ACM Multimedia 2024
Melbourne, Australia

Paper ID: 2398
Posterboard: P161

HiVG: Hierarchical Multimodal Fine-grained Modulation for Visual Grounding

Linhui Xiao[1,2,3], Xiaoshan Yang[1,2,3], Fang Peng[1,2,3], Yaowei Wang[2], Changsheng Xu[1,2,3*]
[1] MAIS, Institute of Automation, Chinese Academy of Sciences [2] Pengcheng Laboratory
[3] School of Artificial Intelligence, University of Chinese Academy of Sciences
E-mail: xiaolinhui16@mails.ucas.ac.cn, Github: https://github.com/linhuixiao/HiVG

MOTIVATION

Visual grounding, which aims to ground a visual region via natural language. Existing works utilized uni-modal pre-trained models to transfer visual or linguistic knowledge separately while ignoring the multimodal corresponding information. Motivated by recent advancements in contrastive language-image pre-training and low-rank adaptation (LoRA) methods, we aim to solve the grounding task based on multimodal pre-training. However, there exists significant task gaps between pre-training and grounding. Therefore, in this paper, we propose a concise and efficient hierarchical multimodal fine-grained modulation framework, namely **HiVG**.

Fig.1 Data granularity gaps between pre-training and grounding.

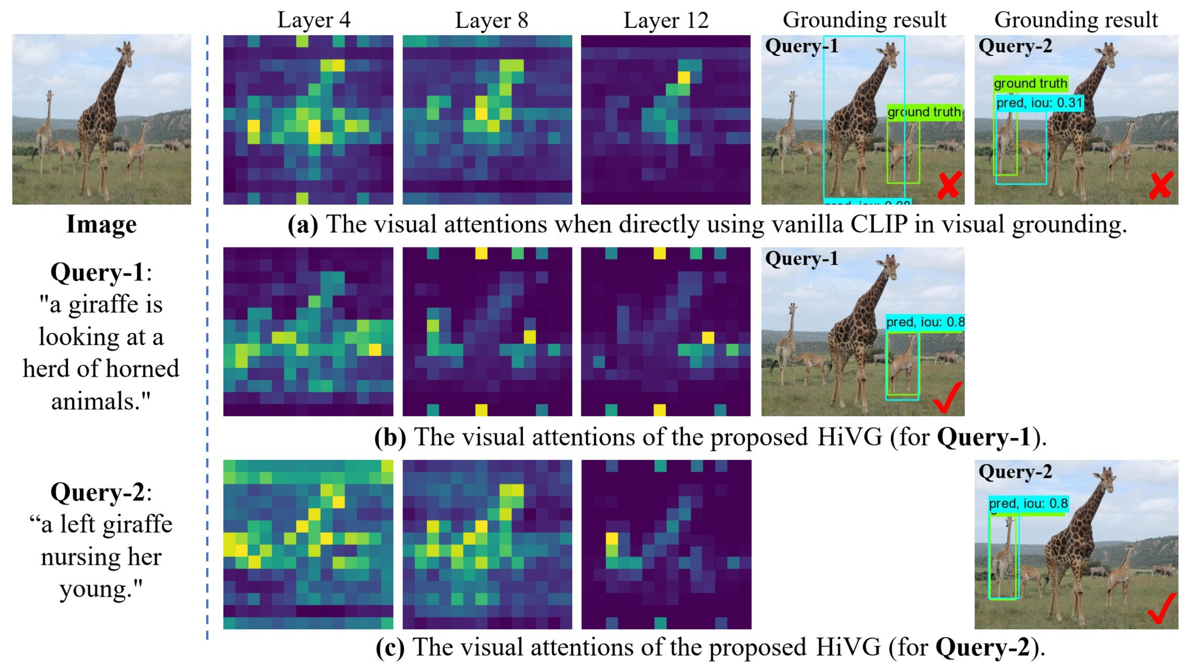

Fig.2 Visual attentions and grounding results of CLIP and the proposed HiVG.

