Bilateral Collaboration with Large Vision-Language Models for Open Vocabulary Human-Object Interaction Detection

Supplementary Material

This supplementary material includes five sections. Section A provides a more comprehensive performance comparison in the closed setting. Section B presents our prompt for MLLM to generate captions. Section C shows the visualization of attention maps. Section D offers a more thorough ablation study on the token weights. More implementation details of the baseline and our model (BC-HOI) are presented in Section E and Section F, respectively. Section G shows the implementation details of "EF Only" and "EF+LSG" settings in the ablation studies, while Section H demonstrates the implementation of token-level supervision. Additional ablation studies for "EF+LSG" are presented in Section I.

A. More Performance Comparisons

In this section, we compare our method with more existing approaches on HICO-DET [**?**], as shown in Table A. To better demonstrate the effectiveness of our method, we provide additional experimental results in the Known-Object (KO) setting. It is shown that our method achieves state-of-the-art performance in both the Default and KO settings. Specifically, in the KO setting, our method outperforms the second-best approach by 3.15%, 5.34%, and 2.49% in the Full, Seen, and Unseen categories, respectively. These experimental results fully demonstrate our method's strong modeling capability for both rare and non-rare HOI categories. The analysis of experimental results on the Default setting is already presented in Section **??**. Similarly, we provide the performance comparison with more existing methods on V-COCO [**?**], as shown in Table B.

B. Prompts of MLLM for Captioning

In this section, we demonstrate how we prompt MLLM to generate captions with rich image context and locality information. Specifically, we first provide GPT-4o [**?**] with a detailed description of the HOI task to obtain a basic prompt, which is then manually refined. Second, we enhance the prompt to enable MLLM to generate captions with locality information using the modified prompt. Additionally, we offer an example in the prompt so as to explain how to describe multi-person scenarios (e.g., using 'a group of'). The final prompt is illustrated in Figure A.

Task: Given an image, identify the human-object interaction triplets present in the scene. Each triplet should include a human, an object, and the action being performed.
Instructions: Analyze the Image: Carefully observe all entities in the image, focusing on both humans and objects.
Identify Interactions: For each interaction you see, focus on the following components:
1.Human: the individual or group performing the interaction, describe their position in the image.
2. Object: the object being interacted with, identify the object class.
3. Action: Specify the action that the human is performing with the object.
Format Your Response: For each human, present your findings in consistent sentences without listing: Detailed descriptions of the human's position and interaction with the object(s) in several sentences, capturing the context involved.
Examples: The man on the left side of the image is holding a leash and walking a dog along the sidewalk. A group of people on the center of the image are sitting on the chairs.

Figure A. The prompt we adopted to instruct the MLLM to caption images in the LSG component.

C. Visualization of Attention Maps

To validate the effectiveness of our proposed method, we visualize the attention maps from our interaction decoder, our ViT encoder, and the vanilla ViT encoder of BLIP-2, as shown in Figure B. Specifically, we train our model in the Default setting of HICO-DET [**?**], and extract the cross-attention map for one selected HOI query from the interaction decoder. Then, we obtain the self-attention map of the corresponding cls_token embedding in $\boldsymbol{C}_{\text{ho}}$, and the self-attention map of the cls_token embedding in the vanilla ViT encoder of BLIP-2. It is shown that the HOI query and the cls_token embedding in $\boldsymbol{C}_{\text{ho}}$ can effectively focus on the interaction area of a specific human-object pair in the image. In comparison, the cls_token embedding in the vanilla ViT encoder of BLIP-2 attend to all the foreground objects in the image. This demonstrates that our HOI detector provides high-quality attention bias via ABG, resulting in a significant HOI detection performance improvement.

D. Study on the Token Weights

In this section, we further conduct more experiments on the token weights w_n in Eq.**??** of LSG. Experiments are conducted on the NF-UC setting of the HICO-DET database. The experimental results are presented in Tables C. It is shown that the combination of $\alpha = 1.5$, $\beta = 2$, and $else = 1$ performs the best. We assert that it is necessary to impose a different loss weight to each token according to its part-of-speech, so that the model can focus more on HOIs contained in the caption.

