HOLa: Zero-Shot HOI Detection with Low-Rank Decomposed VLM Feature Adaptation

Supplementary Materials

Qinqian Lei[1] Bo Wang[2] Robby T. Tan[1,3]
[1]National University of Singapore [2]University of Mississippi
[3]ASUS Intelligent Cloud Services (AICS)
qinqian.lei@u.nus.edu hawk.rsrch@gmail.com robby_tan@asus.com

1. Vision Branch

The detailed vision branch is illustrated in Fig. 1. We first leverage a pre-trained DETR [2] model for object detection, identifying all detected humans and objects h_i, o_j, where $i \in \{1, 2, \cdots, n_h\}, j \in \{1, 2, \cdots, n_o\}$. Here, n_h and n_o represent the total number of detected humans and objects, respectively. For each detection, we extract human and object features from the DETR decoder, denoted as f_{h_i} and f_{o_j}. We then generate all possible human-object feature pairs, represented as (f_{h_i}, f_{o_j}). The human-object tokens are computed as :

$$T_{\text{ho}_{ij}} = \frac{f_{h_i} + f_{o_j}}{2} + f^{\text{spatial}}_{\text{ho}_{ij}}, \tag{1}$$

where $f^{\text{spatial}}_{\text{ho}_{ij}}$ is derived from human-object bounding boxes, incorporating the center coordinates, width, height of each box, pairwise intersection-over-union (IoU), and relative area, which are then processed through an MLP together to obtain $f^{\text{spatial}}_{\text{ho}_{ij}}$. Thus, $T_{\text{ho}_{ij}}$ integrates both appearance and spatial cues to enhance interaction representation for each human-object feature pair (f_{h_i}, f_{o_j}). The complete set of human-object tokens is denoted as T_{ho}, where $T_{\text{ho}} = \{T_{\text{ho}_{ij}} \mid 1 \leq i \leq n_h, 1 \leq j \leq n_o\}$.

To further incorporate interaction prior knowledge, we leverage an LLM to generate descriptions of human body configurations, object attributes, and their spatial relationships with humans. These descriptions are then encoded by the VLM text encoder to obtain prior knowledge features $f^{\text{pr}}_{\text{ho}}$. An example of the generated descriptions used to capture human-object interaction prior knowledge is provided at the end of this section.

To integrate prior-knowledge features $f^{\text{pr}}_{\text{ho}}$ with human-object tokens, we design a cross-attention module: First, down and up projection layers are used to reduce computational cost. Next, human-object tokens serve as the query, while prior-knowledge features $f^{\text{pr}}_{\text{ho}}$ act as the key and value in the cross-attention mechanism. Finally, a residual connection adds the input human-object tokens back to the cross-attention output, refining interaction representation while preserving the original information.

The output human-object tokens are concatenated with input image patches and fed into the VLM visual encoder, guiding it to focus on human-object interactions and improving action distinction. To enhance adaptability, we insert the adapter [9] between each layer of the visual encoder. The output includes adapted human-object tokens $\hat{T}_{\text{ho}}$ and an image feature map $f^{\text{glb}}_{\text{img}} \in \mathcal{R}^{H \times W \times d}$. The final HOI image feature, denoted as $f_{\text{img}} \in \mathcal{R}^{d \times 1}$ and used by the weight adaptation, is defined as follows:

$$f_{\text{img}} = \frac{1}{H * W} \sum_{i=1}^{H} \sum_{j=1}^{W} f^{\text{glb}}_{\text{img}}(i, j, :). \tag{2}$$

We denote the detected bounding boxes for the human, object, and their union regions as b_h, b_o, and b_u, respectively. To extract features focused on specific human-object interaction regions within the image, we first apply RoI pooling to obtain region-specific features of dimension $p \times p \times d$, where p is set to 7. We then apply spatial average pooling to each region-specific feature to obtain $f^h_{\text{img}}, f^o_{\text{img}}, f^u_{\text{img}}$.

The image fusion module is designed to combine the human and object features f^h_{img} and f^o_{img}, respectively. The image fusion process takes the $\text{concat}(f^h_{\text{img}}, f^o_{\text{img}})$ as input and outputs $f^{\text{ho}}_{\text{img}}$. Here, concat denotes concatenation of two features along the first dimension. To reduce computational cost, the image fusion module incorporates down and up projection layers. The concatenated input features then pass through a self-attention module, integrating action and object visual features. Finally, a residual connection adds the input back to the output, refining the fusion while preserving input information.