METHODS

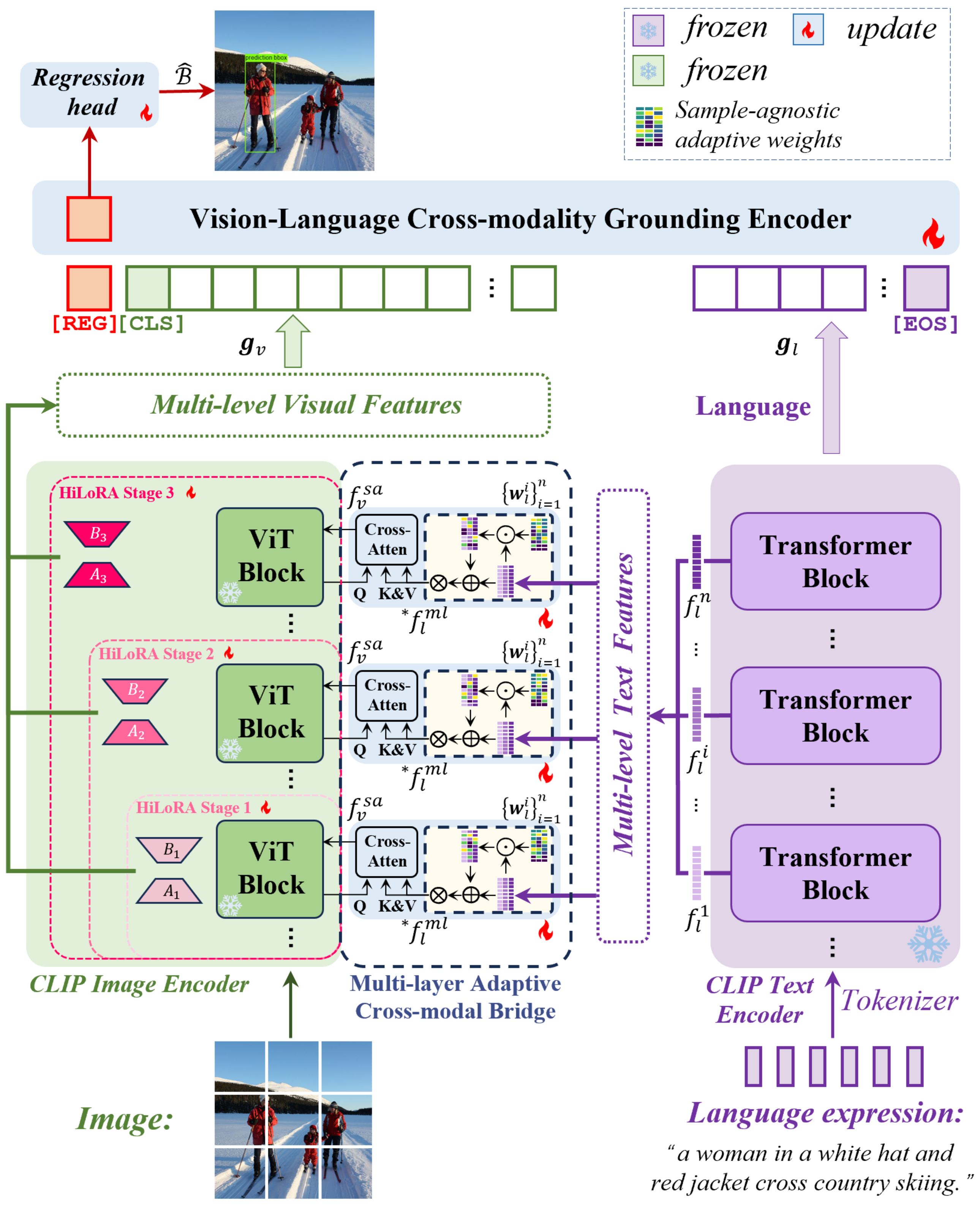

Fig.3 The HiVG framework architecture.

HiVG consists of a multi-layer adaptive cross-modal bridge **(MACB, Fig.4)** and a hierarchical multimodal low-rank adaptation **(HiLoRA, Fig.5)** paradigm.

