SGC-Net: Stratified Granular Comparison Network for Open-Vocabulary HOI Detection

Xin Lin Chong Shi Zuopeng Yang* Haojin Tang* Zhili Zhou
Guangzhou University
linxin94@gzhu.edu.cn, shichong@e.gzhu.edu.cn, yzpeng@gzhu.edu.cn,
tanghaojin@gzhu.edu.cn, zhou_zhili@163.com

Abstract

Recent open-vocabulary human-object interaction (OV-HOI) detection methods primarily rely on large language model (LLM) for generating auxiliary descriptions and leverage knowledge distilled from CLIP to detect unseen interaction categories. Despite their effectiveness, these methods face two challenges: (1) feature granularity deficiency, due to reliance on last layer visual features for text alignment, leading to the neglect of crucial object-level details from intermediate layers; (2) semantic similarity confusion, resulting from CLIP's inherent biases toward certain classes, while LLM-generated descriptions based solely on labels fail to adequately capture inter-class similarities. To address these challenges, we propose a stratified granular comparison network. First, we introduce a granularity sensing alignment module that aggregates global semantic features with local details, refining interaction representations and ensuring robust alignment between intermediate visual features and text embeddings. Second, we develop a hierarchical group comparison module that recursively compares and groups classes using LLMs, generating fine-grained and discriminative descriptions for each interaction category. Experimental results on two widely-used benchmark datasets, SWIG-HOI and HICO-DET, demonstrate that our method achieves state-of-the-art results in OV-HOI detection. Codes will be released on GitHub.

1. Introduction

Human-Object Interaction (HOI) detection aims to localize human-object pairs and recognize their interactions, providing an efficient way for human-centric scene understanding. It plays a crucial role in various computer vision tasks, such as assistive robotics [35, 36] and video analysis [25, 51]. Recently, the emerging field of open-vocabulary HOI (OV-HOI) detection, which broadens the scope of HOI detection by recognizing and associating objects beyond predefined categories, has gained increasing attention.

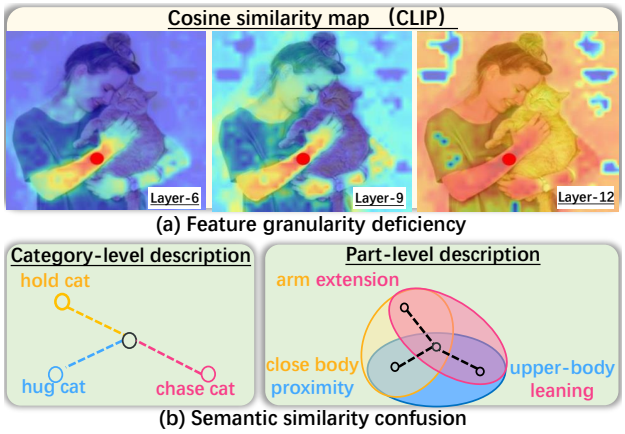

Figure 1. (a) The last layer capture high-level global semantics but contain fewer low-level local details compared to intermediate layers. The red dot marks the selected patch. (b) Category-level and part-level descriptions overlook inter-class similarity, leading to difficulty in distinguishing semantically similar classes.

OV-HOI detection has made notable progress, largely due to the success of vision-language models (VLMs) like CLIP [42]. Unlike zero-shot learning, which prohibits the access of any unannotated unseen classes during training, open-vocabulary learning leverages class name embeddings from VLMs for auxiliary supervision [50, 63]. Therefore, existing CLIP-based HOI detection methods naturally fall under OV-HOI detection. However, these methods face challenges in open-world scenarios because they rely on pretrained object detectors, and using category names as classifiers struggles to capture the variability of HOIs. To overcome these challenges, recent works [24, 48] have removed the need for pretrained detectors by extracting HOI features directly through image encoders while incorporating language priors from LLMs via text encoders.

*Corresponding author.

Despite these advancements, current OV-HOI detection methods without pretrained object detectors still suffer from feature granularity deficiency. This issue arises because CLIP, trained with image-level alignment, produces globally aligned image-text features lacking the local detail needed for effective OV-HOI detection. As shown in Figure 1(a), the last layer focuses on high-level semantic features but neglects low-level local details like arm and face postures, compared to intermediate layers. Additionally, feature dissimilarity between shallow and deep layers increases with network depth, posing significant challenges in aligning intermediate visual features with text embeddings. Moreover, OV-HOI methods that rely on category-level or part-level text classifiers are prone to semantic confusion due to several inherent limitations. First, CLIP's training on large-scale, long-tail datasets tends to introduce biases towards certain classes [49]. For example, as shown on the left side of Figure 1(b), CLIP often confuses "hug cat" with "hold cat." Second, descriptions generated by LLMs based solely on labels may fail to distinguish between semantically similar classes. As illustrated on the right side of Figure 1(b), LLMs produce similar part-level descriptions like "arm extension" for both "hold cat" and "chase cat".

To address the aforementioned challenges, we propose an end-to-end OV-HOI detection network named SGC-Net. First, we design the granularity sensing alignment (GSA) module, leveraging multi-granularity visual features from CLIP to improve the transferability for novel classes. The primary challenge in this process is effectively aggregating global semantic features with local details to refine coarse interaction representations while ensuring alignment between intermediate visual features and text embeddings. Accordingly, we introduce a block partitioning strategy that groups layers based on their relative distances. Specifically, we divide the visual encoder of CLIP into several blocks, ensuring a little variation in the features within each block. To align local details more precisely, we perform a fusion of multi-granularity features within each block, assigning distinct and trainable Gaussian weights to each layer. Once these intra-block features are fused, we proceed to aggregate the fused features across all blocks to maintain global semantic consistency. Furthermore, to preserve the pre-trained alignment between text and visual features in CLIP, we employ visual prompt tuning [7, 27, 62] by adding learnable tokens to the visual features of each layer.

