FreeA: Human-object Interaction Detection using Free Annotation Labels

Yuxiao Wang, Zhenao Wei, Xinyu Jiang, Yu Lei, Weiying Xue, Jinxiu Liu, Qi Liu*, *Senior Member, IEEE*

***Abstract*—Recent human-object interaction (HOI) detection approaches rely on high cost of manpower and require comprehensive annotated image datasets. In this paper, we propose a novel self-adaption language-driven HOI detection method, termed as FreeA, without labeling by leveraging the adaptability of CLIP to generate latent HOI labels. To be specific, FreeA matches image features of human-object pairs with HOI text templates, and *a priori* knowledge-based mask method is developed to suppress improbable interactions. In addition, FreeA utilizes the proposed interaction correlation matching method to enhance the likelihood of actions related to a specified action, further refine the generated HOI labels. Experiments on two benchmark datasets show that FreeA achieves state-of-the-art performance among weakly supervised HOI models. Our approach is +8.58 mean Average Precision (mAP) on HICO-DET and +1.23 mAP on V-COCO more accurate in localizing and classifying the interactive actions than the newest weakly model, and +1.68 mAP and +7.28 mAP than the latest weakly+ model, respectively. Code will be available at https://drliuqi.github.io/.**

***Index Terms*—Human-object interaction detection, weakly supervised, object detection.**

I. Introduction

HUMAN-OBJECT interaction (HOI) aims to localize and classify the interactive actions between a human and an object, enabling a more advanced understanding of images [1]. Specifically, the HOI detection task involves taking an image as input to generate a series of triplets (⟨"human", "interaction", "object" ⟩). Consequently, the success of this task is mainly attributed to the accurate localization of human and object entities, correct classification of object categories, and precise delineation of interaction relationships between humans and objects.

Whether one-stage [2]–[12] or two-stage [1], [13]–[22] HOI detection models, predominantly relies on computationally heavy training and requires extensive-annotation datasets (Figure 1(a) and Figure 1(b)). Taking the HICO-Det dataset as an example, the weakly+ or weakly supervised approaches need to annotate 117,871 interaction labels from (117,871×600) or (117,871×23) potential combinations. It is quite resource-intensive. As shown in Figure 1(b), HOI models can be categorized into twofolds [23]: weakly+ using ⟨"interaction", "object"⟩ label [24]–[26], e.g., eat-banana, and weakly with only ⟨"interaction"⟩ labels, e.g., eat. Although weakly supervised HOI detection models typically employs image-level labels, with only interaction action categories, they still require abundant annotations for large-scale datasets to achieve satisfactory performance. Compared to weakly+ supervised methods, weakly supervised ones are more challenging since they require to explore objects within a broader interaction space [23]. For instance, when considering the action "riding", there could be a wide range of objects involved. Weakly+ supervised models with ⟨"ride", "motorcycle"⟩ labels effectively focus on motorcycles, while it is not easy for weakly supervised models with ⟨"ride"⟩ to determine the accurate object because of multiple matching choices, such as riding bicycle. The proposed method, called as FreeA, belongs to the latter[1]. In

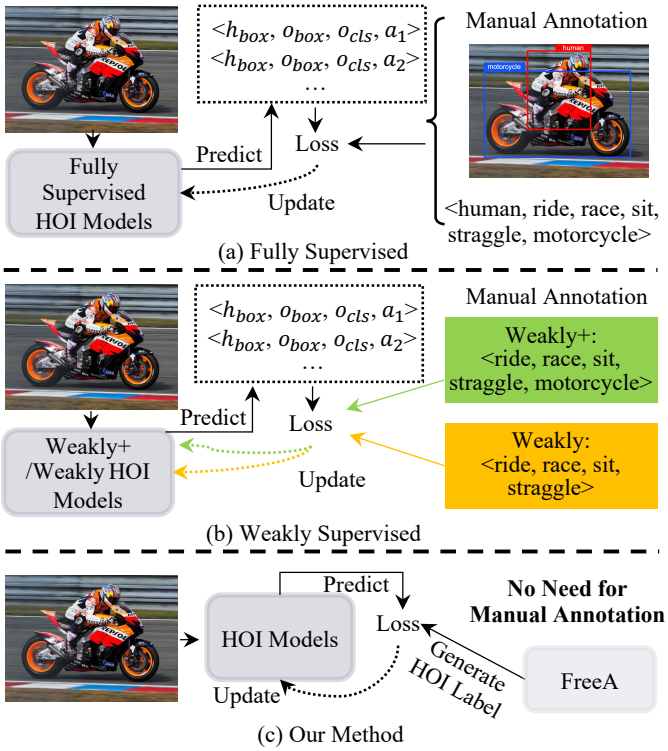

Fig. 1. HOI methods overview: (a) Fully supervised HOI models. Labels consist of human bounding boxes, object bounding boxes, object categories, and interaction actions of each human-object pair. (b) HOI weakly supervised. It divides into twofolds: weakly+ (using ⟨"interaction", "object"⟩ labels), and weakly (using ⟨"interaction"⟩ labels). (c) Our method, i.e., FreeA, automatically generates HOI labels for HOI model training without the need for any manual annotation.