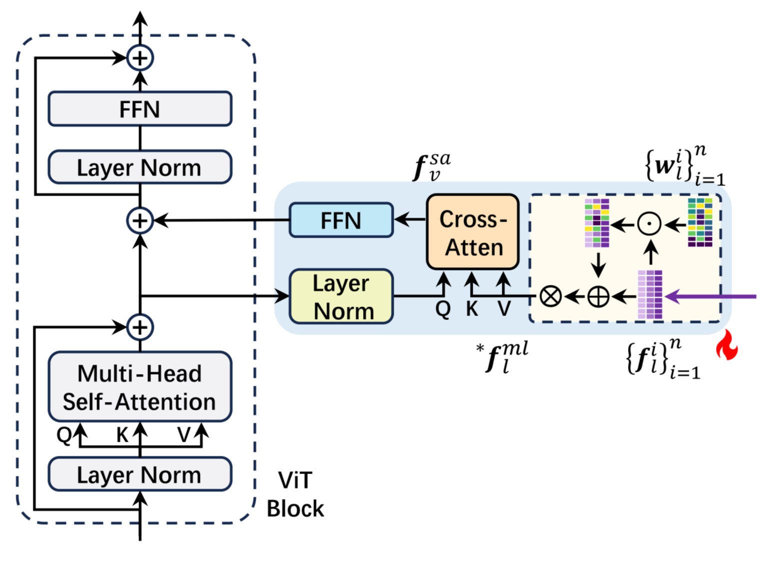

Fig.4 The multi-layer adaptive cross-modal bridge (MACB).

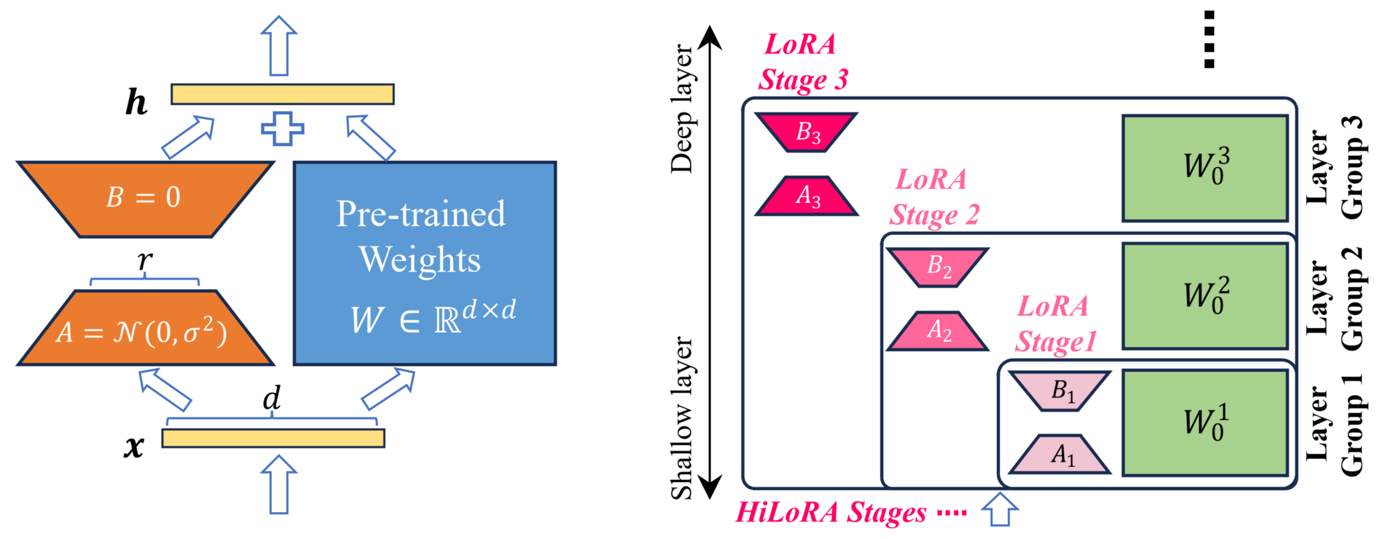

Fig.5 Our proposed HiLoRA and vanilla LoRA.