Secondly, we propose the hierarchical group comparison (HGC) module to effectively differentiate semantically similar categories by recursively grouping and comparing classes via LLMs. Specifically, we utilize clustering algorithms (*e.g.*, K-means [33]) for grouping. Even with extensive prior knowledge of LLMs, comparing numerous categories for OV-HOI detection remains challenging due to the quadratic growth of the description matrix with the number of categories. To address this challenge, we adopt group-specific comparison strategies to improve efficiency. For smaller groups, we directly query LLMs for comparative descriptions, whereas for larger groups, we summarize the group's characteristics using LLMs and incorporate these summaries into new queries to highlight distinctive features of each class within the group. By recursively construct a class hierarchy, we can classify HOI by descending from the top to the bottom of the hierarchy, comparing HOI and text embeddings at each level.

In summary, the innovation of the proposed SGC-Net is three-fold: (1) The GSA module effectively aggregates multi-granular features while aligning the local details and global semantic features of OV-HOI; (2) The HGC module iteratively generates discriminative descriptions to refine the classification boundaries of labels; (3) The efficacy of the proposed SGC-Net is evaluated on two widely-used OV-HOI detection benchmark datasets, *i.e.*, HICO-DET [5] and SWIG-HOI [46]. Experimental results show that our SGC-Net consistently outperforms state-of-the-art methods.

2. Related Work

2.1. HOI Detection

Existing HOI detection works can be divided into two categories, two-stage methods [3, 9, 10, 23, 28, 40, 43, 56, 57] and one-stage methods [6, 12, 20, 21, 30, 44, 59]. Two-stage methods typically use an object detector to recognize humans and objects, followed by specialized modules to associate humans with objects and identify their interactions, *e.g.*, multi-steams [12, 13, 29], graphs [10, 41, 52, 55] or compositional learning [14–16]. In contrast, the one-stage approach detects human-object pairs and their interactions simultaneously, without requiring stepwise processing. In particular, RLIP [61] proposes a pre-training strategy for HOI detection based on image captions. PPDM [30] reformulates the HOI detection task as a point detection and matching problem and achieves simultaneous object and interaction detection. However, existing HOI detection approaches are constrained by a closed-set assumption, which restricts their ability to recognize only predefined interaction categories. In contrast, our work aims to detect and recognize HOIs in the open-vocabulary scenario.

2.2. CLIP-based Open Vocabulary HOI Detection

To alleviate the closed-set limitation, many studies [11, 24, 31, 34, 39, 47, 53] have been devoted to OV-HOI detection, aiming to identify both base and novel categories of HOIs while only base categories are available during training. Specifically, they transfer knowledge from the large-scale visual-linguistic pre-trained model CLIP to enhance interaction understanding. For instance, GEN-VLKTs [31] and HOICLIP [39] convert HOI labels into phrase descriptions

to initialize the classifier and extract visual features from the CLIP image encoder to guide the interaction visual feature learning. MP-HOI [53] incorporates the visual prompts into language-guided-only HOI detectors to enhance its generalization capability. THID [48] proposes a HOI sequence parser to detect multiple interactions. CMD-SE [24] detects HOIs at different distances using distinct feature maps. However, the adopted loss function in [24] involves minimizing differences between continuous and discrete variables therefore hard to optimize. Besides, aligning intermediate features with text embeddings may break the alignment in pre-trained CLIP due to the substantial differences. The proposed approach is a method free from pretrained object detectors but has the following advantages compared with existing works. First, it aggregates multi-granularity features from the visual encoder more appropriately. Second, it is easy to optimize and deploy to existing models.

2.3. Leverage LLM for visual classification

Recently, LLMs [1, 2, 8, 18] have made remarkable progress, revolutionizing natural language processing tasks with hundreds of billions of parameters and techniques such as reinforcement learning. Existing works [4, 19, 45, 60] have shown their effectiveness in generating comprehensive descriptions, particularly in classification and detection tasks. Specifically, Menon and Vondrick [37] generate textual descriptions directly from labels to assist VLMs in image classification. I2MVFormer [38] leverages LLM to generate multi-view document supervision for zero-shot image classification. ContexDET [54] utilizes contextual LLM tokens as conditional object queries to enhance the visual decoder for object detection. RECODE [26] leverages LLMs to generate descriptions for different components of relation categories. However, descriptions generated solely by LLMs from labels often tend to be generic and lack sufficient discriminability among semantically similar classes. To address this issue, our work leverages LLM to generate text descriptions by comparing different classes to reduce inter-class similarity and make the decision boundary of each class more compact.

3. Method

This section provides details of the proposed stratified granular comparison network (SGC-Net). Specifically, we first describe the network architecture, followed by an explanation of the training and inference pipeline. As shown in Figure 2, SGC-Net comprises a granularity sensing alignment (GSA) module and a hierarchical group comparison (HGC) module. The GSA module enhances coarse HOI representations by utilizing distance-Aware Gaussian weighting (DGW) and visual prompt tuning to combine multi-granularity features from CLIP's image encoder. The HGC module refines decision boundaries for semantically similar classes by recursively comparing and grouping them using LLM-derived knowledge. In the below, we will describe these two components sequentially.

3.1. Granularity Sensing Alignment

Existing OV-HOI methods [24, 31, 39] typically use features from the last layer of CLIP's image encoder to model HOIs. While these deep features effectively capture high-level semantics, they often lack the local details necessary for HOI detection. One approach to address this issue is to aggregate feature maps from different levels and feed them into an interaction decoder. However, the significant differences between shallow and deep features can disrupt the pre-trained vision-language alignment, ultimately reducing CLIP's zero-shot capability for OV-HOI tasks.