* Corresponding author: Qi Liu (drliuqi@scut.edu.cn)
Y. Wang (ftwangyuxiao@mail.scut.edu.cn), Z. Wei, X. Jiang, W. Xue, J. Liu, and Q. Liu are with the School of Future Technology, South China University of Technology, China 511400.
Y. Lei is with the School of Information Science & Technology, Southwest Jiaotong University, China 611730 (e-mail:leiyu1117@my.swjtu.edu.cn).

[1]The reason why FreeA is considered as one weakly supervised HOI model is that the used labels in FreeA are automatically generated by CLIP.

addition, another sub-direction of research in HOI detection is zero-shot HOI detection [9], [15], [18], [27]–[36], where the model is trained using only a subset of annotated datasets but is tested on interaction categories not seen during training. It can be observed that zero-shot HOI detection focuses on discovering unseen samples. However, it still relies on training with manually annotated labels.

Motivated by the ability of text CLIP (contrastive language image pretraining) estimates best paired with a given image, FreeA is developed for achieving HOI detection task without the need for manual labels (Figure 1(c)). The FreeA mainly comprises three-folds: candidate image construction (CIC), human-object potential interaction mining (PIM), and human-object interaction inference (HII). In the phase of CIC, FreeA is plug-and-play applying existing object detection methods for all potential instances localization, and utilizes spatial denoising and pairing techniques to establish candidate interaction pairs within the image. The PIM module extensively leverages the adaptability of the text-image model, i.e., CLIP, in the target domain to align the high-dimensional image features with HOI interaction templates, which generates the similarity vectors of candidate interaction relationships. Then, the HII module combines the resulting similarity vectors with a priori HOI action masks to mitigate interference from irrelevant relationships, and augments the likelihood of specific actions through the proposed interaction correlation matching method. Moreover, an adaptive threshold in HII is used to dynamically generate HOI labels for training.

Our key contributions are summarized as threefolds:

1) We propose a novel HOI detection method, namely FreeA, automatically generates HOI labels. To the best of our knowledge, it is the first framework to successfully achieve HOI detection task without the need for manual labeling.
2) HOI detection includes various interactions among multiple instances. Three key challenges are required to tackle when using CLIP to generate labels, namely, multiple actions selection, filtering out irrelevant actions, and refining CLIP's coarse labels. To address that, three corresponding modules are presented that significantly improve the effectiveness of the interactions localization and classification.
3) A broad variety of experiments are conducted to demonstrate the remarkable results of the proposal on HOI detection task, and ours performs the best among all weakly+, weakly and several fully supervised HOI models.

II. Related Work

Supervised HOI Detection. Supervised HOI models train their networks with the help of manually annotation ⟨"person", "object", "action"⟩. Their networks are presented as two-stage or single-stage pipelines. The two-stage methods first use pre-trained object detection network [1], [37]–[39] to detect humans and objects, and then pair them one by one into the interactive discrimination network to achieve HOI detection [1], [13]–[22]. However, it is pretty inefficient for one-to-one pairing between humans and objects [2], [8]. To address that, single-stage HOI detection based on transformer is gradually developed via end-to-end solution [2]–[12]. HOITrans [8] and QPIC [40] applied extractors and encoders in DETR [41] to extract features for global feature encoding. It had been verified that text information enables to improve the HOI detection performance [9], [42]. Current HOI approaches, e.g., GEN-VLKT [9] and HOICLIP [35], took use of CLIP [43] to train the encoding network via image-text pairs.

Weakly Supervised HOI Detection. Weakly supervised HOI detection generally uses image-level interaction labels for training [23], [24], [44]. They can be divided into twofolds: "weakly+" with ⟨"interaction", "object"⟩ annotations [24], [26], [44], and "weakly" with only ⟨"interaction"⟩ annotations [23]. MX-HOI [24] proposed a momentum independent learning framework using weakly+ supervised (⟨"interaction", "object"⟩). Align-Former [44] proposed an "align layer" to achieve pseudo alignment for training based on transformer architecture. Nevertheless, these methods suffer from noisy human-object association and ambiguous interaction types [26]. Therefore, PGBL [26] used CLIP as an interactive prompt network and HOI knowledge to enhance interaction judgment at the instance level. Weakly-HOI [23] applied a language model to query for unnecessary interaction pairing reduction with image-level ⟨"interaction"⟩ annotations. Both of them, however, are still dependent on pre-annotated datasets at the cost of manpower. Unlike the latest methods [23] and [26], we propose a method for HOI detection that does not require manual annotation.