(1) The **MACB (Fig.4)** can address the inconsistency between visual features and those required for grounding and establish a connection between multi-level visual and text features.

$$^{*}f_l^i = w_l^i \odot f_l^i + f_l^i \qquad ^{*}f_l^{ml} = \mathrm{concat}[^{*}f_l^1, ^{*}f_l^2, \cdots, ^{*}f_l^n] \otimes W_{proj}$$

(2) The **HiLoRA (Fig.5)** prevents the accumulation of perceptual errors by adapting the cross-modal features from shallow to deep layers in a hierarchical manner.

Vanilla LoRA: $h = W_0 x + \Delta W x = W_0 x + BAx$

HiLoRA:
$$h_j^l = \begin{cases} W_0^l x^l, & when\ l > j \cdot L/G, \\ W_0^l x^l + \sum_{k=\lceil l \cdot G/L \rceil}^{j} B_k^l A_k^l x^l, & when\ l \le j \cdot L/G \end{cases}$$

HiLoRA with MACB:

When $l > j \cdot L/G$:
$$h_j^l = \begin{cases} W_0^l f_v^{l-1}, & when\ l \notin C, \\ W_0^l (f_v^{l-1} + f_v^{sa}), & when\ l \in C \end{cases}$$

While in $l \le j \cdot L/G$:
$$h_j^l = \begin{cases} W_0^l f_v^{l-1} + \sum_{k=\lceil l \cdot G/L \rceil}^{j} B_k^l A_k^l f_v^{l-1}, & when\ l \notin C, \\ W_0^l (f_v^{l-1} + f_v^{sa}) + \sum_{k=\lceil l \cdot G/L \rceil}^{j} B_k^l A_k^l (f_v^{l-1} + f_v^{sa}), & when\ l \in C. \end{cases}$$

(3) Training objectives.

$$\mathcal{L}_{BOX} = \lambda_{l_1} \mathcal{L}_{\text{smooth-l1}}(\hat{\mathcal{B}}, \mathcal{B}) + \lambda_{giou} \mathcal{L}_{\text{giou}}(\hat{\mathcal{B}}, \mathcal{B}) \qquad \mathcal{L}_{total} = \mathcal{L}_{BOX} + \mathcal{L}_{CLC} + \mathcal{L}_{RTCC}$$

RESULTS

Experimental results on five datasets demonstrate the effectiveness of our approach and showcase the significant grounding capabilities as well as promising energy efficiency advantages.

Tab.1 Comparison HiVG with latest SoTA methods on RefCOCO/+/g *etc.*.

