

Supplementary Materials for "HiVG: Hierarchical Multimodal Fine-grained Modulation for Visual Grounding"

Linhui Xiao
¹MAIS, Institute of Automation,
Chinese Academy of Sciences
²Pengcheng Laboratory
³School of Artificial Intelligence,
University of Chinese Academy of
Sciences
xiaolinhui16@mails.ucas.ac.cn

Xiaoshan Yang
¹MAIS, Institute of Automation,
Chinese Academy of Sciences
²Pengcheng Laboratory
³School of Artificial Intelligence,
University of Chinese Academy of
Sciences
xiaoshan.yang@nlpr.ia.ac.cn

Fang Peng
¹MAIS, Institute of Automation,
Chinese Academy of Sciences
²Pengcheng Laboratory
³School of Artificial Intelligence,
University of Chinese Academy of
Sciences
pengfang21@mails.ucas.ac.cn

Yaowei Wang
¹Pengcheng Laboratory
²Harbin Institute of Technology
(Shenzhen)
wangyw@pcl.ac.cn

Changsheng Xu*
¹MAIS, Institute of Automation,
Chinese Academy of Sciences
²Pengcheng Laboratory
³School of Artificial Intelligence,
University of Chinese Academy of
Sciences
csxu@nlpr.ia.ac.cn

1 ANALYSIS OF THE DATASETS

We present the statistical analysis of the five datasets employed in our experimental study. Tab. 1 presents the detailed statistics.

RefCOCO/RefCOCO+/RefCOCOg. These three datasets belong to the Referring Expression Comprehension (REC), and the images of these three datasets derived from MSCOCO [8]. Expressions in RefCOCO [19] and RefCOCO+ [19] are also collected by the two-player game proposed in ReferItGame [5]. There are two test splits called "testA" and "testB". Images in "testA" only contain multiple people annotation. In contrast, images in "testB" contain all other objects. Expressions in RefCOCOg [9] are collected on Amazon Mechanical Turk in a non-interactive setting. Thus, the expressions in RefCOCOg are longer and more complex. RefCOCOg has "google" and "umd" splits. The "google" split does not have a public test set, and there is an overlap between the training and validation image sets. The "umd" split does not have this overlap. Therefore, we followed the previous studies [12, 18] and tested the RefCOCOg dataset only on the "umd" split.

ReferItGame. ReferItGame [5] contains images from SAIAPR12 [3] and collects expressions through a two-player game. In this game, the first player is shown an image with an object annotation and is asked to write a natural language expression referring to the object. The second player is then shown the same image along with the written expression and is asked to click on the corresponding area of the object. If the clicking is correct, both players receive points and swap roles. If not, a new image will be presented.

Flickr30k Entities. Flickr30k Entities (Flickr30k for short) [10] contains images in Flickr30k dataset. The query sentences are short noun phrases in the captions of the image. The queries are simpler and easier to understand compared to RefCOCO+/g. Therefore, the ambiguity of the expression is heightened simultaneously, resulting in a relative increase in noise.



Figure 1: An illustration of dataset granularity gaps between pre-training task and downstream grounding task. The samples are derived from LAION-400M [11] and RefCOCOg [9] datasets, respectively.

Table 1: The detailed statistics of RefCOCO [19], RefCOCO+ [19], RefCOCOg [9], ReferItGame [5] and Flickr30k[10] datasets. We represent test and testA split in same column.

Dataset	Images	Instances	total queries	train queries	val queries	test(A) queries	testB queries
RefCOCO [19]	19,994	50,000	142,210	120,624	10,834	5,657	5,095
RefCOCO+ [19]	19,992	49,856	141,564	120,191	10,768	5,726	4,889
RefCOCOg [9]	25,799	49,822	95,010	80,512	4,896	9,602	-
ReferItGame [5]	20,000	19,987	120,072	54,127	5,842	60,103	-
Flickr30k [10]	31,783	427,000	456,107	427,193	14,433	14,481	-

Dataset Granularity Gaps between Pre-training and Downstream Grounding. CLIP utilizes the LAION-400M dataset [11] for self-supervised pre-training, which is a noisy web dataset containing 400 million image-text pairs. As shown in Fig. 1, we present an illustration of task granularity gaps between pre-training task and downstream grounding task. It can be observed that the self-supervised pre-training typically learns coarse-grained visual and linguistic concepts from noisy web data (Fig. 1-(a)), while visual grounding requires more refined and complex interaction and alignment between linguistic and visual information (Fig. 1-(b)). The samples are derived from LAION-400M [11] and RefCOCOg [9] datasets, respectively.

Table 4: Ablation study of the training loss, includes Contrastive Learning Constraint (CLC) and Region-Text Contrastive Constraint (RTCC).