Table A. Performance comparisons on HICO-DET in the Default setting and Known-Object (KO) setting. **Bold** represents the best performance, and underline indicates the second-best performance.

	Method	Backbone	Default Setting			Known Object		
			Full	Rare	Non-Rare	Full	Rare	Non-Rare
Two-Stage	ATL [**?**]	ResNet-50	23.81	17.43	25.72	27.38	22.09	28.96
	VSGNet [**?**]	ResNet-152	19.80	16.05	20.91	-	-	-
	DJ-RN [**?**]	ResNet-50	21.34	18.53	22.18	23.69	20.64	24.60
	VCL [**?**]	ResNet-101	23.63	17.21	25.55	25.98	19.12	28.03
	DRG [**?**]	ResNet-50-FPN	24.53	19.47	26.04	27.98	23.11	29.43
	IDN [**?**]	ResNet-50	26.29	22.61	27.39	28.24	24.47	29.37
	UPT [**?**]	ResNet-50	31.66	25.94	33.36	35.05	29.27	36.77
	CLIP4HOI [**?**]	ResNet-50	35.33	33.95	35.74	37.19	35.27	37.77
	CMMP (w/ ViT-L) [**?**]	ResNet-50	38.14	37.75	38.25	-	-	-
	ViPLO [**?**]	ViT-B/16	37.22	35.45	37.75	40.61	38.82	41.15
	EZ-HOI [**?**]	ResNet-50	38.61	37.70	38.89	-	-	-
	BCOM [**?**]	ResNet-50	39.34	39.90	39.17	42.24	42.86	42.05
One-Stage	PPDM [**?**]	Hourglass-104	21.73	13.78	24.10	24.58	16.65	26.84
	HOI-Trans [**?**]	ResNet-101	26.61	19.15	28.84	29.13	20.98	31.57
	AS-Net [**?**]	ResNet-50	28.87	24.25	30.25	31.74	27.07	33.14
	QPIC [**?**]	ResNet-101	29.90	23.92	31.69	32.38	26.06	34.27
	CDN [**?**]	ResNet-50	31.44	27.39	32.64	34.09	29.63	35.42
	DOQ [**?**]	ResNet-50	33.28	29.19	34.50	-	-	-
	GEN-VLKT [**?**]	ResNet-50	33.75	29.25	35.10	36.78	32.75	37.99
	MP-HOI [**?**]	ResNet-50	36.50	35.48	36.80	-	-	-
	HOICLIP [**?**]	ResNet-50	34.69	31.12	35.74	37.61	34.47	38.54
	RLIPv2-ParSeDA [**?**]	ResNet-50	35.38	29.61	37.10	-	-	-
	QAHOI [**?**]	Swin-L	35.78	29.80	37.56	37.59	31.66	39.36
	DP-HOI[**?**]	ResNet-50	36.56	34.36	37.22	39.37	36.59	40.20
	FGAHOI [**?**]	Swin-L	37.18	30.71	39.11	38.93	31.93	41.02
	UniHOI (w/ BLIP-2) [**?**]	ResNet-50	40.06	39.91	40.11	42.20	42.60	42.08
	Ours (w/ BLIP-2)	ResNet-50	**43.01**	**45.76**	**42.18**	**45.35**	**47.94**	**44.57**

Table B. Comparisons on V-COCO in the closed setting.

