NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID:379

Code and Model

SimVG: A Simple Framework for Visual Grounding with Decoupled Multi-modal Fusion

Ming Dai[1], Lingfeng Yang[2], YiHao Xu[1], Zhenhua Feng[3], Wankou Yang[1,4]

[1]Southeast Univerisity, [2]Nanjing University of Science and Technology,

[3]Jiangnan University, [4]Advanced Ocean Institute of Southeast Univerisity, Nantong

E-mail: mingdai@seu.edu.cn,

Github: https://github.com/Dmmm1997/SimVG

Dec 10th –Dec 15th , 2024

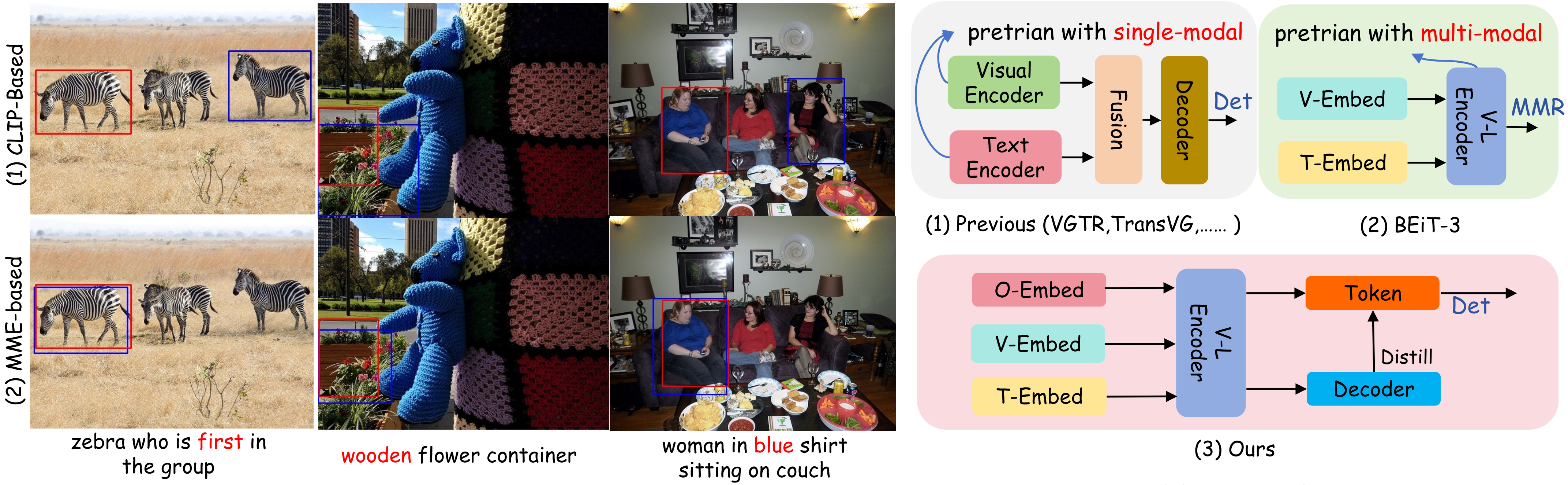
NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2024
Vancouver, Canada
Paper ID:379

Motivation

Starting Point: Independent encoding of image and text features, and then perform multimodal representation in limited downstream data makes **the** multimodal representation capability less general. This leads to insufficient understanding of the model for some fine-grained spatial and physical properties of images and text.

Core Solution: Decouples multimodal representation from limited downstream data to massive upstream data.