RTCC	CLC	Accu@0.5(%) val	test
✗	✗	training unstable	
✓	✗	training unstable	
✗	✓	77.21	77.48
✓	✓	78.29	78.79

Table 5: More ablation study on the implementation of the multi-layer adaptive cross-modal bridge (MACB) on RefCOCOg dataset. w. denotes with. (Accu@0.5(%))

Architecture	Accu@0.5(%) val	test
MACB w. sample-aware weights (with 12 layers)	77.07	77.98
MACB w. only last layer of text features	76.84	77.02
MACB w. (6 th , 12 th) layer of text features	77.08	77.82
MACB w. (1 th , 4 th , 8 th , 12 th) layer of text features	77.65	78.45
MACB w. (1th - 12th) layer of text features	78.29	78.79

Table 6: Ablation study of HiVG by utilizing multi-level visual features of CLIP on RefCOCOg dataset. (Accu@0.5(%))

Architecture	Accu@0.5(%) val	test
HiVG w. (12 th) layer	68.69	67.43
HiVG w. (2 th , 5 th , 9 th) layer	71.02	71.98
HiVG w. (2 th , 5 th , 9 th , 12 th) layer	71.63	72.01
HiVG w. (1th, 4th, 8th, 12th) layer	72.37	72.15
HiVG w. (3 th , 6 th , 9 th , 12 th) layer	72.08	72.04
HiVG w. (6 th - 12 th) layer	71.49	71.75
HiVG w. (1 th - 12 th) layer	71.25	71.15

Following YORO, the FPS of RCCF [7], MattNet [18] and DGA [14] are copied from RCCF work [7], which measures the speed on a Titan Xp GPU (identical to YORO) and Intel Xeon E5-2680v4 CPU@2.4GHZ. For a fair comparison, we normalize the results of TransVG++ [2], VG-LAW [12], CLIP-VG [13], and our HiVG by dividing them with a factor of 3.4 to account for the higher computational capabilities of our NVIDIA A100 GPU and Intel Xeon Gold 6240R CPU@2.4GHZ setup. This normalization factor is derived from comparing the FPS achieved by TransVG on our device (*i.e.*, 59.55, as in Table 2 of the main text) with the reported FPS in YORO (*i.e.*, 17.51). As can be seen in the figure, both our HiVG and CLIP-VG are based on small-resolution images and achieve significantly faster inference speed. Meanwhile, our HiVG achieves the best trade-off between performance and speed.

Analysis of the Computational Complexity in Figure 4-(b) of the Main Text. According to Table 2 of the main text, the number of parameters in existing models (except for QRNet [17]) is not significantly different, roughly ranging from 150M to 210M.

However, the computational complexity of the Transformer architecture heavily depends on the length of input token sequences, *i.e.*, there is an $\mathcal{O}(n^2)$ complexity. For example, TransVG++ [2] utilizes a 640×640 resolution image as input with a patch size of 16×16, resulting in a sequence length of $(640/16)^2 = 40^2 = 1600$ in the vision backbone. In contrast, our HiVG employs a smaller resolution image of 224×224 with a patch size of 16×16; thus, our visual sequence length is only $(224/16)^2 = 14^2 = 196$. Taking [CLS] and [REG] tokens into account, HiVG’s vision sequence length is merely $(196 + 1)/(1600 + 2) = 12.29\%$ compared to that of TransVG++ (*i.e.*, TransVG++ is 8.13× larger than HiVG), demonstrating a dominant difference. Unlike the other works [2, 12], our framework can obtain state-of-the-art results without relying on high-resolution images. This significantly reduces the calculation complexity and greatly accelerates the training and reasoning computation of our HiVG framework.

Details of Figure 4-(d) of the Main Text. In the Figure 4-(d) of the main text, the legend for “original CLIP” represents that we only utilize the image and text encoder from vanilla CLIP as the backbone of our grounding framework while without using HiLoRA, cross-modal bridge, and RTCC constraint *etc.*. Besides, it only uses the final layer of visual and text features for the grounding encoder. The legend for “CLIP w. vanilla LoRA” represents that we additionally utilize the vanilla LoRA when compare to the legend for “original CLIP”. The legend for “HiVG w/o HiLoRA & MACB” represents that our HiVG framework does not utilize the main module of HiLoRA and MACB but utilize the RTCC constraint and multi-level visual features. The legend for “HiVG w/o HiLoRA” represents that our HiVG framework without utilizing HiLoRA but utilizing MACB, RTCC constraint and multi-level visual features. The legend for “HiVG w. vanilla LoRA” represents that our HiVG framework uses vanilla LoRA along with MACB, RTCC constraint and multi-level visual features. The legend for “HiVG w. HiLoRA stage 1, 2, 3” represents our full model under the three stages of HiLoRA.

4 EXTRA ABLATION STUDY

Ablation Study of Training Loss. As presented in Tab. 4, we extend the Table 3 of the main text, which serves as our ablation study for the two framework constraints. After the training of HiLoRA, we observed that without the contrastive learning constraint, the performance sometimes starts to degrade or even catastrophically forgets after reaching a certain level of training accuracy. It can be seen from the table that CLC enhances stability during HiLoRA training. Additionally, since RTCC is a token-wise constraint on the aggregated multi-level visual features, it enables a more fine-grained perception of these features.