According to Eq.(8) of the main paper, we compute the action prediction s_a. For convenience, we reproduce the

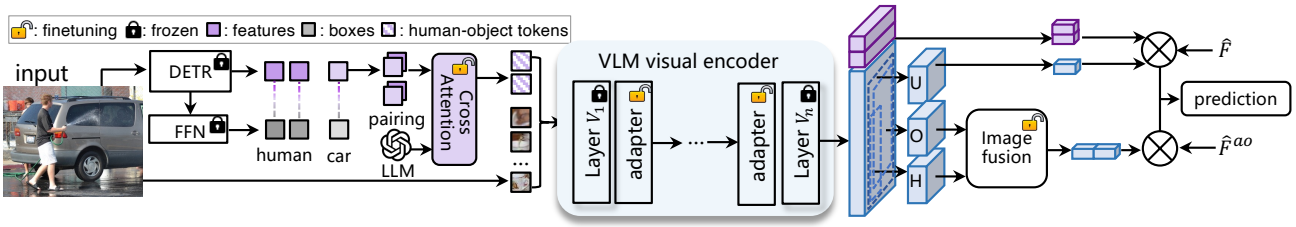

Figure 1. Overview of our vision branch.

equation below:

$$s_a = \gamma_1 * (\text{sim}(f^u_{\text{img}}, \hat{\mathbf{F}}) + \text{sim}(\hat{T}_{\text{ho}}, \hat{\mathbf{F}})) * l^R_u + \gamma_2 * \text{sim}(f^{\text{ho}}_{\text{img}}, \hat{\mathbf{F}}^{\mathbf{ao}}) * l^R_{ao}. \quad (3)$$

Here is an example prompt used with an LLM to generate prior knowledge for the human-object pair ⟨ human, car ⟩.

Provide a detailed description of the physical relationship between a given human-object pair, focusing on various possible configurations and spatial relationships without assuming or naming specific interactions. For the pair (human, car), describing the following perspectives:

*1. **Human Body Description:** - Describe the positioning and orientation of key body parts (e.g., hands, feet, arms, legs, torso, head) in relation to the object. - Highlight the possible roles of specific body parts (e.g., hands gripping, feet pressing, or knees bending) without specifying actions.*

*2. **Object Description:** - Provide a clear and concise description of the object, focusing on its relative size, shape, and structure compared to the human in the image. - Include details about key components (e.g., wheels, deck from skateboard) and their spatial relationship to the human body in various scenarios. - Highlight how the object might be positioned (e.g., sliced, tilted, vertical) and how its components could interact with specific body parts of the human.*

The goal is to provide a comprehensive pool of descriptive information for the human, object, and their possible configurations in various scenarios. Avoid limiting the scope by naming specific interactions or actions. Focus instead on a rich and versatile set of physical relationships. Focus on critical details, avoiding redundant or non-essential information to ensure clarity and precision.

The following is the prior knowledge generated by the LLM for the human–car pair:

1. Human Body Description: (a) Positioning and Orientation: - Head : The human's head can be oriented in various directions relative to the car, such as facing towards the car, away from it, or to the sides. The head's angle might vary, possibly tilted up or down, depending on the human's position relative to the car's height, such as looking over the roof or under the chassis. - Torso : The torso may be upright, leaning forward, or angled sideways. Its position could vary significantly depending on proximity to the car, such as standing next to it, bending over the hood, or leaning against the side. - Arms and Hands : The arms might be extended towards the car or resting at the sides. Hands can be positioned near various parts of the car, such as the door handles, mirrors, or hood, suggesting a potential for contact or proximity. - Legs and Feet : Legs could be positioned straight, bent, or spread apart, depending on the stance relative to the car. Feet might be placed firmly on the ground, perhaps transitioning from one side of the car to the other, or positioned close to the wheels or undercarriage. (b) Roles of Specific Body Parts: - Hands : The hands might appear poised to make contact with the car's surface, potentially hovering over or near tactile features like handles or mirrors. - Feet : The feet might be aligned parallel to the car or angled towards it, suggesting readiness for movement or balance. - Knees : Knees might be straight or slightly bent, indicating a stance that accommodates reaching or observing different parts of the car.

2. Object Description: (a) Size, Shape, and Structure: - General Structure : The car is a relatively large object compared to a human, with a defined structure typically including a body, wheels, and windows. It is generally elongated with a front (hood), middle (cabin), and rear (trunk). - Key Components : - Wheels : Typically four, located at the corners of the car, often near the ground and could be aligned with the human's feet or legs. - Body Panels : Including doors, hood, trunk, and roof, which vary in height and shape, potentially aligning with the human's torso or head. - Windows and Mirrors : Positioned at varying heights, likely matching the human's eye level or above, affecting the head's orientation.

2. Language Branch

We leverage the LLM to generate descriptions for both HOI and action classes in the language branch.

HOI Description Generation The descriptions for all given HOI classes are utilized to obtain the VLM text features $\mathbf{F}$ through the VLM text encoder, as discussed in Section 3.1 of the main paper. For a detailed illustration, con-

sider the HOI class "hosing a car" as an example. The description generated for this HOI class is as follows:

"Hosing a car" description: The photo shows a person using a hose to clean or wash a car. Common elements include water, soap, cleaning tools, and the person's attention to various parts of the car's exterior.

