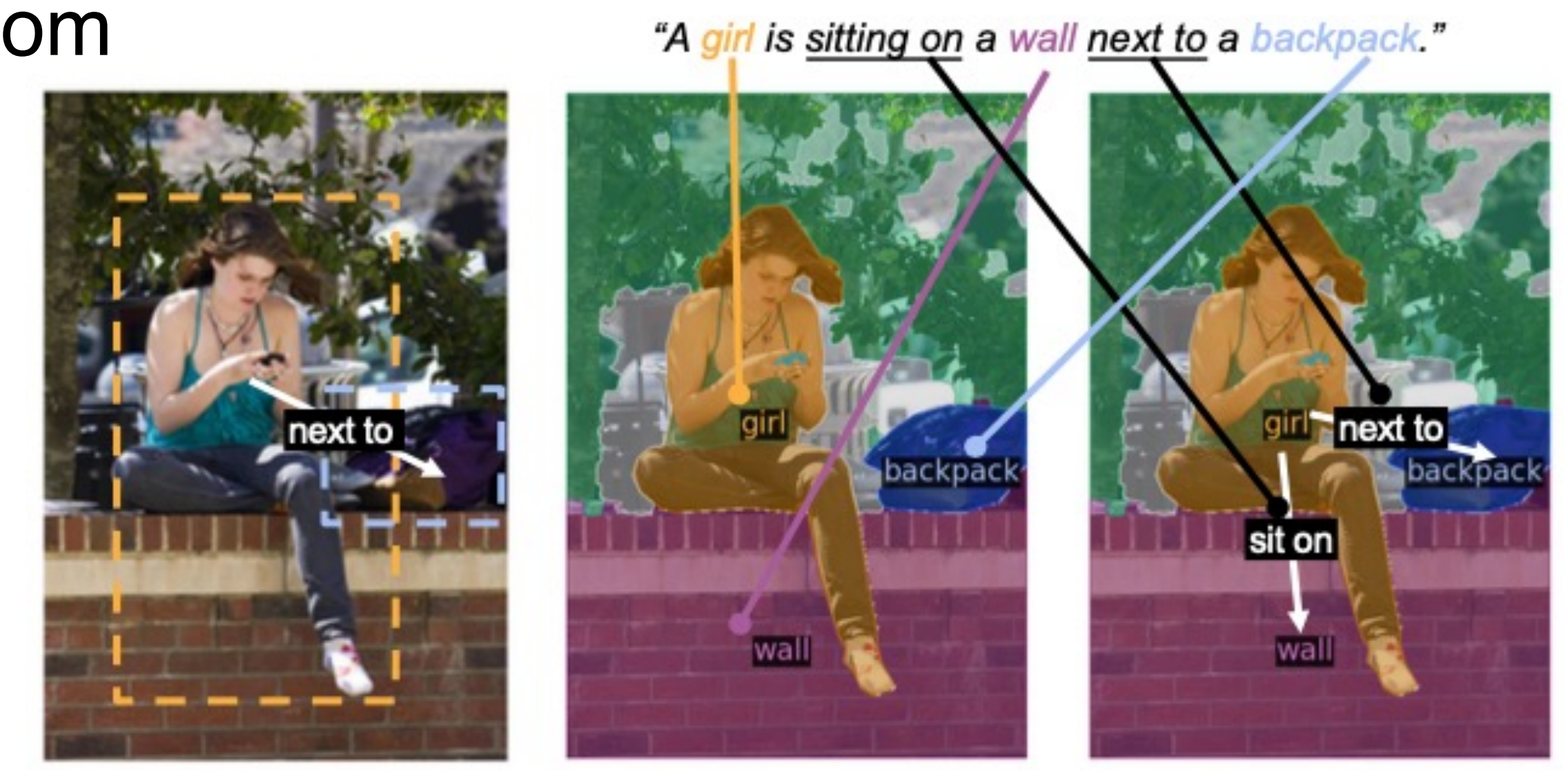
TextPSG: Panoptic Scene Graph Generation from Textual Descriptions

Chengyang Zhao [1] Yikang Shen [2] Zhenfang Chen [2] Mingyu Ding [3] Chuang Gan [2,4]
[1]Peking University [2] MIT-IBM Watson AI Lab [3] UC Berkley [4] UMass Amherst