Zero-shot HOI Detection. The goal of zero-shot HOI detection is to train on a subset of labels and test using another set of unseen labels to detect interactions that were not encountered during training. Many methods [9], [15], [18], [27]–[36] are investigated to handle zero-shot HOI detection. Shen et al. [32] pioneered the application of zero-shot learning methods to address the long-tail problem in HOI detection. They introduced a decomposition model for HOI detection, separately inferring verbs and objects, enabling the detection of new verb-object combinations during testing. S2S [36] embedded both semantic and spatial information into the visual stream, proposing a person-object interaction detection based on verb-object relation reasoning. GEN-VLKT [9] is a simple yet effective framework that utilizes CLIP for knowledge transfer [23], [26], [35], thereby discovering unknown samples. The introduction of zero-shot learning enhances the adaptability of these methods to real-world scenarios. However, these methods still require manually annotated complete HOI labels.

III. Method

As shown in Figure 2, we propose a novel plug-and-play HOI framework, namely, FreeA, to reduce the requirements of annotations. The proposed framework includes candidate image construction (Sec. III-B), human-object potential relationship mining (Sec. III-C), and human-object relationship inference (Sec. III-D). Details are introduced below.

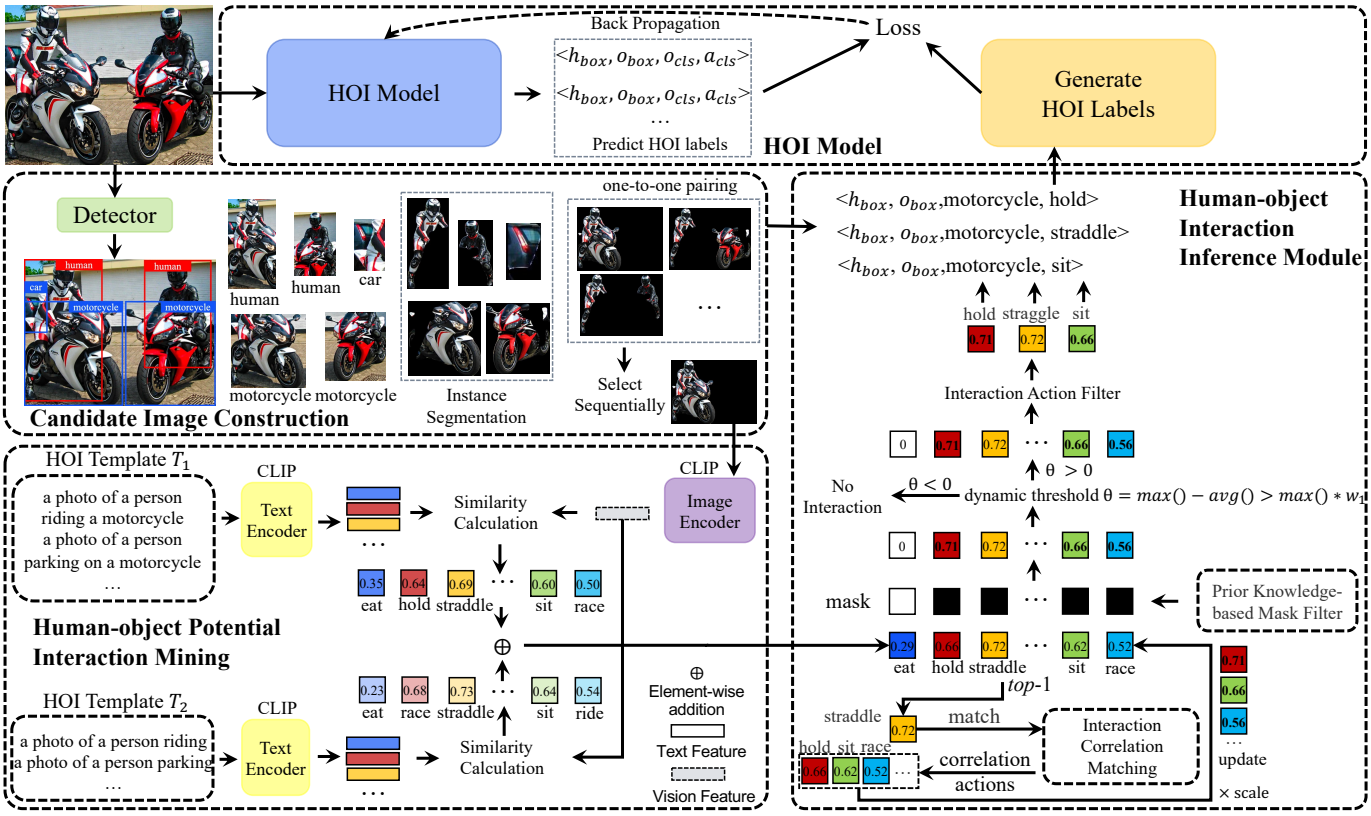

Fig. 2. Method overview. Starting from an existing HOI model, we apply the candidate image construction to extract humans and objects by detection and segmentation, and establish one-to-one human-object pairing. The human-object potential interaction mining module gets initial HOI interaction labels from candidate image pairs, and uses CLIP for domain adaptation. The human-object interaction inference module further refines these interaction labels by using prior knowledge-based mask to eliminate implausible actions and using interaction correlation matching method to enhance relevant actions similarity. Finally, HOI labels are generated for model training through dynamic threshold selector and interaction action filter.