To address this issue, we propose the GSA module, which effectively captures fine-grained details in human-object interactions while preserving the vision-language alignment in the pre-trained CLIP model. Specifically, we first divide the CLIP's image encoder into S blocks, ensuring a little variation in the features within each block. Then, within each block, distinct distance-aware Gaussian weights are assigned to the transformer layers based on their relative neighborhood distances. These weights are trainable, enabling the model to adaptively learn layer- and block-specific information from the training data. Finally, we aggregate the features across different blocks. For each block containing d transformer layers, the aggregated feature $\boldsymbol{Z}$ can be represented as follows:

$$\begin{aligned} \alpha_l^s &= \exp\left(-\frac{1}{2}\frac{(d-l)^2}{\sigma^2}\right), \quad l \in [1, d], \\ \boldsymbol{Z} &= \sum_{s=1}^{S} \alpha_s \left(\sum_{l=1}^{d} \alpha_l^s \boldsymbol{F}_l\right) \end{aligned} \tag{1}$$

Here, l represents the layer index. $\boldsymbol{F}_l$ is a feature map from the l-th layer of CLIP image encoder. α_l^s and α_s denote the distance-aware Gaussian weights between layers within the same block and across different blocks, respectively. By appropriately setting variance parameter σ, the Eq. (1) can assign a higher weight to nearby layers and a lower weight to distant ones, facilitating more effective and flexible integration of features across different depth levels. Notably, the last layer of CLIP is treated as a separate block with a large weight. This design aims to keep the original visual-language correlations established in the pre-trained model. Different blocks can capture varying levels of granularity. Specifically, the early blocks focus more on fine-grained local details, while the later block emphasizes coarse-grained global features.

To better align the intermediate visual features with the text embeddings, we employ visual prompt tuning [60] by introducing learnable tokens onto the visual features of each

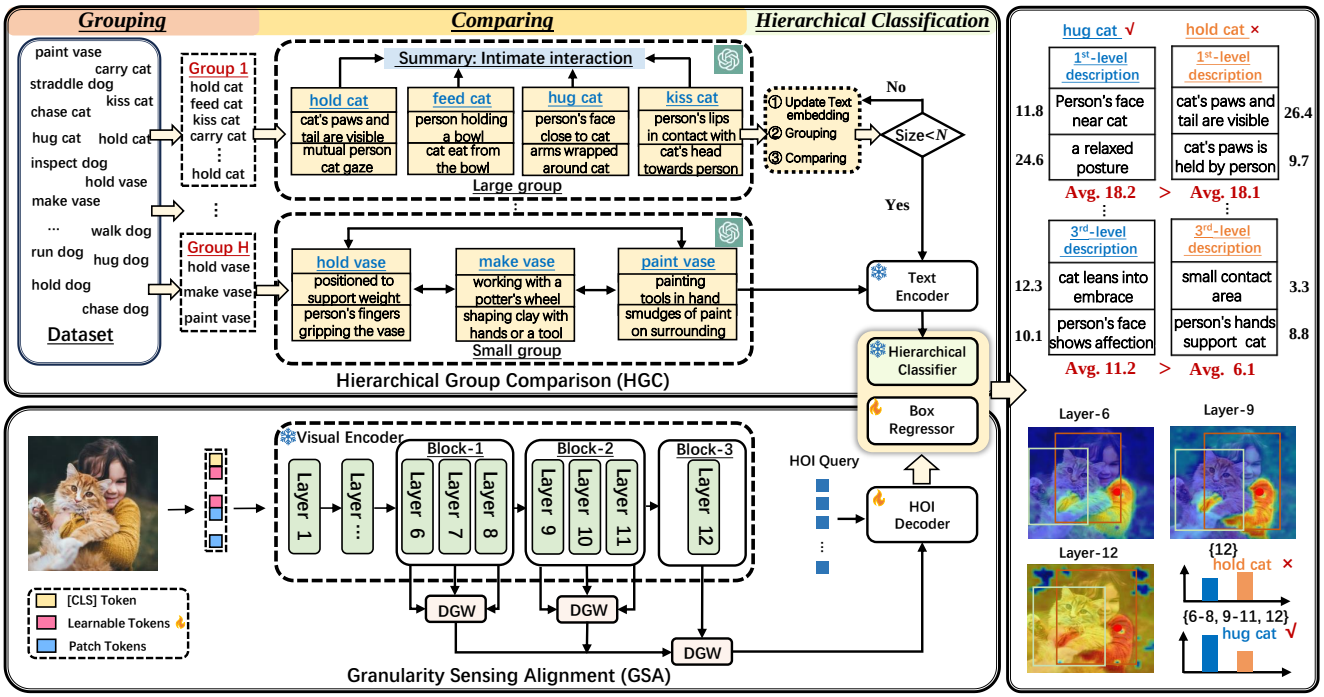

Figure 2. The framework of SGC-Net. It eliminates the need for a pretrained object detector and includes two new modules for OV-HOI detection: (1) The GSA module partitions the CLIP visual encoder into blocks, aggregates features via DGW, and integrates an HOI decoder for fine-grained HOI representations. (2) The HGC module uses LLM to recursively construct class hierarchy, enabling HOI classification by traversing the hierarchy from top to bottom and comparing HOI representations with text embeddings at each level.

layer in the frozen encoder. During visual prompt tuning, the GSA module enables the gradients to be directly back-propagated to the middle layers of the visual encoder. This can promote the alignment of mid-layer features and text embeddings, substantially enhancing the similarity across different layers.