Methods	Venue	Visual Backbone	Language Backbone	Multi-task	RefCOCO val	RefCOCO testA	RefCOCO testB	RefCOCO+ val	RefCOCO+ testA	RefCOCO+ testB	RefCOCOg val	RefCOCOg test	ReferIt test	Flickr test
Fine-tuning *w.* uni-modal pre-trained close-set detector and language model: (traditional setting)														
TransVG [9]	ICCV'21	RN101+DETR	BERT-B	✗	81.02	82.72	78.35	64.82	70.70	56.94	68.67	67.73	70.73	79.10
SeqTR [79]	ECCV'22	DN53	BiGRU	✗	81.23	85.00	76.08	68.82	75.37	58.78	71.35	71.58	69.66	81.23
RefTR* [27]	NeurIPS'21	RN101+DETR	BERT-B	✓	82.23	85.59	76.57	71.58	75.96	62.16	69.41	69.40	71.42	78.66
Word2Pix [77]	TNNLS'22	RN101+DETR	BERT-B	✗	81.20	84.39	78.12	69.74	76.11	61.24	70.81	71.34	–	–
QRNet [72]	CVPR'22	Swin-S[40]	BERT-B	✗	84.01	85.85	82.34	72.94	76.17	63.81	71.89	73.03	74.61	81.95
VG-LAW [56]	CVPR'23	ViT-Det [29]	BERT-B	✗	86.06	88.56	82.87	75.74	80.32	66.69	75.31	75.95	76.60	–
TransVG++[10]	TPAMI'23	ViT-Det [29]	BERT-B	✗	86.28	88.37	80.97	75.39	80.45	66.28	76.18	76.30	74.70	81.49
Fine-tuning *w.* vision-language self-supervised pre-trained model:														
CLIP-VG [64]	TMM'23	CLIP-B	CLIP-B	✗	84.29	87.76	78.43	69.55	77.33	57.62	73.18	72.54	70.89	81.99
JMRI [80]	TIM'23	CLIP-B	CLIP-B	✗	82.97	87.30	74.62	71.17	79.82	57.01	71.96	72.04	68.23	79.90
Dynamic-MDETR	TPAMI'23	CLIP-B	CLIP-B	✗	85.97	88.82	80.12	74.83	81.70	63.44	74.14	74.49	70.37	81.89
HiVG (ours)	ACM MM'24	CLIP-B	CLIP-B	✗	87.32	89.86	83.27	78.06	83.81	68.11	78.29	78.79	75.22	82.11
HiVG-L (ours)	ACM MM'24	CLIP-L	CLIP-L	✗	**88.14**	**91.09**	**83.71**	**80.10**	**86.77**	**70.53**	**80.78**	**80.25**	**76.23**	**82.16**
Fine-tuning *w.* box-level dataset-mixed open-set detection pre-trained model / multi-task mix-supervised pre-trained model:														
MDETR † [20]	ICCV'21	RN101+DETR	RoBERT-B	✗	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89	–	83.80
YORO† [16]	ECCV'22	ViLT [24]	BERT-B	✗	82.90	85.60	77.40	73.50	78.60	64.90	73.40	74.30	71.90	–
DQ-DETR † [33]	AAAI'23	RN101+DETR	BERT-B	✗	88.63	91.04	83.51	81.66	86.15	73.21	82.76	83.44	–	–
Grounding-DINO†	Arxiv'23	Swin-T	BERT-B	✗	89.19	91.86	85.99	81.09	87.40	74.71	84.15	84.94	–	–
UniTAB † [70]	ECCV'22	RN101+DETR	RoBERT-B	✓	86.32	88.84	80.61	78.70	83.22	69.48	79.96	79.97	–	79.38
OFA-B † [61]	ICML'22	OFA-B	OFA-B	✓	88.48	90.67	83.30	81.39	87.15	74.29	82.29	82.31	–	–
OFA-L † [61]	ICML'22	OFA-L	OFA-L	✓	90.05	92.93	85.26	85.80	89.87	**79.22**	85.89	86.55	–	–
HiVG† (ours)	ACM MM'24	CLIP-B	CLIP-B	✗	90.56	92.55	87.23	83.08	89.21	76.68	84.52	85.62	77.75	82.08
HiVG-L† (ours)	ACM MM'24	CLIP-L	CLIP-L	✗	**90.77**	**92.94**	**88.03**	**86.78**	**89.91**	78.02	**86.61**	**86.60**	**78.16**	**82.63**