	Method	$mAP^{\#1}_{role}$	$mAP^{\#2}_{role}$
Two-Stage	VCL [**?**]	48.3	-
	DRG [**?**]	51.0	-
	VSGNet [**?**]	51.8	57.0
	IDN [**?**]	53.3	60.3
	UPT [**?**]	59.0	64.5
	CLIP4HOI [**?**]	-	66.3
	CMMP (w/ ViT-L) [**?**]	-	64.0
	VIPLO [**?**]	62.2	68.0
	EZ-HOI [**?**]	60.5	66.2
	BCOM [**?**]	65.8	69.9
One-Stage	HOI-Trans [**?**]	52.9	-
	AS-Net [**?**]	53.9	-
	HOTR [**?**]	55.2	64.4
	QPIC [**?**]	58.8	61.0
	CDN [**?**]	61.68	63.77
	GEN-VLKT [**?**]	62.41	64.46
	RLIPv2-ParSeDA [**?**]	65.9	68.0
	HOICLIP [**?**]	63.50	64.80
	FGAHOI (Swin-T) [**?**]	60.5	61.2
	MP-HOI [**?**]	66.2	67.6
	DP-HOI [**?**]	66.6	-
	UniHOI (w/ BLIP2) [**?**]	65.58	68.27
	Ours (w/ BLIP2)	**68.20**	**70.61**

Table C. Study on the value of token weight w_n

α	β	$else$	Unseen	Seen	Full
1	1	0	32.47	36.56	35.74
1	1	1	32.79	36.98	36.14
2	2	1	32.89	37.20	36.37
2	1.5	1	32.82	37.07	36.25
1.5	2	1	**33.01**	**37.24**	**36.40**

E. More Details of Baseline

We illustrate the structure of the baseline model in Figure C. To construct the baseline model, we remove the EF, ABG, and LSG components from BC-HOI. The same as BC-HOI, the baseline adopts one transformer layer to fuse the visual features produced by BLIP-2's Vision Tower and the interaction features produced by the HOI detector. Specifically, it adopts $\boldsymbol{E}_{\text{ho}}$ as the query, while the output of Q-Former serve as the key and value, respectively. Besides, the baseline adopts the same classifier as BC-HOI.

F. More Details of BC-HOI

In BC-HOI, the BLIP-2 OPT-2.7B [**?**] is adopted as the large VLM for all experiments following UniHOI [**?**]. Specifically, each image is first down-sampled to 224×224 pixels and then being divided into 16×16 patches. These patches are concatenated with $\boldsymbol{C}$, $\boldsymbol{C}_{\text{f}}$, and $\boldsymbol{C}_{\text{ho}}$ as the input of the ViT encoder, producing embeddings with the dimension of $(1 + 16 \times 16 + 64 + 32) \times 1408$. The output embeddings from the ViT encoder are then fed into the Q-Former