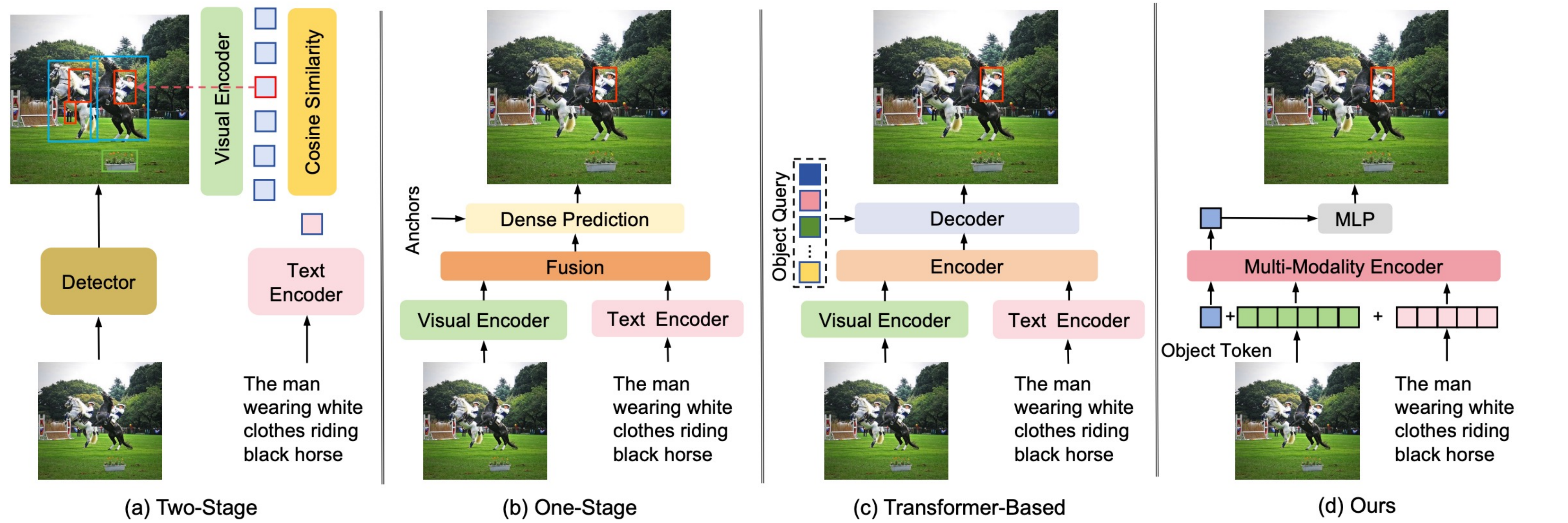
(a) Visualization Cases

(b) Framework

Architecture Comparision

Previous: All previous methods (including One-stage and Transformer-based) use the paradigm of encoding the image and text modalities separately, then performing multi-modal fusion representation, and finally predicting the result through the Decoder.

SimVG: This paper proposes a simpler paradigm, which directly encodes the image and text through a Multi-Modality Encoder and additionally defines an object token to predict the box.



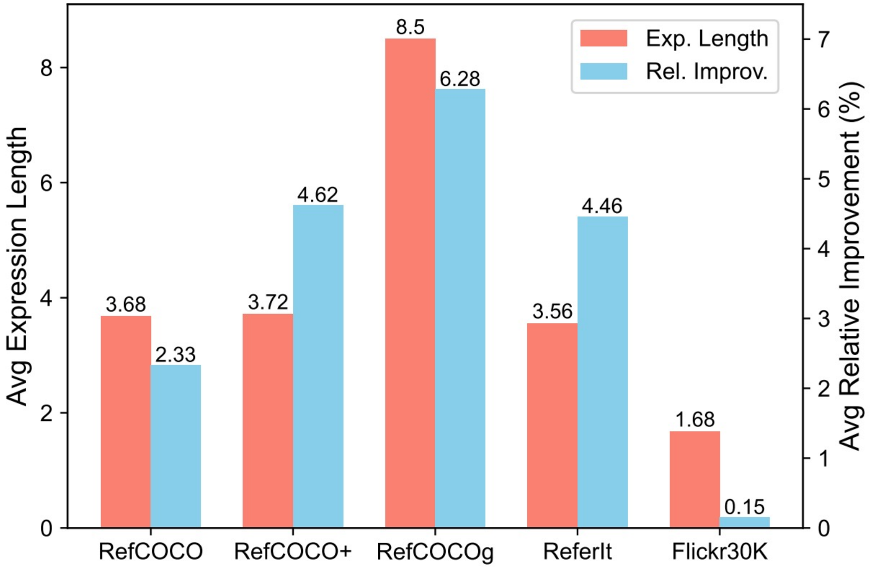
NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2024
Vancouver, Canada
Paper ID:379