A. HOI Model

Motivated by the success of GEN-VLKT [9], we use it as the HOI model shown in Figure 2, and training is initialized with the parameters of COCO trained DETR [41]. It is worth noting that a better HOI model can promote the overall effect of FreeA, as verified in Sec IV-B.

B. Candidate Image Construction

To achieve automatic labels generation, we need to accurately localize humans and objects before, where Yolov8[2] model is used for localization. Given an input image I, we obtain a collection of instance bounding boxes $\boldsymbol{B} = \{(\boldsymbol{b}_i | i = 1, 2, ..., N)\}$, where $\boldsymbol{b}_i = (x_i, y_i, W_i, H_i, c_i)$ from Yolov8, with x_i and y_i representing the center coordinates of the ith bounding box. W_i and H_i are the width and height of the bounding box, and c_i denotes the category of the instance within the box. Here, N represents the number of detected instances.

Subsequently, we trim the images using the bounding boxes and then pair humans with objects to create candidate images. N_I candidate images are obtained, where $N_I = N_h \times N_o$, and $N_h + N_o = N$. N_h and N_o represent the number of humans and objects, respectively. To eliminate interference from redundant instances and background information, we apply instance segmentation to assist the PIM module to focus on interest of interactions.

[2]https://github.com/ultralytics/ultralytics

C. Human-object Potential Interaction Mining

As shown in Figure 3, the PIM module aims to mine potential relationships between humans and objects. The pre-trained CLIP is used to transfer knowledge from source domain to target domain, where features are extracted from both texts and images. Then by computing cross-modal similarity, text-image pairs are matched.

CLIP Image Encoder. The image encoder $\mathbf{F}_{IE}$ of CLIP is employed to process the set of N_I candidate images, with the result of image encoding $\boldsymbol{I}_E \in \mathbb{R}^{N_I \times C_{IE}}$. The $\boldsymbol{C}_{IE}$ denotes the dimension of the image encoder, and $\boldsymbol{I}_B$ represents the collection of candidate images. We have:

$$\boldsymbol{I}_E = \mathbf{F}_{IE}(\boldsymbol{I}_B). \tag{1}$$

CLIP Text Encoder. We start with a text template creation, denoted as $\boldsymbol{T}_1$, in the format "a photo of a person verbing an object". For example, the triplet ⟨ "human", "ride", "motorcycle"⟩ is transformed into "a photo of a person riding a motorcycle". After that, $\boldsymbol{T}_1$ is input into the text encoder $\mathbf{F}_{TE}$, leading to a text information matrix $\boldsymbol{T}_{E1} \in \mathbb{R}^{N_T \times C_{TE}}$. The N_T represents the number of texts, i.e., the number of HOI interaction action categories, and C_{TE} is the dimensionality of the encoded text.