ICCV23 PARIS

Motivation

Problem: Panoptic Scene Graph Generation from **Purely** Textual Descriptions (Caption-to-PSG)



previous box-based our mask-based scene graph generation from captions

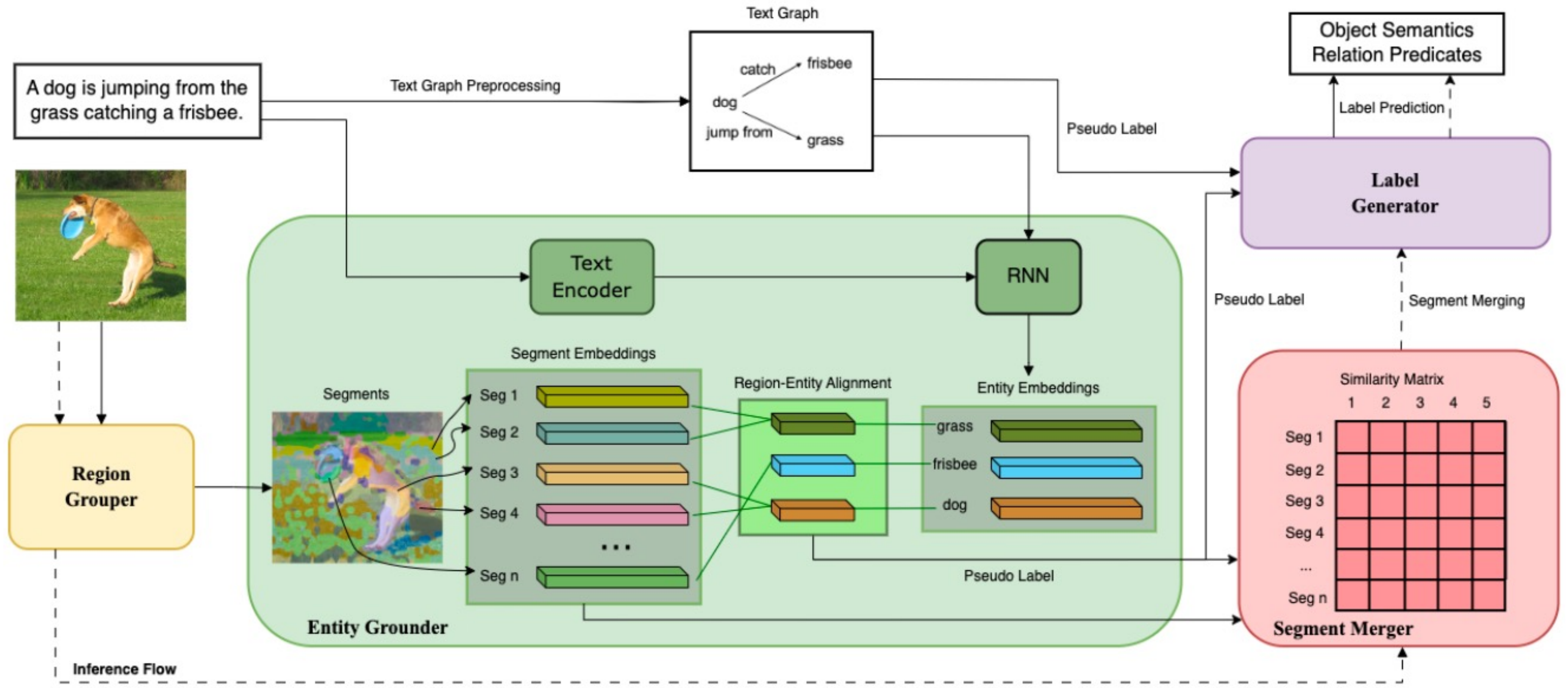
"Purely" for Three Constraints:

- No location priors
- No explicit region-entity links
- No pre-defined concept sets

Two Key Challenges:

- Learning to the ground entities in language onto the visual scene, developing the ability to perform partitioning and grounding purely from textual descriptions
- Learning the object semantics and relation predicates from textual descriptions, without pre-defined fixed object and relation vocabularies

Framework



Region Grouper: partitioning the image in a hierarchical way

Entity Grounder: grounding textual entities onto the image segments through a fine-grained contrastive learning strategy