Compared with CMD-SE [24], we effectively aggregate multi-granularity features from the CLIP image encoder using trainable Gaussian weights, which have the following advantages. First, it allows intermediate layer visual features to complement last layer features while preserving the visual-language associations of the pre-trained CLIP. This approach addresses the issue in CMD-SE [24], where relying solely on discrete single-layer features for HOI modeling results in a loss of fine-grained information. Second, our method is easy to optimize. We eliminate the need to use loss functions that align discrete transformer layer indices with continuous human-object interaction distances, thereby avoiding the complex optimization processes.

Subsequently, following previous works [24, 31, 39], the final interaction representation $\boldsymbol{X}$ is formulated as follows:

$$\boldsymbol{X} = \mathrm{Dec}\left(\boldsymbol{Q}, \boldsymbol{Z}\right), \tag{2}$$

where $\boldsymbol{Q}$ refers to HOI queries, while $\boldsymbol{Z}$ is treated as both the key and value input to the HOI Decoder. The output $\boldsymbol{X}$ is later passed to the bounding box regressor and classifier to predict the bounding boxes of human-object pairs and their interaction categories.

3.2. Hierarchical Group Comparison

After obtaining the fine-grained interaction representation by the GSA module, designing a discriminative classifier remains a challenging task. Existing methods [24, 31, 39] typically design classifiers with the manual prompt "a photo of a person [verb] [object]". However, these classifiers primarily rely on category names, often neglecting the contextual information provided by language. Recent research [24] has introduced the category descriptions produced by LLMs to enhance the model's generalization capability. However, these descriptions directly generated by LLMs from labels are often generic and semantically similar.

To address this issue, we propose the hierarchical group comparison (HGC) module. Inspired by the coarse-to-fine approach humans use to recognize objects, HGC employs three strategies: Grouping, Comparison, and Hierarchical Classification. It identifies semantically similar descriptions and refines decision boundaries from generic to more discriminative. Specifically, we first utilize LLMs to generate initial descriptions for each HOI category.

```
Q: What features are useful to distinguish
```

{HOI category} in a photo?
A: -

Grouping: Following [37], we employ the pretrained CLIP text encoder to map the LLM-generated descriptions into the latent space. Utilizing these features, we apply a clustering algorithm, such as K-means [33], to perform grouping and identify semantic neighbors. Specifically, the value of K is calculated as the number of categories divided by a predefined grouping threshold N. Even with extensive prior knowledge of LLMs, comparing a large number of categories presents a formidable challenge, as it needs to generate a comprehensive description matrix that grows quadratically with the number of categories. To address this issue, we employ tailored strategies for generating descriptions based on the group size, allowing for more efficient and effective comparisons. Specifically, we adopt summary-based comparison for larger groups and direct comparison for smaller ones.

Comparing: For groups involving a large number of categories (*i.e.*, exceeding half of the grouping threshold), we leverage the LLM to summarize their overall characteristics. By incorporating these summarized characteristics into a new query, we can generate a comparative description that captures the distinctive features and relationships among the categories. This approach enhances our ability to effectively compare and contrast the elements within a group.

Q: Summarize the following interactions with one sentence: {category list}?
A: {subset description}

Q: What features are useful to distinguish {HOI category} from {subset description}?
A: -

When the number of elements within a group is relatively small, a direct query to the LLM enables us to obtain an exceptionally comparative description. This approach capitalizes on the LLM's capabilities to provide a comprehensive and insightful analysis, enhancing the comparative understanding of the elements within the group.

Q: What features are useful to distinguish {target category} from {other categories} in a photo?
A: -

Utilizing these detailed and comparative descriptions, we can achieve more nuanced text embedding. Subsequently, we can continue to perform clustering, followed by additional comparisons to derive new descriptors.

Hierarchical Classification: After constructing the class hierarchy, we can obtain text features at multiple hierarchical levels. More specific, the hierarchical text features can be represented as $\mathcal{T} = \{\boldsymbol{D}_1, \boldsymbol{D}_2, \cdots, \boldsymbol{D}_T\}$, where T is the number of test classes. $\boldsymbol{D}_i \in \mathbb{R}^{M_j \times C}$ represents the hierarchical text embeddings for category i. C is the embedding dimension. M_j is the number of comparative descriptions for category i. Accordingly, the cosine similarity score between the HOI feature $\boldsymbol{x}$ and the j-th description of the i-th HOI category, denoted as $\boldsymbol{D}_i^j$, can be expressed as follows:

$$p_i^j = \boldsymbol{D}_i^j \cdot \boldsymbol{x}^T. \tag{3}$$

However, unreliable low-level descriptions can introduce errors and redundancies, degrading the quality of high-level descriptions and ultimately impairing the model's classification performance. To address this issue, we define an iterative evaluator $\boldsymbol{u}_i^k$ to evaluate the acceptability of the current k-th level description for the i-th category as follows:

$$\boldsymbol{u}_i^k = \mathbb{I}\left(p_i^{k+1} > p_i^k + \tau\right), \tag{4}$$

where τ denotes the tolerance parameter, $\mathbb{I}$ represents indicator function. Eq. (4) ensures that a new score is merged only if a subsequent discriminative description yields a higher score than the current one. Subsequently, we obtain the running average of the longest sequence of monotonically increasing p values as follows:

$$r(\boldsymbol{x}, i) = \frac{p_i^1 + \sum_{j=2}^{M_i} p_i^j \prod_{k=1}^{j-1} \boldsymbol{u}_i^k}{1 + \sum_{j=2}^{M_i} \prod_{k=1}^{j-1} \boldsymbol{u}_i^k}. \tag{5}$$

Finally, the similarity score $s(\boldsymbol{x}, i)$ between HOI feature $\boldsymbol{x}$ and the i-th interaction label embedding is computed using a weighted average function, as follows:

$$s(\boldsymbol{x}, i) = (1-\lambda)(\boldsymbol{p}_i^1 + \boldsymbol{t} \cdot \boldsymbol{x}^T) + \lambda\left(r(\boldsymbol{x}, i)\right). \tag{6}$$

Here, we apply prompt tuning [48] with learnable text tokens $\boldsymbol{t}$ to learn the sentence format rather than using manually defined context words. Besides, our fusion method takes into account the score of the initial description $\boldsymbol{p}_i^1$ as an offset, and we introduce a constant hyperparameter $\lambda \in [0, 1]$ to balance the two terms.