Action Description Generation Additionally, we introduce action descriptions for all given action classes to obtain action regularization, as discussed in Section 3.2 of the main paper.

Here is an example prompt provided to the LLM to generate a description for the action "hose".

Please describe the visual features for action: hose. The objects can be "car, dog, potted plant, elephant". Please summarize the common visual features for the action, shared by the objects. Use LESS than 250 characters for description. "

The following is the action description generated by the LLM for "hose":

The visual features for the action "a person is hosing the object" include: 1. Person holding a hose or water spray nozzle. 2. Object being hosed (e.g. car, dog, potted plant, elephant) in close proximity to the person. 3. Water spray or stream visible, potentially creating a mist or splash. 4. Person's arm or hand in motion, indicating the action of spraying water. 5. Background context suggesting an outdoor or yard setting.

3. Experiment Results

3.1. Implementation Details

In weight adaptation, image and text fusion and human-object tokens design, all down projection layers reduce feature dimension from 512 to 64, while all up projection layers expands it back from 64 to 512. The head number of self attention and cross attention modules is 2. The temperature of KL divergence used in our method is 0.1. We use AdamW [16] as the optimizer and the initial learning rate is 1e-3. For all experiments, our batch size is set as 64 on 4 A5000 GPUs. Training takes 7 hours on 4 A5000 GPUs (22.5 GB VRAM each) with only 4.0M trainable parameters. Inference time is 82 ms per image.

We use three types of descriptions generated by foundation models: (1) HOI class descriptions from EZ-HOI [8], generated using LLaVA [15]. These descriptions are encoded by a VLM text encoder to produce $\mathbf{F}$, as described in Section 3.1 (VLM Feature Decomposition and Adaptation) of the main paper. An example is also included in the HOI Description Generation subsection of the language branch (Sec. 2); (2) action descriptions for LLM-derived action regularization, generated using the LLaMA-3-8B model [4], and used in the language branch (Sec. 2); and (3) Prior knowledge descriptions for human-object pairs, also generated by LLaMA-3-8B, and used in the vision branch (Sec. 1).

Method	HICO-DET		
	Full	Rare	Nonrare
One-stage Methods			
GEN-VLKT (CVPR'22) [13]	33.75	29.25	35.10
EoID (AAAI'23) [19]	31.11	26.49	32.49
HOICLIP (CVPR'23) [18]	34.69	31.12	35.74
LogicHOI (NeurIPS'23) [12]	35.47	32.03	36.22
UniHOI (NeurIPS'23) [1]	**35.92**	**34.39**	**36.26**
Two-stage Methods			
FCL (CVPR'21) [7]	29.12	23.67	30.75
ATL (CVPR'21) [6]	23.81	17.43	25.72
ADA-CM [9] (ICCV'23)	33.80	31.72	34.42
CLIP4HOI (NeurIPS'23) [17]	35.33	33.95	**35.75**
CMMP (ECCV'24) [11]	33.24	32.26	33.53
Ours (HOLa)	**35.41**	**34.35**	35.73
ADA-CM$_l$ (ICCV'23)	38.40	37.52	38.66
CMMP$_l$ (ECCV'24)	38.14	37.75	38.25
EZ-HOI$_l$ (NeurIPS'24)	38.61	37.70	38.89
Ours$_l$ (HOLa)	**39.05**	**38.66**	**39.17**

Table 1. Quantitative comparison of HOI detection with state-of-the-art methods in the fully-supervised setting on HICO-DET. ***Ours$_l$*** denotes our scaled-up version utilizing the ViT-L/14 backbone.

Datasets We evaluate our method on the HICO-DET dataset [3], a widely-used benchmark in human-object interaction detection. HICO-DET contains 47,776 images in total, consisting of 38,118 training images and 9,658 test images. The dataset includes 600 HOI classes combined from 117 action categories and 80 object categories. We also provided the evaluation on the V-COCO [14], a subset of COCO, comprises 10,396 images, with 5,400 trainval images and 4,946 test images, and includes 24 action classes and 80 object classes. Note that V-COCO only contains evaluation under fully-supervised setting, but our focus is on the zero-shot HOI detection.

3.2. Quantitative Results

Fully Supervised Setting on HICO-DET We evaluate our method against HOI approaches with zero-shot HOI detection ability in the fully supervised settings , excluding methods that do not support unseen-action HOI detection. Table 1 demonstrates that our method sets a new state-of-the-art performance on the HICO-DET dataset in the fully supervised setting. Using the ViT-B backbone, the same as those used in existing methods [1, 6, 7, 11–13, 17–19], our method achieves a 35.41 mAP, surpassing all state-of-the-art two-stage HOI detection methods. Switching to a ViT-L backbone further enhances performance, reach-

reconstruction score	rank	mAP		
		Unseen	Seen	Full
0.80	17	26.32	32.69	31.80
0.90	42	25.71	33.01	31.98
0.95	71	25.47	33.59	32.46
0.98	119	25.17	32.82	31.75

Table 2. Ablation study for the rank of basis features $\mathbf{B}$ and weights $\mathbf{W}$ in the unseen-verb zero-shot setting.

ing 39.05 mAP. Although primarily designed to focus on zero-shot HOI detection and improve generalization to unseen classes, our method also shows competitive results in the fully supervised setting, underscoring its effectiveness across diverse evaluation scenarios.