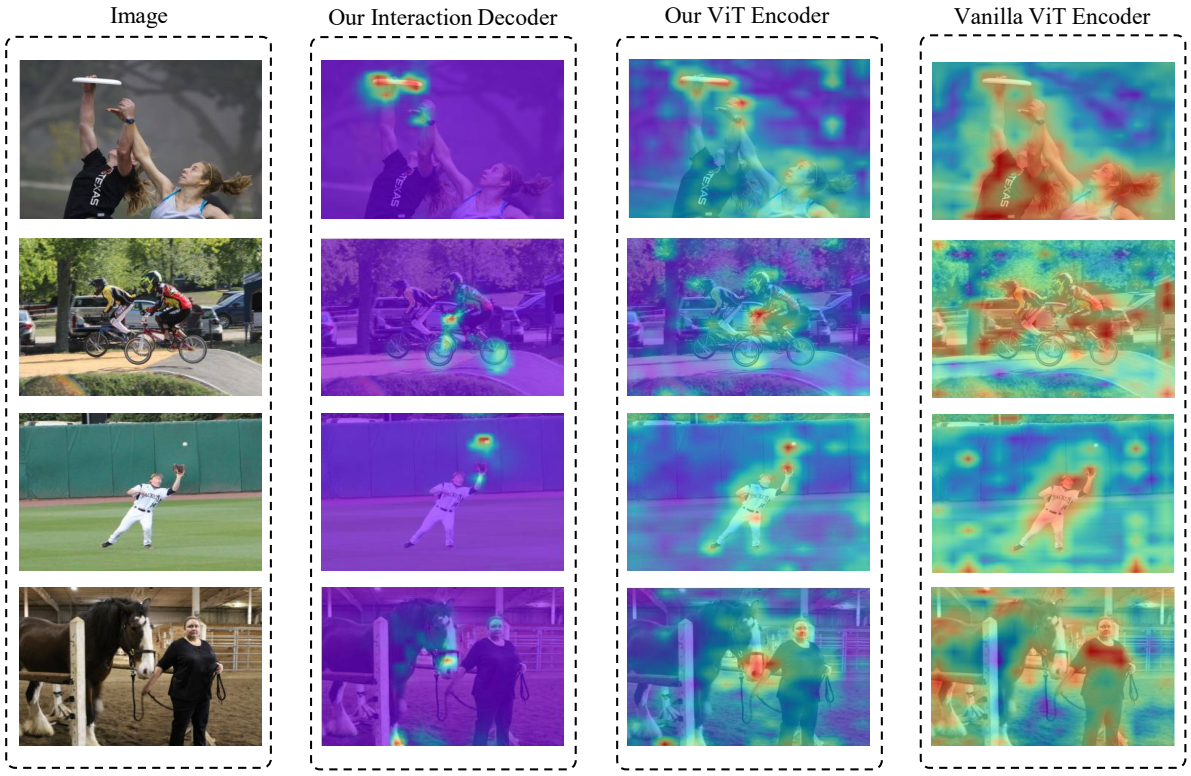

Figure B. Visualization of the attention maps. Each row demonstrates the original image, the cross-attention map for one selected HOI query in our interaction decoder, the self-attention map of the corresponding cls_token embedding in $\boldsymbol{C}_{\text{ho}}$, and the self-attention map of the cls_token embedding in the vanilla ViT encoder of BLIP-2. It is shown that with the guidance by ABG, the self-attention map produced by the cls_token embedding in $\boldsymbol{C}_{\text{ho}}$ effectively focuses on the interaction areas of an interested human-object pair. In contrast, the self-attention map produced by the vanilla ViT encoder of BLIP-2 attend to all foreground objects in the image.

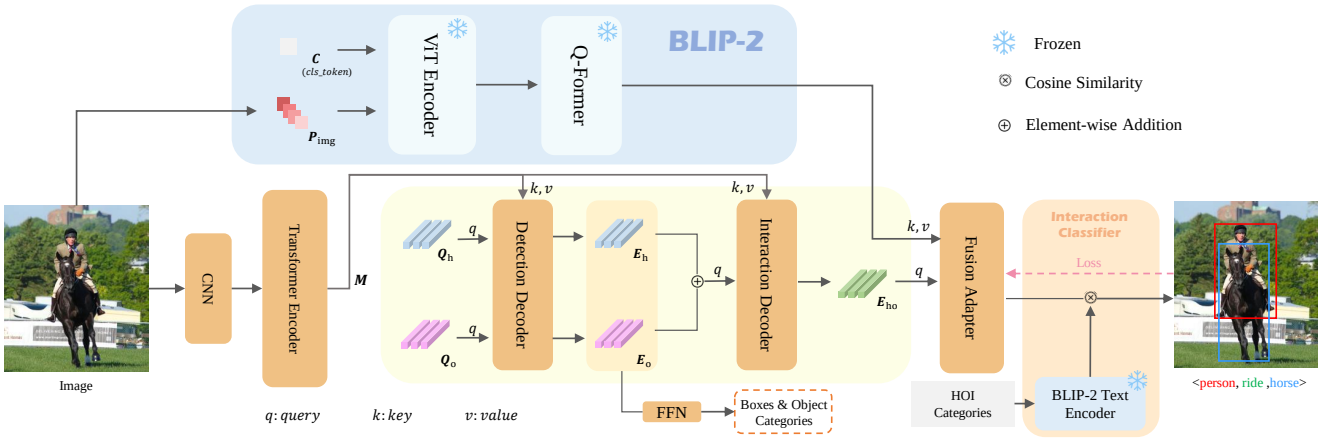

Figure C. Model structure of the baseline model.