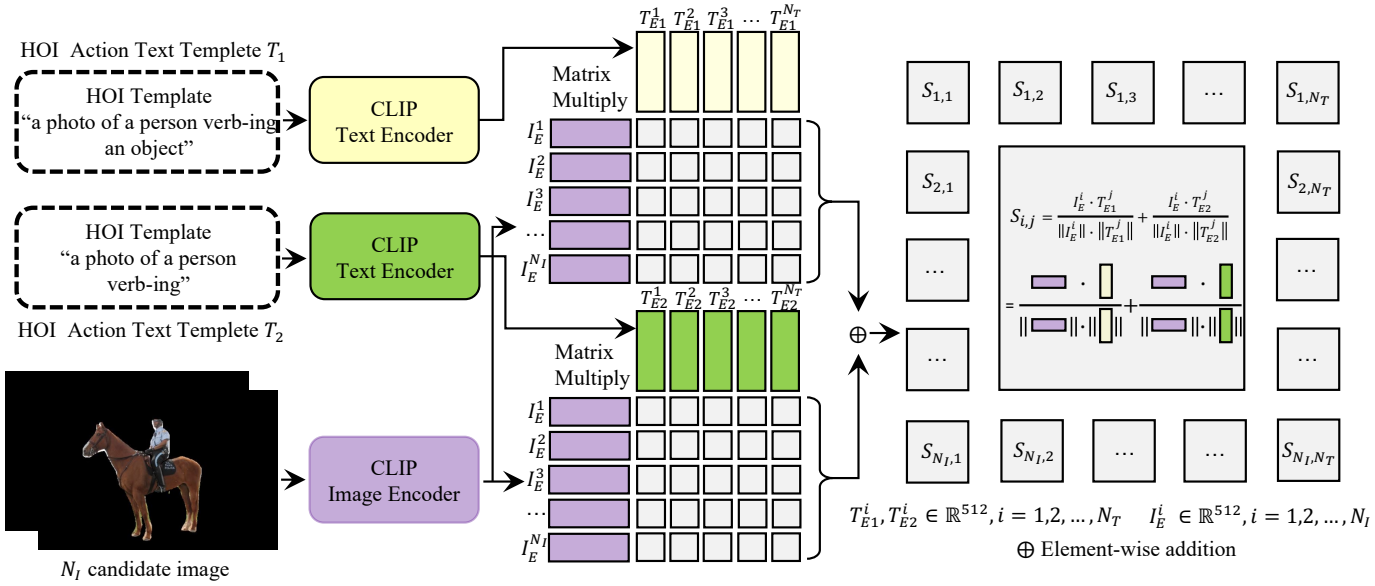

Fig. 3. Human-object potential interaction mining module. The candidate image and HOI action text prompt templates are separately fed into the image encoder and text encoder of CLIP to obtain high-level feature encodings. The image encoding, denoted as I_E, is paired with the text encodings T_{E1} and T_{E2} to calculate textual similarities. The similarity calculations are specified by Eq. 3 and 4.

To emphasize the importance of verbs in HOI relationships, another type of text template $\boldsymbol{T}_2$ in the format "a photo of a person verb-ing", has been constructed. $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ are distinct text templates for different HOI actions, with corresponding to information matrices $\boldsymbol{T}_{E1}$ and $\boldsymbol{T}_{E2}$, respectively. They are formulated as:

$$\boldsymbol{T}_{Ei} = \mathbf{F}_{TE}\left(\boldsymbol{T}_i\right), i = 1, 2. \tag{2}$$

Next, we calculate the cosine similarity between the image encoding information $\boldsymbol{I}_E$ and the text information $\boldsymbol{T}_E$. That is:

$$\boldsymbol{sim}_{Ei}(\boldsymbol{I}_E, \boldsymbol{T}_{Ei}) = \frac{\boldsymbol{I}_E \cdot \boldsymbol{T}_{Ei}^T}{\|\boldsymbol{I}_E\| \cdot \|\boldsymbol{T}_{Ei}\|}, i = 1, 2, \tag{3}$$

$$\boldsymbol{S} = \boldsymbol{sim}_{E1} + \boldsymbol{sim}_{E2}, \tag{4}$$

where $\boldsymbol{I}_E \in \mathbb{R}^{N_I \times C_{IE}}$, $\boldsymbol{T}_{Ei} \in \mathbb{R}^{N_T \times C_{TE}}$, $C_{IE} = C_{TE}$, $\boldsymbol{sim}_{Ei} \in \mathbb{R}^{N_I \times N_T}$, and $\boldsymbol{S} \in \mathbb{R}^{N_I \times N_T}$.

D. Human-object Interaction Inference

The HII module consists of interaction correlation matching, prior knowledge-based mask filter, dynamic threshold selector, and interaction action filter.

Interaction Correlation Matching (ICM). If a certain action occurs, other actions may also occur concurrently. For instance, when a person "racing a motorcycle", he is also "riding and sitting on the motorcycle". Inspired by that, we propose interaction correlation matching to infer other behaviors strongly correlated with the *top*-1 selected initial interaction action, as shown in Figure 4. When "race" is selected based on the highest similarity, we will also extract highly correlated actions, such as "ride", "straddle", "sit", etc. Afterward, we amplify the similarity of the selected action and update the similarity vector. To be specific, for each row vector in $\boldsymbol{S}$, we employ a *top*-1 selection strategy to choose the initial interaction action with the highest image-text similarity.