3.3. Training and Inference

In this subsection, we elaborate on the processes of training and inference of our model.

Training. During the training stage, we follow the query-based methods [24, 31, 48] to assign a bipartite matching prediction with each ground-truth using the Hungarian algorithm [22]. The matching cost $\mathcal{L}$ for the matching process and the targeting cost for the training back-propagation share the same strategy, which is formulated as follows:

$$\mathcal{L} = \lambda_b \sum_{i \in \{h,o\}} \mathcal{L}_b^i + \lambda_{iou} \sum_{i \in \{h,o\}} \mathcal{L}_{iou}^i + \lambda_{cls} \mathcal{L}_{cls}, \tag{7}$$

where $\mathcal{L}_b$, $\mathcal{L}_{iou}$, and $\mathcal{L}_{cls}$ denote the box regression, intersection over union, and classification losses, respectively.

During the training stage, we follow the query-based methods [24, 31, 39] to assign a bipartite matching prediction with each ground-truth using the Hungarian algorithm [22]. λ_b, λ_{iou}, and λ_{cls} are the hyper-parameter weight.

Inference. For each HOI prediction, including the bounding-box pair $(\hat{b_h^i}, \hat{b_o^i})$, the bounding box score $\hat{c}_i$ from the box regressor, and the interaction score $\hat{s}_i$ from the interaction classifier, the final score $\hat{s_i}'$ is computed as:

$$\hat{s_i}' = \hat{s}_i \cdot \hat{c}_i^{\gamma} \tag{8}$$

where $\gamma > 1$ is a constant used during inference to suppress overconfident objects [55, 56].

4. Experiment

4.1. Experimental Setting

Datasets. Our experiments are conducted on two datasets: namely, SWIG-HOI [46] and HICO-DET [5]. The SWIG-HOI dataset provides diverse human interactions with large-vocabulary objects, comprising 400 human actions and 1,000 object categories. During the testing, we utilize approximately 5,500 interactions, including around 1,800 interactions that are not found in the training set. The annotations of HICO-DET include 600 combinations of 117 human actions and 80 objects. Among the 600 interactions, follow [14, 24, 48], we simulate the open-vocabulary detection setting by holding out 120 rare interactions.

Evaluation Metric. Following in [5, 24, 31, 32, 48], we use the mean Average Precision (mAP) for evaluation. An HOI triplet prediction is classified as a true-positive example when the following criteria are satisfied: 1) The intersection over union of the human bounding box and object bounding box are larger than 0.5 *w.r.t.* the GT bounding boxes; 2) the predicted interaction category is accurate.

Implementation Details. We follow the settings of previous work [24, 48] to build our model upon the pretrained CLIP. Specifically, for the visual encoder, we employ the ViT-B/16 version as our visual encoder and apply 12 learnable tokens in each layer to detect human-object interactions. For the text encoder, we introduce 8 prefix tokens and 4 conjunctive tokens to connect the words of human actions and objects for adaptively learning categories' information. Our model is optimized using AdamW with an initial learning rate of 10^{-4}, using 64 as the batch size for SWIG-HOI dataset and 32 for HICO-DET. We set the cost weights λ_b, λ_{cls} and λ_{iou} to 5, 2, and 5, respectively. The LLM we utilize is GPT-3.5. We set the hyperparameter λ in Eq. (6) as 0.5. In all experiments, the variance parameter σ is set to 1, and the tolerance parameter τ is set to 0.

4.2. Comparisons with State-of-the-Art Methods

HICO-DET. To facilitate fair comparison, we first compare our model with state-of-the-art methods without pretrained detectors on the HICO-DET dataset. As shown in Table 1, SGC-Net outperforms CMD-SE [24] by 6.57%, 4.39%, and 4.87% in mAP on the Unseen, Seen, and Full categories, respectively. Furthermore, even when compared to methods using pretrained detectors, SGC-Net achieves competitive performance. Although recent OV-HOI methods (*e.g.*, GEN-VLKT [31], OpenCat [58], HOICLIP [39]) leverage CLIP text embeddings for interaction classification, they typically rely on a DETR architecture with pretrained weights. It is important to note that comparing our method with these approaches on HICO-DET dataset is not entirely fair, as the COCO dataset used for DETR pretraining shares the same object label space as HICO-DET.

SWIG-HOI. Table 2 shows that SGC-Net outperforms

Method	Pretrained Detector	Unseen	Seen	Full
FCL [16]	✓	13.16	24.23	22.01
SCL [17]	✓	19.07	30.39	28.08
GEN-VLKT [31]	✓	21.36	32.91	30.56
OpenCat [58]	✓	21.46	33.86	31.38
HOICLIP [39]	✓	**23.48**	**34.47**	**32.26**
THID [48]	✗	15.53	24.32	22.38
CMD-SE [24]	✗	16.70	23.95	22.35
SGC-Net	✗	**23.27**	**28.34**	**27.22**

Table 1. Comparison of our proposed SGC-Net with state-of-the-art methods on the HICO-DET dataset.