More Detailed Ablation Results on MACB. In Table 4 of the main text, the weights in the table denotes the sample-agnostic weights. In the line 7 of Table 4 of the main text, “layer-to-layer linear connect” represents direct connect the corresponding layer of the image and text encoder by a MLP and a cross-attention module. In the line 8 of Table 4 of the main text, “only last layer of text features” represents only utilizing the last layer of text features with our multi-layer adaptive cross-modal bridge, and the shape of



Figure 3: Additional qualitative results of our HiVG framework on the RefCOCOg-val split. The CLIP-VG model is compared. We present the prediction box with IoU (in cyan) and the ground truth box (in green) in a unified image to visually display the grounding accuracy. We show the [REG] token’s attention over vision tokens from the last grounding block of each framework. The examples exhibit the relatively more challenging instances for grounding, thereby showcasing HiVG’s robust semantic comprehension capabilities.

the sample-agnostic weights are $1 \times L_l \times H_l$. As shown in Tab. 5, we provide more ablation study on the implementation of the multi-layer adaptive cross-modal bridge. In the line 1 of Tab. 5, “sample-aware weights” represents that we replace the sample-agnostic weights with a MLP structure, while also utilizing the 1^{th} - 12^{th} layer of text features. The table shows that our designed structure can effectively select the multi-level text features and can achieve the best performance when utilizing all the 12 layer of text feature.

Ablation Study of Multi-level Visual Features. We perform an ablation study on the utilization of multi-level visual features. We conduct the ablation study on the HiVG model without utilizing all the MACB, HiLoRA, and RTCC methods. As observed from Tab. 6, any approach that incorporates the intermediate layer of visual features outperforms solely relying on the final layer features. This confirms that some lower-level useful visual information may be discarded in the final layer, which is crucial for grounding tasks. It demonstrates that employing features from layers 1, 4, 8, and 12 yields the most favorable results.

5 ADDITIONAL QUALITATIVE RESULTS

As shown in Fig. 3, we present the grounding qualitative results with several additional challenging examples. All these results demonstrate the strong capability of our HiVG model in complex text understanding and cross-modal grounding.

6 FUTURE WORK

In the future, as a task-agnostic hierarchical adaptation paradigm, Hi LoRA can be further investigated across diverse downstream transfer scenarios. In this paper, we only explore the implementation of a simple progressive version. Additionally, there should be further research on the settings of layer groups and LoRA stages, such as exploring the adaptive selection of the both. Finally, it is also important to explore the broader application of hierarchical LoRA for visual, linguistic, and cross-modal tasks beyond grounding and detection tasks.

REFERENCES

- [1] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. 2021. TransVG: End-to-End Visual Grounding with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1769–1779.
- [2] Jiajun Deng, Zhengyuan Yang, Daqing Liu, Tianlang Chen, Wengang Zhou, Yanyong Zhang, Houqiang Li, and Wanli Ouyang. 2023. Transvg++: End-to-end visual grounding with language conditioned vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
- [3] Hugo Jair Escalante, Carlos A Hernández, Jesus A Gonzalez, Aurelio López-López, Manuel Montes, Eduardo F Morales, L Enrique Sucar, Luis Villaseñor, and Michael Grubinger. 2010. The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding (CVIU)* 114 (2010), 419–428.
- [4] Chih-Hui Ho, Srikanth Appalaraju, Bhavan Jasani, R Manmatha, and Nuno Vasconcelos. 2023. YORO-Lightweight End to End Visual Grounding. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*. Springer, 3–23.
- [5] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 787–798.
- [6] Muchen Li and Leonid Sigal. 2021. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in Neural Information Processing Systems* 34 (2021), 19652–19664.
- [7] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. 2020. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.
- [9] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [10] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2641–2649.
- [11] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* (2021).
- [12] Wei Su, Peihan Miao, Huanzhang Dou, Gaoang Wang, Liang Qiao, Zheyang Li, and Xi Li. 2023. Language adaptive weight generation for multi-task visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10857–10866.
- [13] Linhui Xiao, Xiaoshan Yang, Fang Peng, Ming Yan, Yaowei Wang, and Changsheng Xu. 2023. CLIP-VG: Self-paced Curriculum Adapting of CLIP for Visual Grounding. *IEEE Transactions on Multimedia* (2023).
- [14] Sibe Yang, Guanbin Li, and Yizhou Yu. 2019. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4644–4653.
- [15] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. 2020. Improving one-stage visual grounding by recursive sub-query construction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 387–404.
- [16] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019. A fast and accurate one-stage approach to visual grounding. In *ICCV*. 4683–4693.
- [17] Jiabo Ye, Junfeng Tian, Ming Yan, Xiaoshan Yang, Xuwu Wang, Ji Zhang, Liang He, and Xin Lin. 2022. Shifting More Attention to Visual Backbone: Query-modulated Refinement Networks for End-to-End Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15502–15512.
- [18] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. MATTNet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1307–1315.
- [19] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*. Springer, 69–85.