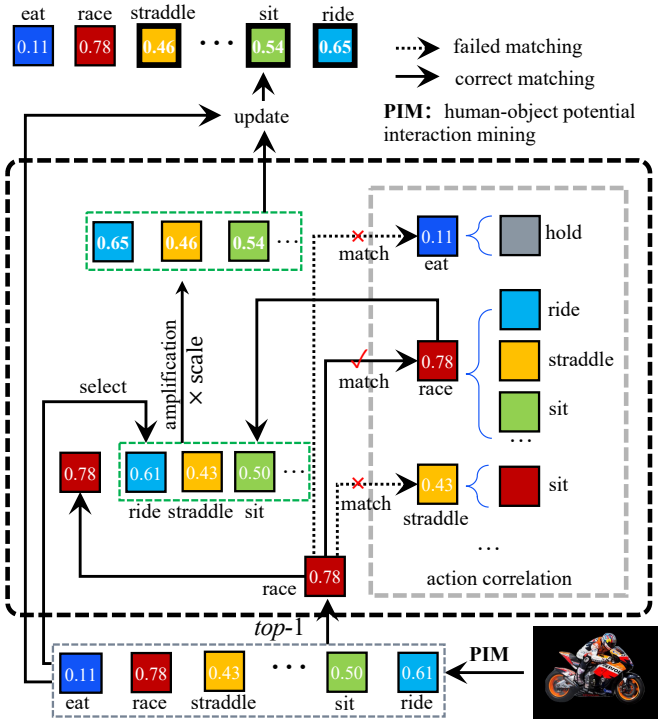

Fig. 4. Details of the interaction correlation matching method. After selecting the action with the highest similarity, we perform action correlation extraction. For example, if "race" is the selected action, related actions such as "ride", "straddle", and "sit" will be identified. We then amplify the similarities of these related actions, making actions highly correlated with "race" more likely to be chosen in the subsequent steps.

Furthermore, we amplify image-text similarity to highlight the similarity between these interaction actions. Detailed steps for applying interaction correlation matching to a single candidate image are:

$$\boldsymbol{k}_a = [j_1, j_2, ..., j_n] = \mathbf{F}_{ICM}(\boldsymbol{k}_{max}(\boldsymbol{S}_i)), \tag{5}$$

$$\boldsymbol{S}_{ij} = \boldsymbol{S}_{ij} \times scale,\ j \in \boldsymbol{k}_a, \tag{6}$$

where $\boldsymbol{k}_{max}(\boldsymbol{S}_i)$ represents the index of interaction action with the highest image-text similarity in the ith candidate image, the function $\mathbf{F}_{ICM}$ selects several other behaviors strongly correlated with $\boldsymbol{k}_{max}(\boldsymbol{S}_i)$, $\boldsymbol{k}_a$ denotes a set of indexes associated with several interaction actions correlated with $\boldsymbol{k}_{max}(\boldsymbol{S}_i)$, and $scale$ denotes the amplification factor.

Prior Knowledge-based Mask (PKM) Filter. As widely acknowledged, specific objects often exhibit clear associations with particular action categories. For instance, common interaction actions with an "apple" include "pick" and "eat", while "ride" or "drive" are highly unlikely. Inspired by that we design *a priori* knowledge-based mask filter, which uses a specific mask mechanism to filter interaction actions for specific objects based on prior knowledge. The similarity score $\boldsymbol{S}_{ij}$ is updated as:

$$\boldsymbol{k}_o = [j_1, j_2, ..., j_n] = \mathbf{F}_{PKM}(\boldsymbol{o}_i), \tag{7}$$

$$\begin{aligned} \boldsymbol{mask} &= [m_1, m_2, \ldots, m_j, \ldots, m_{N_T}] \\ &= \begin{cases} 0, j \notin \boldsymbol{k}_o \\ 1, j \in \boldsymbol{k}_o \end{cases}, j = 1, 2, \ldots, N_T, \end{aligned} \tag{8}$$

$$\boldsymbol{S}_{ij} = \boldsymbol{S}_{ij} \times \boldsymbol{mask}_j, j = 1, 2, ..., N_T, \tag{9}$$

where $\boldsymbol{o}_i$ represents the object category in the ith image. The function $\mathbf{F}_{PKM}$ selects the indexes of all interaction actions related to $\boldsymbol{o}_i$, denoted as $\boldsymbol{k}_o$. m_j = {1, 0}: 1 indicates that the jth action is related to $\boldsymbol{o}_i$, not the other way around. Therefore, we retain interaction actions related to specific object categories after reducing interference from unlikely actions.

Dynamic Threshold Selector. A dynamic threshold selector is employed to assess whether interaction has occurred in a candidate image. Our starting point is that when the difference between the maximum value and the mean value of $\boldsymbol{S}_i$ is greater than the maximum value of $\boldsymbol{S}_i$ multiplied by a weighting factor, it is considered that the maximum value of $\boldsymbol{S}_i$ has a significant gap with the other values. This indicates that there may be an interaction (when there is interaction in the image, the similarity of the specific action tends to be high, while the similarity of unrelated actions tends to be low). The calculation formula is written as:

$$\theta = (max\left(\boldsymbol{S}_i\right) - \frac{\sum_{j=1}^{N_T} \boldsymbol{S}_{ij}}{N_T}) - max(\boldsymbol{S}_i) \times \omega_1, \tag{10}$$

where $max(\boldsymbol{S}_i)$ is the highest image-text similarity value in the ith candidate image, and ω_1 is a parameter to balance the threshold range. A positive θ ($\theta > 0$) indicates the presence of human-object interaction in ith candidate image. This dynamic threshold adjustment enhances the accuracy of interaction relationship detection and recognition for different scenarios, further leading to more precise event determination.