Method	Non-rare	Rare	Unseen	Full
CHOID [46]	10.93	6.63	2.64	6.64
QPIC [43]	16.95	10.84	6.21	11.12
GEN-VLKT [31]	20.91	10.41	-	10.87
MP-HOI [53]	20.28	14.78	-	12.61
THID [48]	17.67	12.82	10.04	13.26
CMD-SE [24]	21.46	14.64	10.70	15.26
SGC-Net	**23.67**	**16.55**	**12.46**	**17.20**

Table 2. Comparison of our proposed SGC-Net with state-of-the-art methods on the SWIG-HOI dataset.

Method	Non-rare	Rare	Unseen	Full
Base	15.69	11.53	7.32	11.45
+ *GSA*	22.74	16.00	11.64	16.49
+ *HGC*	21.18	14.19	10.69	14.81
SGC-Net	**23.67**	**16.55**	**12.46**	**17.20**

Table 3. Ablation studies of the proposed method.

Layer splitting manner	Non-rare	Rare	Unseen	Full
{4-6}, {7-9}, {10-12}	23.03	15.72	11.35	16.35
{6-8}, {9-10}, {11-12}	23.61	16.52	11.64	17.01
{6-8}, {9-11}, {12}	**23.67**	**16.55**	**12.46**	**17.20**

Table 4. Effectiveness of different layer splitting manners. The elements in {} means the layer numbers that are used.

Number of blocks	Non-rare	Rare	Unseen	Full
{12}	21.18	14.19	10.09	14.81
{9-11}, {12}	21.91	14.48	10.01	14.78
{6-8}, {9-11}, {12}	**23.67**	**16.55**	**12.46**	**17.20**
{3-5}, {6-8}, {9-11}, {12}	22.61	15.25	12.01	16.53
Number of layers				
{8-9}, {10-11}, {12}	22.45	15.17	11.55	15.98
{6-8}, {9-11}, {12}	**23.67**	**16.55**	**12.46**	**17.20**
{4-7}, {8-11}, {12}	23.50	16.21	12.13	16.75

Table 5. Evaluation on the number of blocks and layers.

Aggregation	Non-rare	Rare	Unseen	Full
Self-attention	21.37	13.43	7.88	13.90
Concat	21.84	15.52	10.22	15.68
Sum	22.09	14.89	9.59	15.26
DGW	23.40	16.14	11.60	16.71
Sum*	22.21	15.16	10.45	15.64
DGW*	**23.67**	**16.55**	**12.46**	**17.20**

Table 6. Comparison of different aggregation methods. * denotes that the weights are trainable.

all state-of-the-art methods on various setups. Specifically, SGC-Net outperforms CMD-SE [24] by 1.76% and 1.91% in mAP on the Rare and Unseen categories, respectively. Furthermore, SGC-Net outperforms the best OV-HOI method with a pretrained object detector by 4.59% in mAP on the Full category. These results indicate that methods such as GEN-VLKT [31] and MP-HOI [53], which rely on pretrained detectors, perform suboptimally on the SWIG-HOI dataset. This is primarily because these methods struggle to scale effectively with vocabulary size, ultimately limiting their applicability in open-world scenarios. In contrast, SGC-Net overcomes this constraint by not relying on any detection pretraining, demonstrating superior capabilities in detecting and recognizing OV-HOI.

4.3. Ablation Study

To prove the effectiveness of our proposed methods, we conduct five ablation studies on the SWIG-HOI dataset. Results of the ablation studies are summarized in Table 3-6, and Figure 3, respectively.

Effectiveness of the proposed modules. We first perform

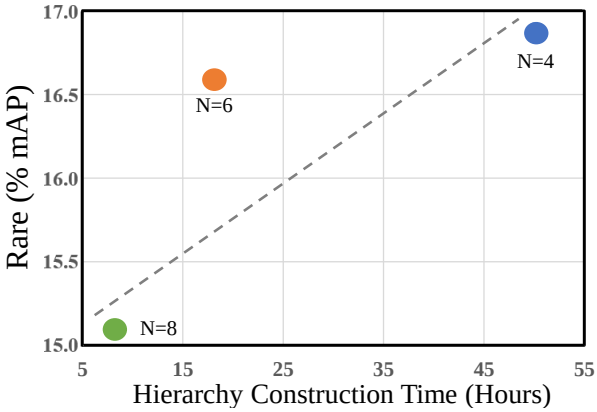

Figure 3. Evaluation on the value of grouping threshold N.

an ablation study to validate the effectiveness of the GSA and HGC modules. Results are summarized in Table 3. Our baseline model follows CMD-SE [24], which utilizes CLIP and removes the need for pretrained object detectors in OV-HOI detection. First, we apply the GSA module on the baseline model to improve its transferability for novel classes, denoted as *+GSA*. This modification achieves mAP improvements of 7.05% for Non-Rare, 2.66% for Rare, 4.32% for Unseen, and 5.04% for Full categories. Moreover, we integrate the HGC module with the baseline to enhance its ability to distinguish semantically similar interactions, denoted as *+HGC*. The results indicate that the HGC module enables the baseline model to achieve 14.81% in mAP for the Full categories. Finally, combining both modules results in the best performance, with an mAP of 12.46% on the Unseen categories. These results demonstrate the significant potential of SGC-Net in enhancing interaction understanding within an open-vocabulary setting.

Effectiveness of the proposed layer splitting manner. The block partitioning strategy plays a crucial role in the GSA module due to its potential to preserve visual-language associations. Experimental results in Table 4 indicate that separating the last layer into an independent block is more effective than other combinations of partitioning strategies. This is because the last layer of CLIP's image encoder has the strongest association with text embeddings, and partitioning it into a separate block helps preserve this correlation while minimizing potential disruptions.