Analysis

Setting: We analyze the relationship between the average sentence length and the relative improvement of SimVG compared to the SOTA model in five grounding datasets.

Conclusion: Decoupling multimodal representations to upstream can effectively improve the understanding of long and difficult texts.

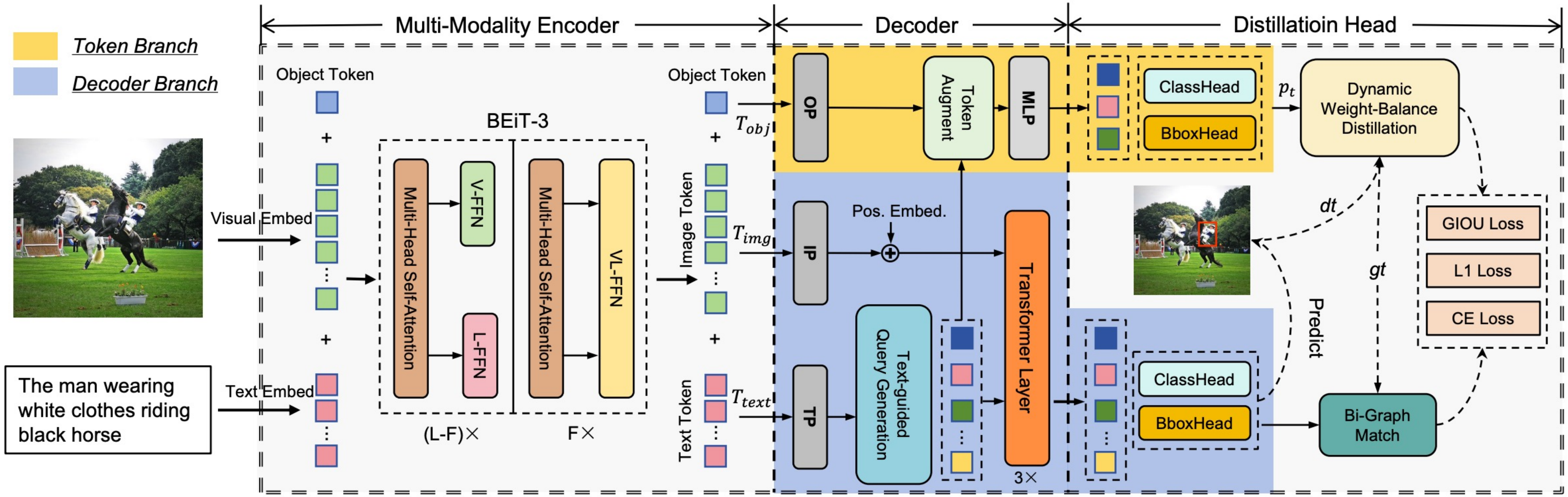


NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2024
Vancouver, Canada
Paper ID:379

SimVG Framework

Overall: BEiT-3 is introduced for feature extraction and integrated representation of Object, Visual, and Text. It is subsequently divided into a Transformer-based Decoder Branch (DB) and an lightweight MLP-based Token Branch (TB). We design a Dynamic Weight-Balance Distillation (DWBD), imparts DB abilities to the TB branch.。



NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS 2024
Vancouver, Canada
Paper ID:379