Interaction Action Filter. In the presence ($\theta > 0$) of interaction in the candidate image, we employ an interaction action filtering to select target actions from N_T interaction actions. The filtering procedure is expressed as:

$$\begin{aligned} &\boldsymbol{a}_{index} \\ &= \{j | (max\left(\boldsymbol{S}_i\right) - max\left(\boldsymbol{S}_i\right) \times \omega_2) < \boldsymbol{S}_{ij} < max\left(\boldsymbol{S}_i\right), \\ &\boldsymbol{S}_{ij} \in \boldsymbol{S}_i\}, \end{aligned} \tag{11}$$

where j represents the index of each interaction action, and $\boldsymbol{a}_{index}$ denotes the set of action indices. Finally, the HOI labels $\mathcal{O}$ are built as:

$$\mathcal{O} = \{(\boldsymbol{h}_{box}, \boldsymbol{o}_{box}, c_o, \boldsymbol{a}_i) | \boldsymbol{a}_i \in \boldsymbol{a}_{index}\}, \tag{12}$$

where $\boldsymbol{h}_{box}$ and $\boldsymbol{o}_{box}$ are the detected bounding boxes of the human and object entities, respectively. c_o denotes the object category, and $\boldsymbol{a}_i$ represents the interaction action index.

The total loss function is consistent with that of GEN-VLKT [9], defined as:

$$\mathcal{L} = \lambda_b \sum_{i \in (h,o)} \mathcal{L}_b^i + \lambda_u \sum_{j \in (h,o)} \mathcal{L}_u^j + \sum_{k \in (o,a)} \lambda_c^k \mathcal{L}_c^k, \tag{13}$$

where $\mathcal{L}_b$, $\mathcal{L}_u$, and $\mathcal{L}_c$ are box regression loss, IoU loss, and classification loss, respectively. λ_b, λ_u and λ_c^k are the hyper-parameters for adjusting the weights of each loss.

IV. Experiments

Datasets. The benchmark datasets, HICO-Det and V-COCO, are used to demonstrate the effectiveness of the proposed method. HICO-Det includes 47,776 images with 38,118 for training and 9,658 for testing, and covers 80 object categories, 117 action categories, and 600 distinct interaction types. V-COCO comprises 10,326 images with 5,400 for training and 4,964 for testing, featuring 80 object categories and 29 action categories, including 4 body actions without object interactions. Both datasets encompass a wide range of object and action categories, making them essential for evaluating HOI detection algorithms.

Evaluation Metric. The mean Average Precision (mAP) is used as the evaluation metric. The correct HOI triplet prediction should satisfy: 1) the predicted human and object bounding boxes have an IoU (Intersection over Union) with the ground truth greater than 0.5, and 2) the predicted category is correct. Evaluation on the HICO-Det dataset encompasses three modes: Full (600 HOI categories), Rare (actions with fewer than 10 training instances), and Non-Rare (the rest HOI categories). For the V-COCO dataset. We assess two scenarios: S1 with 29 action categories (including 4 body actions) and S2 with 25 action categories.

Implementation Details. The learning rate of HOI model is 1e-4. The optimizer is AdamW with a weight decay rate of 1e-4. The encoder has 6 layers, and the decoder has 3 layers. The total number of training epochs is 90, with a learning rate drop of ten times after the 60th epoch. The ω_1 in Eq. 10 and ω_2 in Eq. 11 are set to 0.23 and 0.1, respectively. All experimental trials are executed on 2 NVIDIA A800 (80G) GPUs, employing a batch size of 16. The computational environment runs Ubuntu 22.04, with Python version 3.8, PyTorch version 1.7.1, torchvision version 0.8.2, and CUDA version 11.0.

TABLE I
PERFORMANCE COMPARISONS ON HICO-DET DATASET. * REPRESENTS THE RESULTS GIVEN IN [23], [26].