Number of blocks and number of layers in each block. To show the importance of multi-granularity feature fusion across different layers, we compare the performance of SGC-Net with varying numbers of blocks and layers per block. As shown in Table 5, increasing the number of blocks from 1 to 3 gradually improves the performance, particularly for Unseen interactions. The best performance is achieved with three blocks, each containing three layers. Reducing the number of layers to 2 degrades performance, primarily due to neglecting mid-layer features. It should be

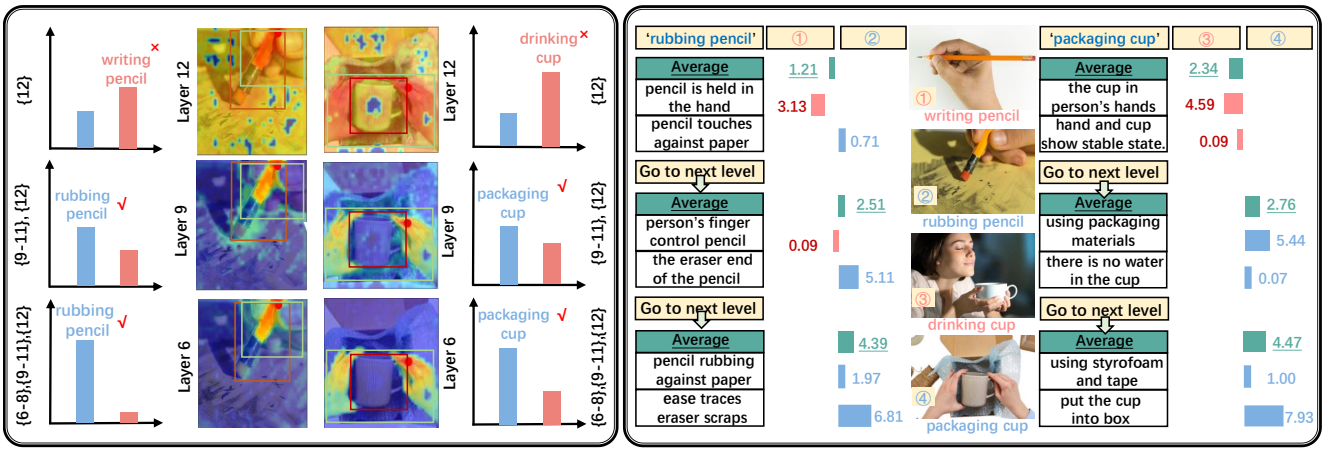

Figure 4. Qualitative results of SGC-Net. On the left-hand side, we visualize feature correspondence through cosine similarity calculations from both deep and shallow layers, and compare the predicted HOI categories. On the right-hand side, we show two inference examples and the absolute score gap between two images at each hierarchical level by querying the description with extracted HOI representations.

noted that the beginning layers are excluded, as they primarily encode features with limited semantics.

Comparison of different aggregation methods. Due to the substantial differences between shallow and deep features, directly integrating multi-granularity features may disrupt the alignment of vision-language embeddings in the pre-trained CLIP model. To address this issue, we introduce the DGW function, which mitigates these disruptions by accounting for the distances between layers. As shown in Table 6, DGW consistently outperforms direct aggregation strategies such as Sum, Concat. Additionally, setting the aggregation weights as learnable further enhances the performance of DGW. These results demonstrate that our method, by assigning smaller weights to features from more distant layers during multi-granularity fusion, effectively improves model performance in open-vocabulary tasks.

Grouping threshold N. A large grouping threshold, such as $N = 8$, tends to produce rudimentary hierarchies and generalized descriptions. In contrast, smaller grouping thresholds increase the time required to construct the class hierarchy due to the more detailed recursive comparisons and groupings involved. As shown in Figure 3, the performance of SGC-Net improves as the grouping threshold increases. To balance computational efficiency and performance gains, we chose to set $N = 6$.

4.4. Qualitative Results

To demonstrate the effectiveness of the proposed GSA module in leveraging CLIP's multi-granularity features, we present cosine similarity maps of features and qualitative OV-HOI detection results. At the left hand side of Figure 4, we show the patch similarity of Unseen classes not included in the training process. We observe that the intermediate layers of our method contain detailed information on local objects, including interaction patterns. Moreover, by leveraging these distinctive features, our SGC-Net improves HOI detection performance for both seen and unseen classes compared to using only last-layer features.

At the right hand side of Figure 4, we show two groups of examples that demonstrate the inference and absolute score gap calculation via the proposed HGC module. Specifically, we input two Unseen class images (② and ④) and two Rare class images (① and ③) to evaluate their similarity with the hierarchical descriptions. The absolute score gap is computed as the absolute difference in similarity scores between paired images: *i.e.*, ① vs. ② and ③ vs. ④. Due to the weak comparative information in the early descriptions, similar images tend to yield similar scores at the beginning. As we progress towards a more comparative description, the disparity between the scores of the two images gradually increases. Besides, this also highlights the interpretability of our method, as it provides the user insights regarding the attributes that elicit stronger responses to the input. Additionally, it indicates the specific layer of description at which the scores of different images diverge, widening the gap.

5. Conclusion

In this paper, we propose the SGC-Net which enhances CLIP's transferability and discriminability to handle two critical issues in OV-HOI: feature granularity deficiency and semantic similarity confusion. Specifically, the GSA module enhances CLIP's image encoder by improving its ability to capture fine-grained HOI features while maintaining alignment between multi-granularity visual features and text embeddings. The HGC module strengthens CLIP's text encoder in distinguishing semantically related HOI categories by incorporating a grouping and comparison strategy into the LMM. Through extensive comparative experiments and ablation studies, we validate the effectiveness of SGC-Net on two datasets.