as keys and values of the cross-attention layers, while $\boldsymbol{Q}^{\text{q}}_{\text{ho}}$ and $\boldsymbol{Q}^{\text{q}}_{\text{f}}$ serve as the queries with dimensions of 64×768 and 32×768, respectively. The output features $\boldsymbol{E}^{\text{q}}_{\text{ho}}$ from the Q-Former are fused with the features extracted by the HOI detector via element-wise addition. Another set of output features $\boldsymbol{E}^{\text{q}}_{\text{f}}$ from the Q-Former is both fed into the OPT-2.7B model for image captioning during training and into the Fusion Adapter as keys and values. Moreover, we adopt "a photo of a person <doing something>" as template to construct a phrase for each HOI category. This phrase is

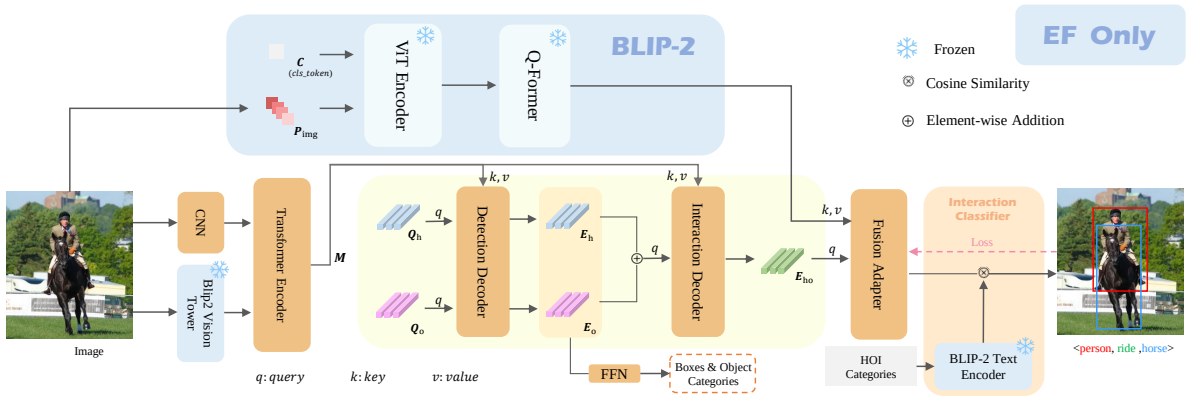

Figure D. Model structure of the "EF Only" setting in the ablation study.

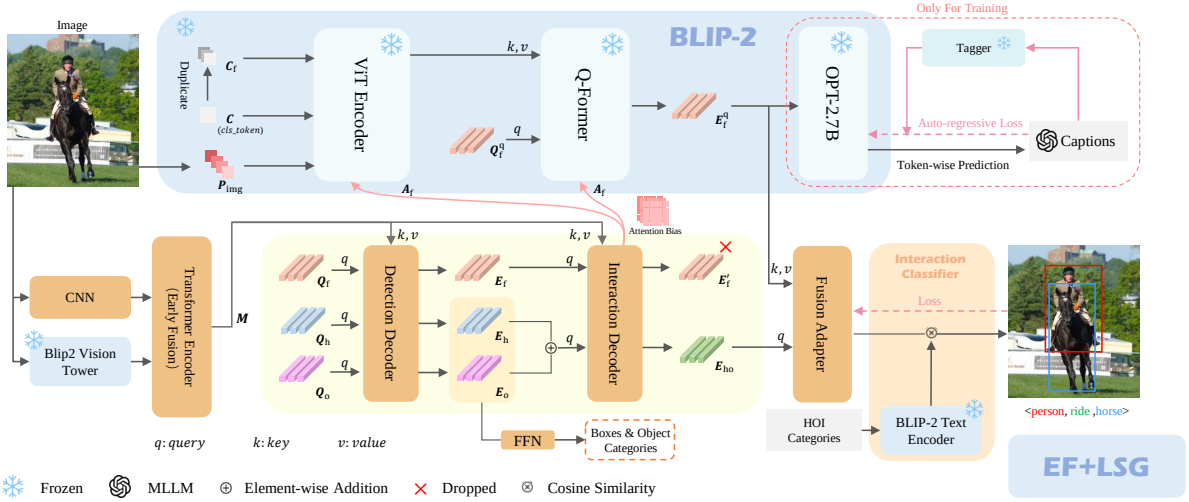

Figure E. Model structure of the "EF+LSG" setting in the ablation study.

then fed into the BLIP-2 ViT-L text encoder to obtain the interaction classifier.