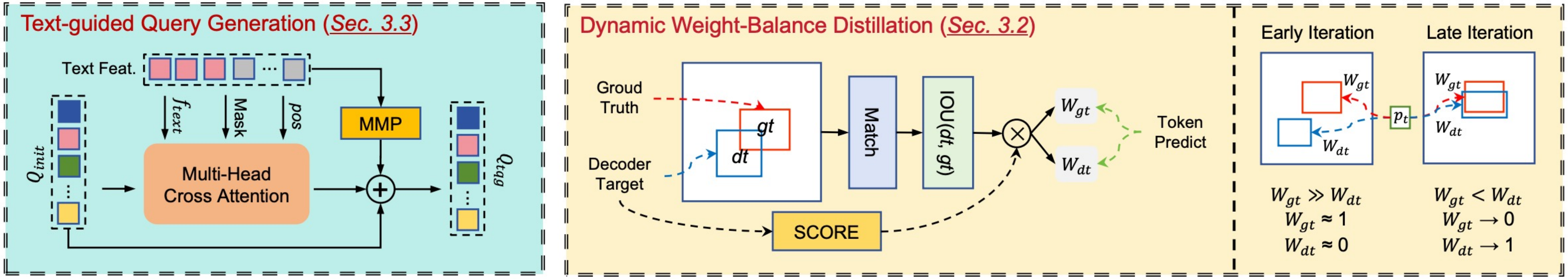
SimVG Framework

Text-guided Query Generation (TQG) :Textual word-level and sentence-level prior information is injected into the Object Query.

$$Q_{tqg} = \mathbf{MCA}(Q_{init}, f_{text} + pos, \mathrm{Mask}) + \mathbf{MMP}(f_{text}, \mathrm{Mask}) + Q_{init},$$

Dynamic Weight-Balance Distillation (DWBD) : the reliability of the teacher branch is adaptively understood through the IOU between DB and GT, and the weights for GT and the teacher branch are dynamically allocated for learning by the student branch.

$$W_{dt} = \frac{1}{N_{gt}} \sum_{i}^{N_{gt}} \left[\mathbf{IOU}(b_i, \hat{b}_{\hat{\sigma}}(i)) \times \mathbf{SCORE}(\hat{p}_{\hat{\sigma}(i)}) \right],$$



NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2024
Vancouver, Canada
Paper ID:379

Experiments Analysis

Description: Both ViLT and BEiT-3 employ fused representation architectures for pre-training, whereas CLIP is a dual-stream approach.

Analysis: The fused representation architecture alleviates the pressure of downstream fitting by decoupling MMR from upstream pre-training.

Conclusion: Decoupling multimodal representations from downstream to upstream not only effectively improves multimodal understanding capabilities but also significantly increases the efficiency of model training.

Method (ViT-B/32)	RefCOCO		
	val	testA	testB
CLIP [49]	73.93	77.14	67.43
ViLT [25]	78.54	82.31	72.47
BEiT-3 [60]	82.35	84.66	78.38
Baseline (BEiT-3)	82.35	84.66	78.38
+VE Interp.	85.37(**+3.02**)	86.67(**+2.01**)	81.57(**+3.19**)
Token Branch	85.47	86.75	81.66
Decoder Branch	86.78	88.19	82.83

Table 4: Some ablation experiments on different multimodal fusion architectures. VE Interp. refers to the downsampling convolution kernel in Visual Embed that performs bilinear interpolation from pre-trained weights.

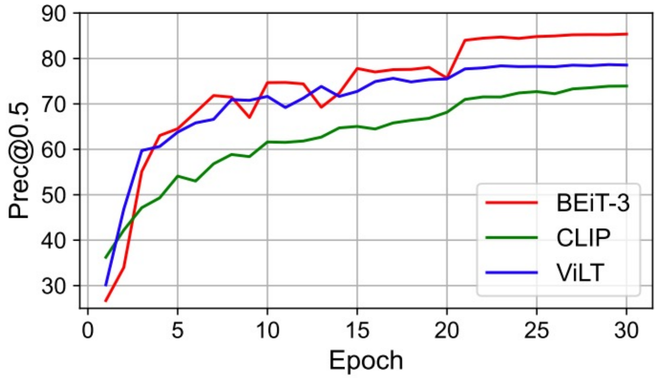


Figure 4: The convergence speed of three different multimodal pretraining architecture models.