Fully Supervised Setting on V-COCO Our method also demonstrates competitive performance on the V-COCO dataset, achieving a 66.0 $AP^{S_2}_{role}$, achieving an improvement of 2.0 mAP over the current state-of-the-art method, CMMP [11]. Our $AP^{S_1}_{role} = 60.3$.

3.3. Ablation Study

Rank Selection for B and W We conduct an ablation study on the rank m of the basis features and weights, as shown in Table 2. This study specifically explores the impact of the selected rank m on the performance, focusing solely on the feature decomposition module. Consequently, other components, such as the action prior and the action-object branch, were excluded from this analysis.

We initialize the weights and basis features using Principal Component Analysis (PCA). Specifically, we achieve reconstruction percentages of 0.80, 0.90, 0.95, and 0.98 for the original VLM text features, $\mathbf{F}$. These percentages correspond to ranks of 17, 42, 71, and 119, respectively, in the obtained weights and basis features.

The evaluation results show that a rank 17 yields the highest unseen mAP (26.32), due to its compact representation that emphasizes class-shared features, enhancing generalization to unseen classes. However, this compactness leads to a drop in seen class performance, due to the loss of some detailed information from $\mathbf{F}$. Conversely, increasing the rank to 119 captures more class-specific details in the reconstructed features but diminishes the shared information across classes, leading to poorer unseen class performance. Consequently, we select the rank of 71 to optimally balance performance between seen and unseen classes.

VLM Feature Decomposition Constraints We conducted an ablation study on the constraints for VLM feature decomposition as shown in Table 3. The first row removes the orthogonal constraint L_{ort} on the basis features, leading to 1.10 mAP drop among unseen classes compared to the third row, indicating that the orthogonal constraint helps the basis features capture class-shared information more effectively, enhancing generalization to unseen classes. Additionally, removing the sparsity constraint L_{sparse} (second row) lowers both seen and unseen performance, indicating that sparsity reduces redundancy in the factorization, leading to a more compact representation.

L_{ort}	L_{sparse}	mAP		
		Unseen	Seen	Full
×	✓	26.81	34.70	33.60
✓	×	27.47	34.45	33.48
✓	✓	27.91	35.09	34.09

Table 3. Ablation study for VLM feature decomposition constraints L_{ort} and L_{sparse} in the unseen-verb zero-shot setting.

L_{sem}	mAP		
	Unseen	Seen	Full
×	27.19	34.68	33.63
✓	27.91	35.09	34.09

Table 4. Ablation study for semantic loss in the unseen-verb zero-shot setting.

Semantic Loss We also design the semantic loss L_{sem} to preserve the distribution of pairwise cosine similarity among VLM text feature of each class. The pairwise cosine similarity demonstrates the relationship between HOI classes indicated by VLM, which, trained on millions of data, generalizes these relationships to unseen classes. Unlike the original VLM features, which primarily emphasize object information and cluster different actions with the same object together, our method explicitly enhances action distinctions. To achieve this, we compute similarity only among HOI classes involving the same object, as shown in Eq.(4), with the mask M excluding interactions with different objects.

$$\begin{aligned} L_{\text{sem}} = \; & D_{\text{KL}}[\, \frac{\text{sim}(\hat{\mathbf{F}}, \hat{\mathbf{F}})}{\tau} * M \;\|\; \frac{\text{sim}(\mathbf{F}, \mathbf{F})}{\tau} * M \,] \\ & + D_{\text{KL}}[\frac{\text{sim}(\hat{\mathbf{F}}^{\mathbf{ao}}, \hat{\mathbf{F}}^{\mathbf{ao}})}{\tau} * M \;\|\; \frac{\text{sim}(\mathbf{F}, \mathbf{F})}{\tau} * M \,], \end{aligned} \tag{4}$$

where we apply a temperature coefficient τ in the KL divergence, setting $\tau = 0.1$ to emphasize action relationships that are underestimated in the original VLM features. As shown in Table 4, without L_{sem}, the overall performance in the unseen-verb setting decreases from 32.66 to 32.41 mAP, with a 0.48 mAP drop among unseen classes.

Human-Object Tokens Table 5 presents the ablation study on interaction prior knowledge generated by the LLM for human-object tokens f_{ho}. In the first row, we remove this prior knowledge and replace cross-attention with self-

LLM-generated Prior Knowledge	mAP		
	Unseen	Seen	Full
×	27.90	34.58	33.65
✓	27.91	35.09	34.09

Table 5. Ablation study for LLM description in human-object token design of the vision branch. "None" means no interaction prior knowledge generated from LLM.