References

[1] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023. 3

[2] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 3

[3] Yichao Cao, Qingfei Tang, Feng Yang, Xiu Su, Shan You, Xiaobo Lu, and Chang Xu. Re-mine, learn and reason: Exploring the cross-modal semantic correlations for language-guided hoi detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23492–23503, 2023. 2

[4] 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

[5] 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. 2, 6

[6] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9004–9013, 2021. 2

[7] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, pages 11583–11592, 2022. 2

[8] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 3

[9] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018. 2

[10] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 696–712. Springer, 2020. 2

[11] Jianjun Gao, Kim-Hui Yap, Kejun Wu, Duc Tri Phan, Kratika Garg, and Boon Siew Han. Contextual human object interaction understanding from pre-trained large language model. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13436–13440. IEEE, 2024. 2

[12] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8359–8367, 2018. 2

[13] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9677–9685, 2019. 2

[14] 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. 2, 6

[15] 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.

[16] 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. 2, 6

[17] Zhi Hou, Baosheng Yu, and Dacheng Tao. Discovering human-object interaction concepts via self-compositional learning. In *European Conference on Computer Vision*, pages 461–478. Springer, 2022. 6

[18] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022. 3

[19] Sheng Jin, Xueying Jiang, Jiaxing Huang, Lewei Lu, and Shijian Lu. Llms meet vlms: Boost open vocabulary object detection with fine-grained descriptors. *arXiv preprint arXiv:2402.04630*, 2024. 3

[20] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 498–514. Springer, 2020. 2

[21] Bumsoo Kim, Jonghwan Mun, Kyoung-Woon On, Minchul Shin, Junhyun Lee, and Eun-Sol Kim. Mstr: Multi-scale transformer for end-to-end human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19578–19587, 2022. 2

[22] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5, 6

[23] 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. 2

[24] 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*, pages 16657–16667, 2024. 1, 2, 3, 4, 5, 6, 7

[25] Huadong Li, Ying Wei, Shuailei Ma, Mingyu Chen, and Ge Li. Ripple transformer: A human-object interaction backbone and a new prediction strategy for smart surveillance devices. *IEEE Transactions on Consumer Electronics*, 2024. 1

[26] Lin Li, Jun Xiao, Guikun Chen, Jian Shao, Yueting Zhuang, and Long Chen. Zero-shot visual relation detection via composite visual cues from large language models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[27] Yunheng Li, ZhongYu Li, Quansheng Zeng, Qibin Hou, and Ming-Ming Cheng. Cascade-clip: Cascaded vision-language embeddings alignment for zero-shot semantic segmentation. *arXiv preprint arXiv:2406.00670*, 2024. 2

[28] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2019. 2

[29] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 382–391, 2020. 2

[30] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490, 2020. 2

[31] 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*, pages 20123–20132, 2022. 2, 3, 4, 5, 6, 7

[32] Xinpeng Liu, Yong-Lu Li, Xiaoqian Wu, Yu-Wing Tai, Cewu Lu, and Chi-Keung Tang. Interactiveness field in human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20113–20122, 2022. 6

[33] J Macqueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*, 1967. 2, 5

[34] 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*, 36:45895–45906, 2023. 2

[35] Esteve Valls Mascaro, Daniel Sliwowski, and Dongheui Lee. Hoi4abot: Human-object interaction anticipation for human intention reading collaborative robots. *arXiv preprint arXiv:2309.16524*, 2023. 1

[36] Jean Massardi, Mathieu Gravel, and Éric Beaudry. Parc: A plan and activity recognition component for assistive robots. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3025–3031. IEEE, 2020. 1

[37] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*, 2022. 3, 5

[38] Muhammad Ferjad Naeem, Muhammad Gul Zain Ali Khan, Yongqin Xian, Muhammad Zeshan Afzal, Didier Stricker, Luc Van Gool, and Federico Tombari. I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15169–15179, 2023. 3

[39] 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. 2, 3, 4, 6

[40] Jeeseung Park, Jin-Woo Park, and Jong-Seok Lee. Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17152–17162, 2023. 2

[41] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 401–417, 2018. 2

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[43] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021. 2, 6

[44] Danyang Tu, Wei Sun, Guangtao Zhai, and Wei Shen. Agglomerative transformer for human-object interaction detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21614–21624, 2023. 2

[45] Mesut Erhan Unal and Adriana Kovashka. Weakly-supervised hoi detection from interaction labels only and language/vision-language priors. *arXiv preprint arXiv:2303.05546*, 2023. 3

[46] Suchen Wang, Kim-Hui Yap, Henghui Ding, Jiyan Wu, Junsong Yuan, and Yap-Peng Tan. Discovering human interactions with large-vocabulary objects via query and multi-scale detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13475–13484, 2021. 2, 6

[47] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognexploiition (CVPR)*, pages 939–948, 2022. 2

[48] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable

human-object interaction detector with natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2022. 1, 3, 5, 6

[49] Xudong Wang, Zhirong Wu, Long Lian, and Stella X Yu. Debiased learning from naturally imbalanced pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14647–14657, 2022. 2

[50] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, et al. Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[51] Nan Xi, Jingjing Meng, and Junsong Yuan. Open set video hoi detection from action-centric chain-of-look prompting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3079–3089, 2023. 1

[52] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2

[53] 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. 2, 3, 6, 7

[54] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *International Journal of Computer Vision*, pages 1–19, 2024. 3

[55] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13319–13327, 2021. 2, 6

[56] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20104–20112, 2022. 2, 6

[57] Frederic Z Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring predicate visual context in detecting of human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10411–10421, 2023. 2

[58] Sipeng Zheng, Boshen Xu, and Qin Jin. Open-category human-object interaction pre-training via language modeling framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19392–19402, 2023. 6

[59] Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13234–13243, 2021. 2

[60] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3

[61] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 843–851, 2019. 2

[62] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023. 2

[63] Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1