Main Results

Our method consistently surpasses existing approaches and achieves SoTA performance on six grounding datasets while maintaining high inference speed.

Tab.1 Comparison with SoTA methods on REC task.

Models	Visual Encoder	RefCOCO val	RefCOCO testA	RefCOCO testB	RefCOCO+ val	RefCOCO+ testA	RefCOCO+ testB	RefCOCOg val-g	RefCOCOg val-u	RefCOCOg test-u	ReferIt test	Flickr30k test	Time (ms)
Two-Stage													
MAttNet [72]	RN101	76.40	80.43	69.28	64.93	70.26	56.00	-	66.58	67.27	29.04	-	320
CM-Att-Erase [42]	RN101	78.35	83.14	71.32	68.09	73.65	58.03	-	67.99	68.67	-	-	-
DGA [66]	VGG16	-	78.42	65.53	-	69.07	51.99	-	-	63.28	-	-	341
NMTree [36]	RN101	76.41	81.21	70.09	66.46	72.02	57.52	64.62	65.87	66.44	-	-	-
One-Stage													
RealGIN [77]	DN53	77.25	78.70	72.10	62.78	67.17	54.21	-	62.75	62.33	-	-	35
FAOA [69]	DN53	71.15	74.88	66.32	56.86	61.89	49.46	-	59.44	58.90	60.67	68.71	39
RCCF [34]	DLA34	-	81.06	71.85	-	70.35	56.32	-	-	65.73	63.79	-	**25**
MCN [44]	DN53	80.08	82.29	74.98	67.16	72.86	57.31	-	66.46	66.01	-	-	56
ReSC$_L$ [67]	DN53	77.63	80.45	72.30	63.59	68.36	56.81	63.12	67.30	67.20	64.60	69.28	36
LBYL [19]	DN53	79.67	82.91	74.15	68.64	73.38	59.49	62.70	-	-	67.47	-	<u>30</u>
Transformer-Based													
TransVG [7]	RN101	81.02	82.72	78.35	64.82	70.70	56.94	67.02	68.67	67.73	70.73	79.10	62
TRAR [78]	DN53	-	81.40	78.60	-	69.10	56.10	-	68.90	68.30	-	-	-
VGTR [11]	RN50	78.29	81.49	72.38	63.29	70.01	55.64	61.64	64.19	64.01	63.63	75.44	-
SeqTR [79]	DN53	83.72	86.51	81.24	71.45	76.26	64.88	71.50	74.86	74.21	69.66	81.23	50
VLTVG [64]	RN50	84.53	87.69	79.22	73.60	78.37	64.53	72.53	74.90	73.88	71.60	79.18	79*
TransCP [57]	RN50	84.25	87.38	79.78	73.07	78.05	63.35	72.60	-	-	72.05	80.04	74*
Dyn.MDETR [54]	ViT-B/16	85.97	88.82	80.12	74.83	81.70	63.44	72.21	74.14	74.49	70.37	81.89	-
SimVG-TB (ours)	ViT-B/32	87.07	89.04	83.57	78.84	83.64	70.67	77.66	79.82	79.93	74.59	81.59	44
SimVG-DB (ours)	ViT-B/32	87.63	90.22	84.04	78.65	83.36	71.82	78.81	80.37	80.51	74.83	82.04	52
SimVG-TB (ours)	ViT-L/32	**90.61**	**92.53**	**87.68**	**85.36**	**89.61**	**79.74**	<u>79.34</u>	**85.99**	**86.83**	**79.30**	<u>82.61</u>	101
SimVG-DB (ours)	ViT-L/32	<u>90.51</u>	<u>92.37</u>	<u>87.07</u>	<u>84.88</u>	<u>88.50</u>	<u>78.66</u>	**80.42**	<u>85.72</u>	<u>86.70</u>	<u>78.75</u>	**83.15**	116

Main Results

Our method consistently surpasses existing approaches and achieves SoTA performance on six grounding datasets while maintaining high inference speed.