$f_{\mathrm{ho}_{ij}}$	mAP		
	Unseen	Seen	Full
$\frac{f_{h_i}+f_{o_j}}{2}$	27.37	34.42	33.43
$f^{\mathrm{spatial}}_{\mathrm{ho}_{ij}}$	27.59	34.63	33.64
$\frac{f_{h_i}+f_{o_j}}{2} + f^{\mathrm{spatial}}_{\mathrm{ho}_{ij}}$	27.91	35.09	34.09

Table 6. Ablation study for human-object token design of the vision branch. "None" means no interaction prior knowledge generated from LLM.

attention process for f_{ho}. The results indicate that interaction prior knowledge primarily improves seen-class performance. This is because the interaction prior knowledge provides all possible human body configurations, object attributes and their spatial relationships. During training, the model is guided by training data to select knowledge mainly for seen HOI classes. Consequently, this interaction prior knowledge does not obviously enhance unseen HOI performance.

Table 6 shows the ablation study on the components of human-object tokens $f_{\mathrm{ho}_{ij}}$. As defined in Eq.(1), $f_{\mathrm{ho}_{ij}}$ consists of two components: human and object appearance features $\frac{f_{h_i}+f_{o_j}}{2}$ from DETR and the spatial features $f^{\mathrm{spatial}}_{\mathrm{ho}_{ij}}$ from detected human and object bounding boxes. We found that we need to combine all components in human-object tokens for the best performance among both seen and unseen classes according to the results shown in Table 6.

Image Fusion Table 7 presents the ablation study for the image fusion module. Removing this module reduces performance from 34.09 to 33.10 mAP, highlighting its effectiveness. The image fusion module adapts and integrates separate action and object visual features, capturing more fine-grained information than human-object union region features. While f^{h}_{img} , f^{o}_{img} , and f^{u}_{img} share the same feature dimension, f^{h}_{img} and f^{o}_{img} focus on smaller, localized regions—human and object separately, rather than their combined union. This processing better preserves action and object details, ultimately improving performance.

VLM Feature Decomposition and Adaptation Table 8 presents an ablation study on the VLM feature decomposition and adaptation. The first row serves as the baseline,

Image Fusion	mAP		
	Unseen	Seen	Full
×	26.46	34.19	33.10
✓	27.91	35.09	34.09

Table 7. Ablation study for image fusion design in the unseen-verb zero-shot setting.

$\mathbf{W}$	$\mathbf{B}$	mAP		
		Unseen	Seen	Full
/	/	23.58	31.55	30.43
$\overline{\mathbf{W}}$	$\overline{\mathbf{B}}$	25.76	31.35	30.57
$\overline{\mathbf{W}}$	$\mathbf{B}$	25.84	31.19	30.44
$\mathbf{W}$	$\mathbf{B}$	22.95	30.30	29.27
$\mathbf{W}$	$\overline{\mathbf{B}}$	25.47	33.59	32.46

Table 8. Ablation study for weights and basis features optimization in the unseen-verb zero-shot setting. $\mathbf{X}$ denotes applying classification loss L_{cls} and feature decomposition loss L_{fd} in training to update $\mathbf{X}$. $\overline{\mathbf{X}}$ denotes applying only L_{fd}. $\mathbf{X} \in \{\mathbf{W}, \mathbf{B}\}$.

where VLM feature decomposition is not applied, and no LLM-derived regularization are used. This baseline ensures that the ablation study specifically analyzes the impact of VLM feature decomposition. In the second row, we apply the only feature decomposition loss L_{fd} to update both weights and basis features ($\overline{\mathbf{W}}$, $\overline{\mathbf{B}}$). This improves unseen mAP by 2.18, indicating that feature decomposition enhances generalization to unseen classes. Applying classification loss L_{cls} only to the basis features ($\overline{\mathbf{W}}$, $\mathbf{B}$), as in the third row, yields results similar to the second row. In the fourth row, adding classification loss L_{cls}, supervised by ground truths from seen classes together with L_{fd} in the training process, to both weights and basis features results in performance degradation ($\mathbf{W}$, $\mathbf{B}$). This suggests that the updating of basis features from training data compromises essential class-shared information necessary for generalization, while the weights do not adapt effectively to distinguish actions within the HOI setting. The last row shows the best results, where L_{cls} is applied only to the weights, while L_{fd} is used for both weights and basis features ($\mathbf{W}$, $\overline{\mathbf{B}}$). This configuration achieves balanced performance across seen and unseen classes, improving the seen mAP by 2.04 and the unseen mAP by 1.89 compared to the baseline.