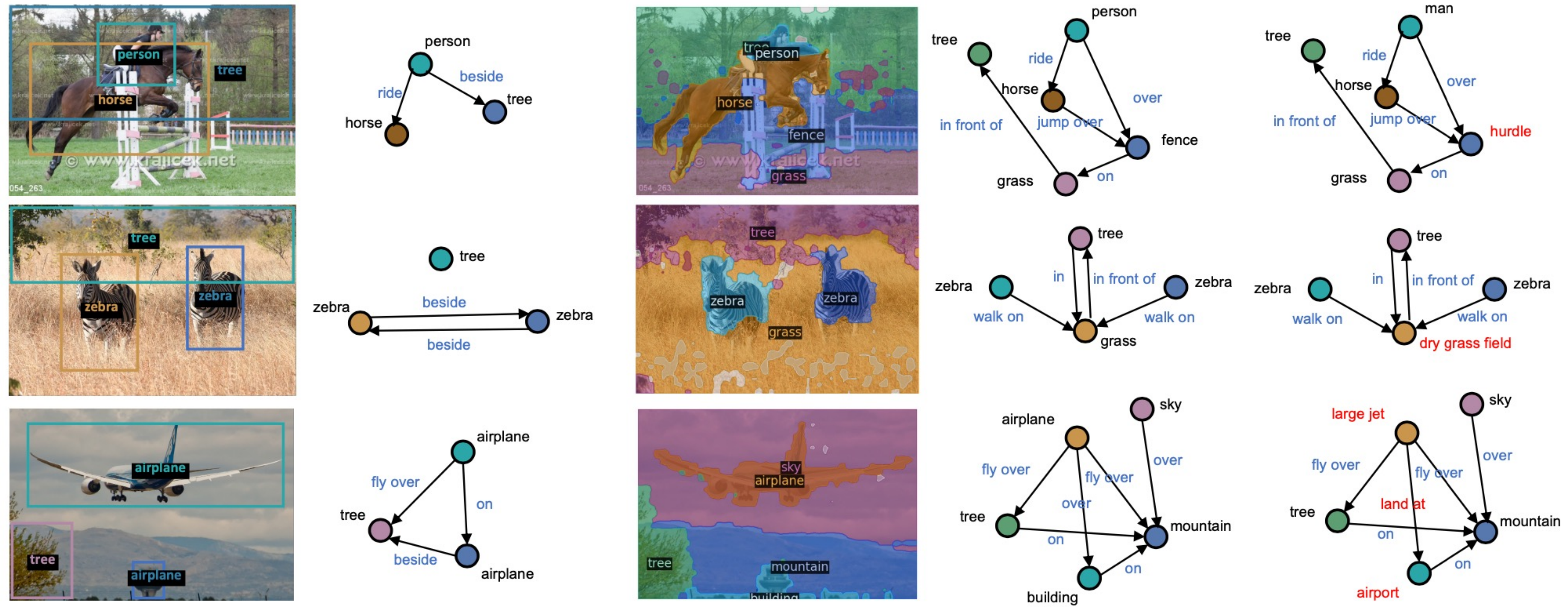
Segment Merger: leveraging the grounding results as explicit supervision to learn similarity matrices for inference-time merging

Label Generator: auto-regressive generation for prediction, leveraging pre-learned common sense from pre-trained language model, PET to better incorporate the common sense

Quantitative Results

Method			Mode	PhrDet				SGDet			
Model	Proposal	Target		N3R50	N3R100	N5R50	N5R100	N3R50	N3R100	N5R50	N5R100
SGGNLS-c	Detector	✔	bbox	9.69	11.45	10.24	12.22	6.76	7.81	7.2	8.65
Random	Selective Search	✘	bbox	0.02	0.03	0.02	0.03	0.01	0.02	0.02	0.03
Prior		✘	bbox	0.04	0.07	0.05	0.07	0.03	0.06	0.05	0.07
MIL		✘	bbox	1.97	2.18	2.04	2.61	1.2	1.35	1.56	1.97
SGCLIP		✘	bbox	3.02	3.45	3.38	3.71	2.13	2.3	2.39	2.7
SGGNLS-o	Detector	✘	bbox	6.2	6.79	6.92	7.93	3.96	4.21	4.53	5.02
Ours	–	✘	mask	8.28	9.16	9.06	10.51	3.32	3.63	3.71	4.18
Ours	–	✘	bbox	**11.37**	**12.74**	**12.24**	**14.37**	**4.29**	**4.77**	**4.82**	**5.48**