Tab.2 Comparison with SoTA methods on GREC task

Methods	Visual Encoder	Textual Encoder	val		testA		testB	
			Prec@(F_1@0.5)	N-acc.	Prec@(F_1@0.5)	N-acc.	Prec@(F_1@0.5)	N-acc.
MCN [44]	DN53	GRU	28.0	30.6	32.3	32.0	26.8	30.3
VLT [9]	DN53	GRU	36.6	35.2	40.2	34.1	30.2	32.5
MDETR [23]	RN101	RoBERTa	42.7	36.3	50.0	34.5	36.5	31.0
UNINEXT [63]	RN50	BERT	58.2	50.6	46.4	49.3	42.9	48.2
SimVG-TB (ours)	ViT-B/32	/	61.3	**56.1**	61.7	**58.0**	53.1	**57.5**
SimVG-DB (ours)	ViT-B/32	/	**62.1**	54.7	**64.6**	57.2	**54.8**	57.2

NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2024
Vancouver, Canada
Paper ID:379

Main Results

Our method consistently surpasses existing approaches and achieves SoTA performance on six grounding datasets while maintaining high inference speed. And less pre-training data is adopted.

Tab.3 Comparison with SoTA methods on REC task using pre-training data

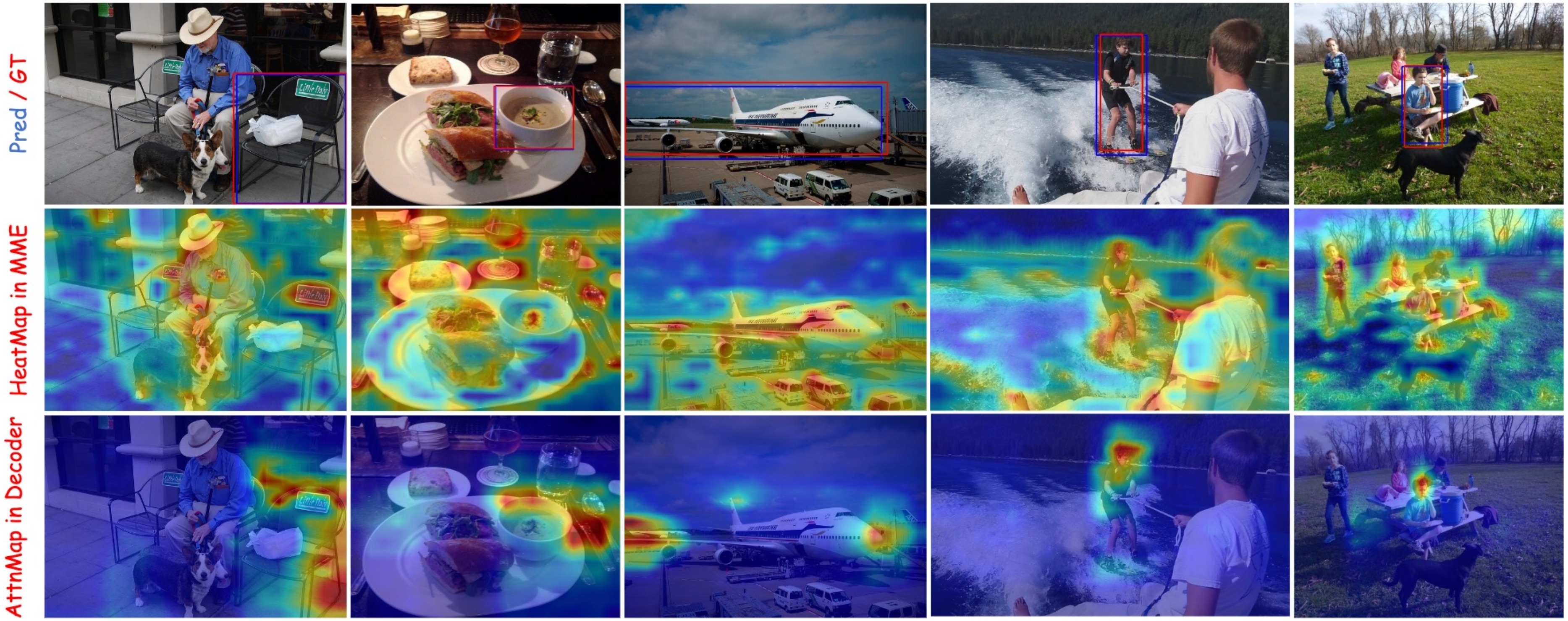
Models	Visual Encoder	Params (M)	Pre-train images	RefCOCO			RefCOCO+			RefCOCOg		Time (ms)
				val	testA	testB	val	testA	testB	val-u	test-u	
UNITER$_L$ [5]	RN101	-	4.6M	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77	-
VILLA$_L$ [13]	RN101	-	4.6M	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71	-
MDETR [23]	RN101	17.36	200K	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89	108
RefTR [31]	RN101	17.86	100K	85.65	88.73	81.16	77.55	82.26	68.99	79.25	80.01	**40**
SeqTR [79]	DN53	7.90	174K	87.00	90.15	83.59	78.69	84.51	71.87	82.69	83.37	50
UniTAB [68]	RN101	-	200K	88.59	91.06	83.75	80.97	85.36	71.55	84.58	84.70	-
DQ-DETR [39]	RN101	-	200K	88.63	91.04	83.51	81.66	86.15	73.21	82.76	83.44	-
GroundingDINO[41]	Swin-T	-	200K	89.19	91.86	85.99	81.09	87.40	74.71	84.15	84.94	120
PolyFormer[38]	Swin-B	-	174K	89.73	91.73	86.03	83.73	88.60	76.38	84.46	84.96	-