G. More Details of "EF Only" and "EF+LSG"

For clarity, we illustrate the structure of "EF Only" and "EF+LSG" of the ablation studies in Figures D and E, respectively. In the "EF Only" setting, we add EF (Early Fusion) to the baseline, which means the branch of BLIP-2 Vision Tower remains in the model. In the "EF+LSG" setting, we remove $\boldsymbol{C}_{\text{ho}}$, $\boldsymbol{Q}^{\text{q}}_{\text{ho}}$, $\boldsymbol{A}_{\text{ho}}$, and $\boldsymbol{E}^{\text{q}}_{\text{ho}}$ from Figure 3. Although all cls_tokens ($\boldsymbol{C}$ and $\boldsymbol{C}_{\text{f}}$) attend to the same features ($\boldsymbol{P}_{\text{img}}$), their attention maps are affected by different bias values within $\boldsymbol{A}_{\text{f}}$.

H. More Details of token-level supervision.

As illustrated in Figure F, in the LSG framework, we feed the embeddings $\boldsymbol{E}^{\text{q}}_{\text{f}}$ into BLIP-2's LLM component, prompting it to generate fine-grained captions in an auto-regressive manner. This produces token-level supervision, enabling the model to produce high-quality attention maps. Notably, the code of the auto-regressive loss is highly mature, and computations for each token are processed in parallel, ensuring sufficient efficiency.

Table D. Study on "EF+LSG" setting in the NF-UC setting

method	Unseen	Seen	Full
EF+LSG w/o ($\boldsymbol{A}_{\text{f}}$ and LLM)	30.38	32.71	32.10
EF+LSG w/o $\boldsymbol{A}_{\text{f}}$	30.62	33.10	32.57
EF+LSG	31.37	33.84	33.38

I. Study on "EF+LSG" setting

To demonstrate the novelty and effectiveness of our LSG method, we provide additional ablation studies in the "EF+LSG" setting, as shown in the Table D. In the "EF+LSG w/o $\boldsymbol{A}_{\text{f}}$" setting, we remove the in-

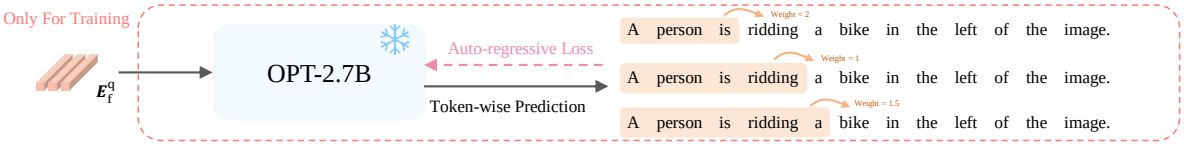

Figure F. Implementation of the token-level supervision.

fluence of $\boldsymbol{A}_{\mathrm{f}}$, and the model performance drops to 30.62%/33.10%/32.57% on Unseen/Seen/Full in the UC-NF setting. This demonstrates that LSG enhances the performance by optimizing the attention maps generated by the model. In the "EF+LSG w/o ($\boldsymbol{A}_{\mathrm{f}}$ and LLM)" setting, we remove both $\boldsymbol{A}_{\mathrm{f}}$ and LLM from "EF+LSG" and retain the learnable $\boldsymbol{Q}_{\mathrm{f}}^{\mathrm{q}}$. Compared with the performance of "EF+LSG", the performance drops by 0.99%/1.13%/1.28% on Unseen/Seen/Full in the UC-NF setting. This means LSG mainly benefits from its LLM supervision instead of refined-query $\boldsymbol{Q}_{\mathrm{f}}^{\mathrm{q}}$, which requires the HOI detector to provide high-quality attention bias $\boldsymbol{A}_{\mathrm{f}}$.