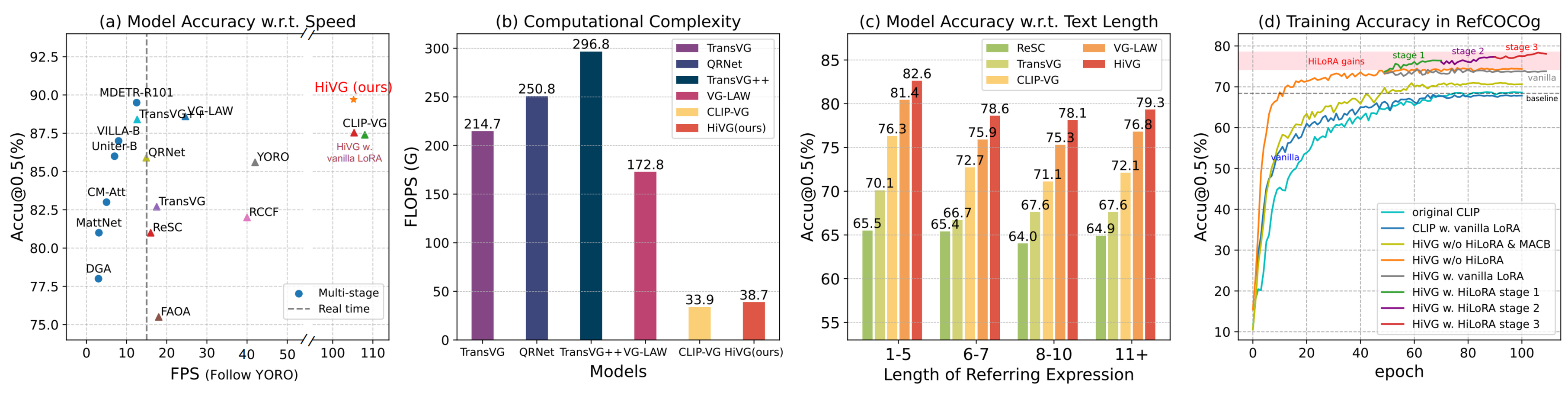

Fig.6 Visualization comparing HiVG (base) with other SoTA models.

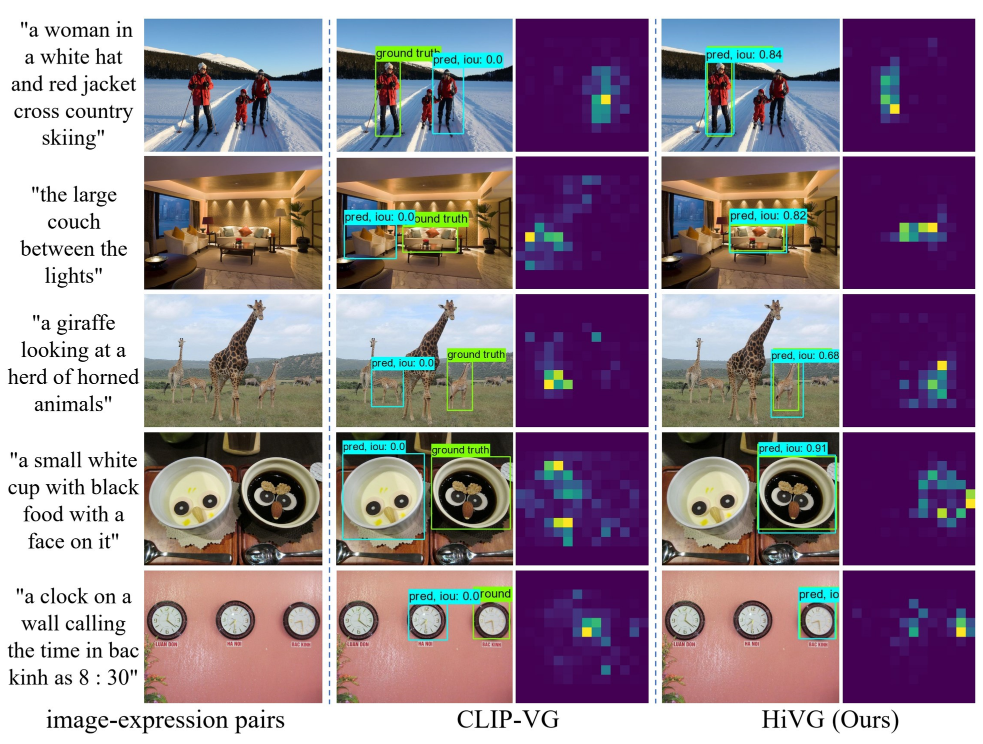

Fig.7 Qualitative results of our HiVG.

CONCLUSION

HiVG effectively implements fine-grained adaptation of the pre-trained model in the complex grounding task. It is a concise and efficient end-to-end framework. Our exploration in hierarchical cross-modal features offer new insights for the future grounding research.