Qualitative Results

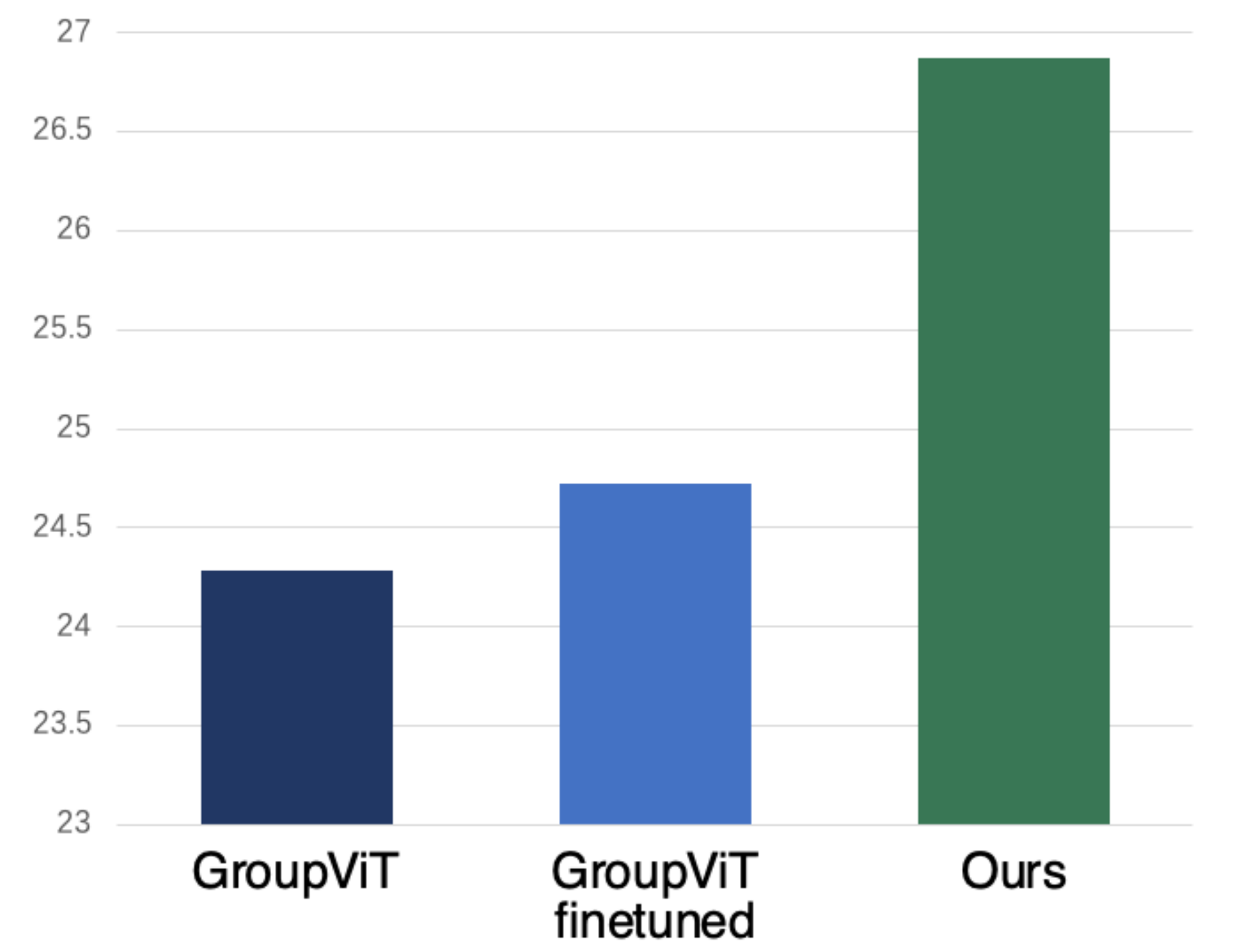


(a) Results of SGGNLS-o (b) Results of Ours

OOD Robustness

Set	Model	Target	Mode	PhrDet		SGDet	
				N3R100	N5R100	N3R100	N5R100
ID	SGGNLS-c	✔	bbox	**16.76**	**18.48**	**10.45**	**11.86**
	SGGNLS-o	✘	bbox	11.55	13.64	7.13	8.47
	Ours	✘	mask	9.27	10.45	3.28	3.76
	Ours	✘	bbox	13.35	14.82	4.63	5.36
OOD	SGGNLS-c	✔	bbox	0	0	0	0
	SGGNLS-o	✘	bbox	0.05	0.06	0	0
	Ours	✘	mask	8.47	9.76	4.07	4.51
	Ours	✘	bbox	**10.18**	**11.69**	**5.23**	**5.72**

Text-supervised Semantic Segmentation



Ablation Studies

Stage	#Seg	Cut	PhrDet		SGDet	
			N3R100	N5R100	N3R100	N5R100
1	64	✘	10.73	11.39	3.18	3.51
1	64	✔	**12.74**	**14.37**	**4.77**	**5.48**
2	8	✘	9.24	11.03	3.53	4.35
2	8	✔	6.78	8.45	2.46	3.21

Ablation Study on the Segment Merger.

Label Prediction	Model	PhrDet		SGDet	
		N3R100	N5R100	N3R100	N5R100
Cls + WordNet	-	8.82	9.36	2.36	2.72
Gen	RNN	9.12	10.44	2.65	3.07
Gen w/o PET	BLIP [21]	2.33	2.58	0.45	0.6
Gen w/ PET	BLIP [21]	**12.64**	**14.28**	**4.77**	**5.49**

Ablation Study on the Label Generator.

Project Page GitHub Code

Scan the QR code for more information and to contact us!