SimVG-DB (ours)	ViT-B/32	6.32	**28K**	90.98	92.68	87.94	84.17	88.58	78.53	85.90	86.23	52
SimVG-TB (ours)	ViT-B/32	**1.58**	174K	90.59	92.80	87.04	83.54	88.05	77.50	85.38	86.28	44
SimVG-DB (ours)	ViT-B/32	6.32	174K	91.47	93.65	87.94	84.83	88.85	79.12	86.30	87.26	52
SimVG-TB (ours)	ViT-L/32	**1.58**	**28K**	**92.99**	**94.86**	90.12	**87.43**	91.02	82.10	87.95	88.96	101
SimVG-DB (ours)	ViT-L/32	6.32	**28K**	92.93	94.70	**90.28**	87.28	**91.64**	**82.41**	**87.99**	**89.15**	116

NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2024
Vancouver, Canada
Paper ID:379

Qualitatity Results

In the MME, the attention mechanism primarily focuses on the foreground objects within the image, while the attention in the decoder concentrates on referring information related to the text describtion.



a chair have black color & a white carrybag is here

a bowl of soup next to a sandwich

airplane sitting ground fedex plane in the background

young boy on water skis

boy with banian and short eating with left hand before a dog

NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2024
Vancouver, Canada
Paper ID:379

Conclusion

Code and Model

We propose a simple Visual Grounding paradigm that migrates multimodal representations from downstream tasks to upstream pre-training, enhancing the comprehension capability of image-text content. Furthermore, we design a Dynamic Weight-Balance Distillation (DWBD) to bolster the performance of the lightweight MLP branch, achieving efficient inference while maintaining high performance. Ultimately, our method has achieved advanced results in six mainstream detection datasets.

E-mail: mingdai@seu.edu.cn,
Github: https://github.com/Dmmm1997/SimVG

Dec 10th -Dec 15th , 2024