Weights for each training loss term Loss weights including $\alpha, \beta_1, \beta_2, \beta_3, \beta_4$ introduced in Section 3.4 of the main paper, are set to keep all loss terms on a comparable scale during early training, ensuring balanced contributions. Table 9 shows an ablation study where we vary one loss weight at a time while keeping the others fixed, where using comparable values for each loss term results in the best overall performance.

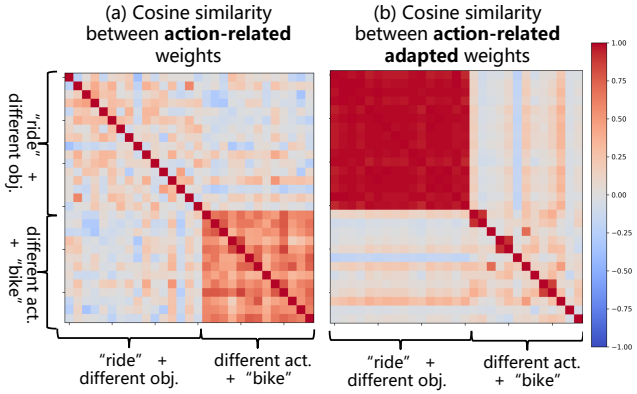

Figure 2. (a) Weight subset similarity visualization related to "ride" and "bike" HOI classes, **before weight adaptation**; (b) Adapted weight subset similarity visualization related to "ride" and "bike" HOI classes, **after weight adaptation**.

α	β_1	β_2	β_3	β_4	mAP		
					Unseen	Seen	Full
320	0.1	0.1	0.001	50	27.50	33.95	33.05
80	0.5	0.1	0.001	50	28.32	34.35	33.50
80	0.1	0.5	0.001	50	28.81	34.64	33.82
80	0.1	0.1	0.005	50	27.12	34.38	33.36
80	0.1	0.1	0.001	250	27.87	33.69	32.88
80	0.1	0.1	0.001	50	27.91	35.09	34.09

Table 9. Ablation study for training loss weights in the unseen-verb zero-shot setting. In each row, one loss weight is varied while others remain fixed. The changed value is shown in blue.

Visual Features for Human, Object and Union Regions We use three visual features from the image feature map $f_{\text{img}}^{\text{glb}}$ for HOI prediction: human (f_{img}^{h}), object (f_{img}^{o}), and union (f_{img}^{u}) features. Ablation results in Table 10 show that using all three yields the best performance.

	mAP		
	Unseen	Seen	Full
H+U	27.92	33.80	32.98
O+U	27.58	33.95	33.06
H+O	27.36	34.81	33.76
H+O+U	27.91	35.09	34.09

Table 10. Ablation study on visual features in the vision branch under the unseen-verb zero-shot setting. "H", "O", and "U" denote f_{img}^{h}, f_{img}^{o}, and f_{img}^{u}, respectively.

Weight Adaptation Visualization Here, we visualize and compare the weights $\mathbf{W}$ before and after the weight adaptation process, especially focusing on the subset of $\mathbf{W}$ applied with the LLM-derived action regularization, as discussed in the main paper Section 3.2. The index set for the subset selection is defined as $\mathcal{I} = \{ i \mid \boldsymbol{b}_i \in \mathbf{B}^a \}$, where $\boldsymbol{b}_i$ is the i-th row of the matrix $\mathbf{B}$ and also belongs to the subset $\mathbf{B}^a$ Before weight adaptation, the subset of $\mathbf{W}$ is obtained by $\{\boldsymbol{w}_i' \mid i \in \mathcal{I}\}$, where $\boldsymbol{w}_i'$ is the i-th column of the matrix $\mathbf{W}$. After weight adaptation, the subset is denoted as $\mathbf{W}_{\text{ar}} = \{\hat{\boldsymbol{w}}_i' \mid i \in \mathcal{I}\}$, where $\hat{\boldsymbol{w}}_i'$ is the i-th column of the adapted matrix $\mathbf{W}$.

As shown in Fig. 2 (a), the subset before the weight adaptation contains limited action-specific information, as indicated by the low cosine similarities between weights for HOI classes associated with the action "ride". This suggests that shared information specific to the action "ride" is not well captured. Moreover, the weights for classes involving the same object, "bike", show high similarity between each other, before weight adaptation. This demonstrates that in the original VLM feature space, actions linked to the same object tend to cluster together. After our proposed weight adaptation, the weight subset $\mathbf{W}_{\mathbf{ar}}$ show noticeably higher similarities among classes that share the "ride" action.

3.4. Qualitative Results

We visualize our method's predictions across four settings in zero-shot HOI detection of HICO-DET: the unseen-verb setting in Fig. 3, the rare-first unseen-composition setting in Fig. 4, the non-rare-first unseen-composition setting in Fig. 5 and the unseen-object setting in Fig. 6. Our HOLa successfully identifies unseen HOI classes in various scenarios, demonstrating its generalization ability to unseen HOI classes. This performance is due to our low-rank decomposed feature adaptation that emphasizes class-shared information, thereby enhancing generalization to unseen classes. Additionally, the incorporation of action priors helps reduce overfitting to seen classes.

3.5. Controlability

While the learned basis features in our low-rank decomposition is not directly interpretable, our method enhances controllability by restricting adaptation to a low-dimensional subspace, spanned by basis vectors $\mathbf{b}_i \in \mathbf{B}$. In this subspace, explicit structures (e.g., orthogonality) are enforced and inspected, instead of modifying the features in the full VLM space.

3.6. Limitations

While our method achieves strong performance in zero-shot HOI detection, it relies on predefined unseen HOI class names, a standard requirement in zero-shot protocols [5–7, 18, 19]. However, this dependency limits flexibility and scalability in real-world scenarios where such predefined classes may be unavailable. To address this, our future work will focus on extending our approach to open-vocabulary HOI detection [10, 20].

Figure 3. Visualization of HOI predictions in the unseen-verb setting on HICO-DET. The purple bar indicates predictions for seen HOI classes and the yellow bar indicates predictions for unseen HOI classes.

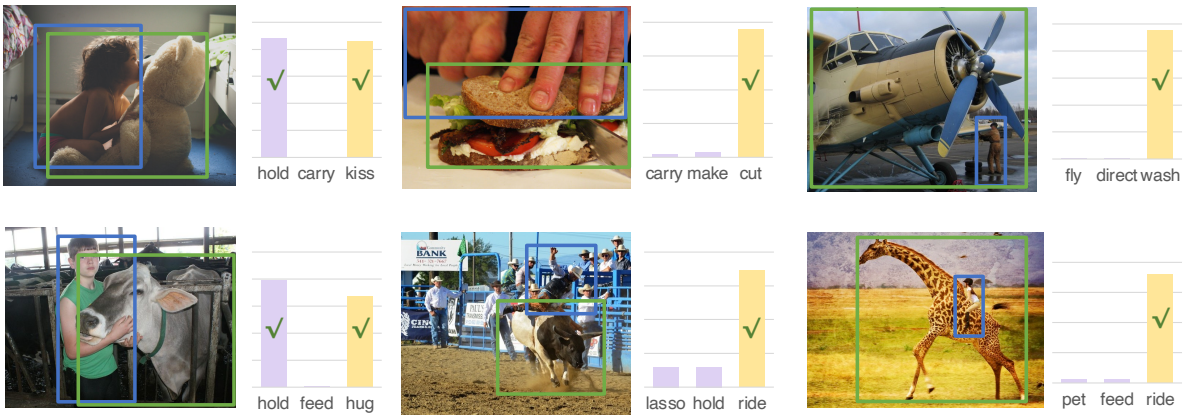

Figure 4. Visualization of HOI predictions in the rare-first unseen-composition setting on HICO-DET. The purple bar indicates predictions for seen HOI classes and the yellow bar indicates predictions for unseen HOI classes.

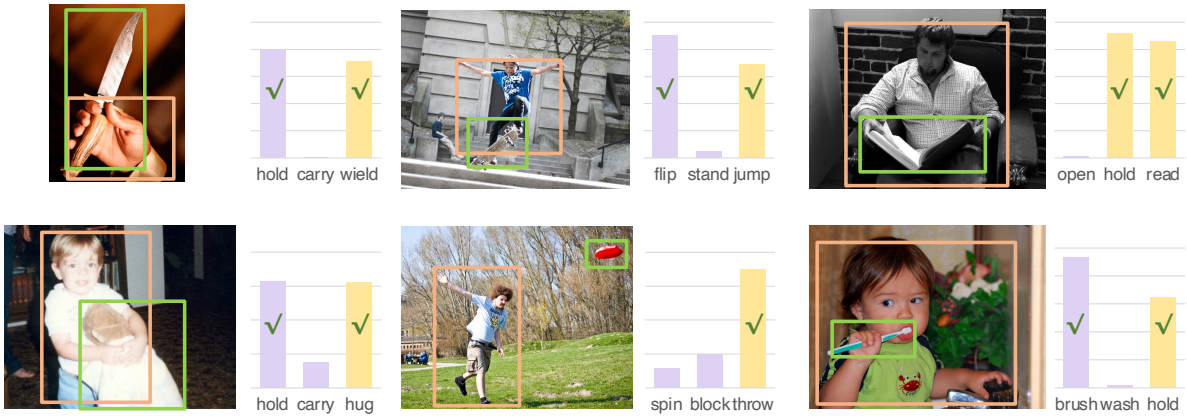

Figure 5. Visualization of HOI predictions in the non-rare-first unseen-composition setting on HICO-DET. The purple bar indicates predictions for seen HOI classes and the yellow bar indicates predictions for unseen HOI classes.

3.7. Future Work Exploration

In our method, adaptation with low-rank decomposition is applied to the language branch, specifically on action and

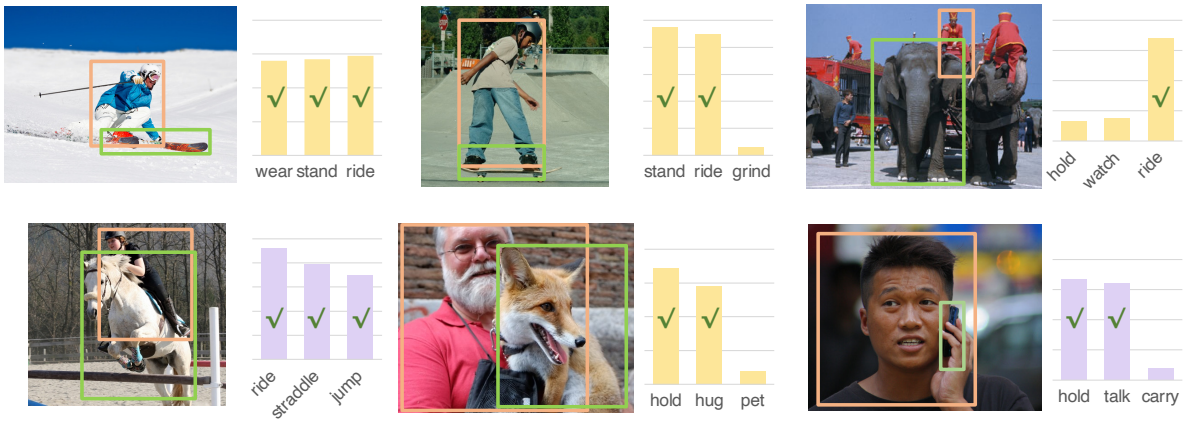

Figure 6. Visualization of HOI predictions in the unseen-object setting on HICO-DET. The purple bar indicates predictions for seen HOI classes and the yellow bar indicates predictions for unseen HOI classes.

interaction features, to enhance generalization to unseen classes. This design leverages the availability of unseen class text descriptions during training, enabling the model to incorporate class-shared knowledge from both seen and unseen HOI classes.

Similar techniques could potentially be extended to object features in the language branch or to visual features. However, in standard two-stage HOI methods [5, 6, 9], object detection is typically handled by an off-the-shelf detector. As a result, the primary challenge in HOI detection lies in modeling unseen actions or novel action-object pairs, rather than object categories, where object generalization is addressed separately in open-vocabulary object detection. However, applying low-rank decomposition to object features may offer a promising direction to benefit open-vocabulary object detection as well.

Furthermore, visual features from unseen classes are not accessible under the standard zero-shot setting, making it infeasible to inject unseen information into the vision branch during training. Exploring decomposition strategies in the vision branch under settings with full or partial visual supervision is another promising avenue for future work.

References

[1] Yichao Cao, Qingfei Tang, Xiu Su, Song Chen, Shan You, Xiaobo Lu, and Chang Xu. Detecting any human-object interaction relationship: Universal hoi detector with spatial prompt learning on foundation models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1

[3] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 3

[4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3

[5] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 584–600. Springer, 2020. 6, 8

[6] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 495–504, 2021. 3, 8

[7] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14646–14655, 2021. 3, 6

[8] Qinqian Lei, Bo Wang, and Robby T. Tan. Ez-hoi: Vlm adaptation via guided prompt learning for zero-shot hoi detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[9] Ting Lei, Fabian Caba, Qingchao Chen, Hailin Jin, Yuxin Peng, and Yang Liu. Efficient adaptive human-object interaction detection with concept-guided memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6480–6490, 2023. 1, 3, 8

[10] Ting Lei, Shaofeng Yin, and Yang Liu. Exploring the potential of large foundation models for open-vocabulary hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16657–16667, 2024. 6

[11] Ting Lei, Shaofeng Yin, Yuxin Peng, and Yang Liu. Exploring conditional multi-modal prompts for zero-shot hoi de-

tection. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024. 3, 4

[12] Liulei Li, Jianan Wei, Wenguan Wang, and Yi Yang. Neural-logic human-object interaction detection. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[13] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20123–20132, 2022. 3

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[17] Yunyao Mao, Jiajun Deng, Wengang Zhou, Li Li, Yao Fang, and Houqiang Li. Clip4hoi: Towards adapting clip for practical zero-shot hoi detection. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[18] Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23507–23517, 2023. 3, 6

[19] Mingrui Wu, Jiaxin Gu, Yunhang Shen, Mingbao Lin, Chao Chen, and Xiaoshuai Sun. End-to-end zero-shot hoi detection via vision and language knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2839–2846, 2023. 3, 6

[20] Jie Yang, Bingliang Li, Ailing Zeng, Lei Zhang, and Ruimao Zhang. Open-world human-object interaction detection via multi-modal prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16954–16964, 2024. 6