Methods	Backbone	Source	Default (**mAP**↑)			Know Object (**mAP**↑)		
			Full	Rare	None-Rare	Full	Rare	None-Rare
fully supervised (using ⟨"human", "interaction", "object" ⟩ labels)								
InteractNet [13]	ResNet-50-FPN	CVPR 2018	9.94	7.16	10.77	-	-	-
iCAN [45]	ResNet-50	BMCV 2018	14.84	10.45	16.15	16.26	11.33	17.73
PMFNet [46]	ResNet-50-FPN	ICCV 2019	17.46	15.56	18.00	20.34	17.47	21.20
DJ-RN [47]	ResNet-50	CVPR 2020	21.34	18.53	22.18	23.69	20.64	24.60
IDN [48]	ResNet-50	NeurIPS 2020	23.36	22.47	23.63	26.43	25.01	26.85
HOTR [49]	ResNet-50	CVPR 2021	25.10	17.34	27.42	-	-	-
QPIC [40]	ResNet-101	CVPR 2021	29.90	23.92	31.69	32.38	26.06	34.27
HRNet [20]	ResNet-152	TIP 2021	21.93	16.30	23.62	25.22	18.75	27.15
MSTR [50]	ResNet-50	CVPR 2022	31.17	25.31	33.92	34.02	28.83	35.57
DisTr [51]	ResNet-50	CVPR 2022	31.75	27.45	33.03	34.50	30.13	35.81
RCL [52]	ResNet-50	CVPR 2023	32.87	28.67	34.12	35.52	30.88	36.45
GEN-VLKT [9]	ResNet-50	CVPR 2022	33.75	29.25	35.10	36.78	32.75	37.99
PBLQG [11]	ResNet-50	TIP 2023	31.64	26.23	33.25	34.61	30.16	35.93
SG2HOI [12]	ResNet-50	TIP 2023	33.14	29.27	**35.72**	35.73	32.01	36.43
TED-Net [10]	ResNet-50	TCSVT 2024	**34.00**	**29.88**	35.24	**37.13**	**33.63**	**38.18**
weakly+ supervised (using ⟨"interaction", "object" ⟩ labels)								
Explanation-HOI* [53]	ResNeXt101	ECCV 2020	10.63	8.71	11.20	-	-	-
MAX-HOI [24]	ResNet101	WACV 2021	16.14	12.06	17.50	-	-	-
Align-Former [44]	ResNet-101	arXiv 2021	20.85	18.23	21.64	-	-	-
PPR-FCN* [25]	ResNet-50	ICCV 2017	17.55	15.69	18.41	-	-	-
Weakly-HOI [23]	ResNet-50	CVPR 2023	19.26	-	-	-	-	-
PGBL [26]	ResNet-50	ICLR 2023	22.89	**22.41**	23.03	-	-	-
Ours	ResNet-50	-	**24.57**	21.45	**25.51**	**26.52**	**23.64**	**27.38**
weakly supervised (using ⟨"interaction" ⟩ labels)								
SCG* [54]	ResNet-50	ICCV 2021	7.05	-	-	-	-	-
Weakly-HOI [23]	ResNet-50	CVPR 2023	8.38	-	-	-	-	-
Ours (no labels)	ResNet-50	-	**16.96**	**16.26**	**17.17**	**18.89**	**18.11**	**19.12**

TABLE II
PERFORMANCE COMPARISONS ON V-COCO DATASET. * REPRESENTS THE RESULTS GIVEN IN [23]. WEAKLY-HOI† IS TRAINED VIA VIDEO CAPTURED LABELS.

Method	Source	AP^{S1}_{role} (**mAP**↑)	AP^{S2}_{role} (**mAP**↑)
weakly+ supervised			
Explanation-HOI [53]	ECCV 2020	-	-
MAX-HOI [24]	WACV 2021	-	-
PPR-FCN [25]	ICCV 2017	-	-
Align-Former [44]	arXiv 2021	15.82	16.34
PGBL [26]	ICLR 2023	42.97	48.06
Ours	-	**50.25**	**52.05**
weakly supervised			
SCG* [54]	ICCV 2021	20.05	-
Weakly-HOI [23]	CVPR 2023	29.59	-
Weakly-HOI† [23]	CVPR 2023	17.71	-
Ours (no labels)	-	**30.82**	**32.60**

A. Effectiveness for Regular HOI Detection

As illustrated in Table I, we conduct comparisons under different supervised scenarios using the HICO-DET dataset. In contrast to the state-of-the-art (SOTA) weakly supervised Weakly-HOI method with image-level ⟨"interaction"⟩ labels for training, our method demonstrates a relative 8.58 mAP improvement in the absence of manual labels. Furthermore, we extend our method to weakly+ supervised field. Compared with the SOTA weakly+ supervised PGBL method, our method achieves better performance in both full and non-rare scenarios under default setting, with a performance improvement of 1.68 mAP and 2.48 mAP, respectively. It is noteworthy that our approach surpasses several fully supervised HOI models (e.g., InteractNet, iCAN, IDN). As shown in Figure 5, the proposed FreeA can almost generate comparable HOI labels with ground truth.

Experimental results on the V-COCO dataset are presented in Table II. The proposed FreeA is far ahead of Weakly-HOI† method using utilized video captured labels, where the result increases from 17.71 mAP to 30.82 mAP in terms of AP^{S1}_{role}. As well, FreeA surpasses the Weakly-HOI model by achieving a 1.23 mAP increase at AP^{S1}_{role}. Moreover, FreeA achieves a 7.28 mAP increase at AP^{S1}_{role} as compared to the SOTA weakly+ supervised PGBL model.

B. Ablation Studies

1) HOI model: To evaluate the effectiveness of HOI model to the FreeA, four up-to-date fully supervised HOI approaches are tested, as shown in Table III. It is observed that a better HOI model can promote the overall effect of FreeA. For example, the QPIC model, with a 29.07 mAP in terms of full under fully supervised, achieves a 20.18 mAP in the weakly+ supervised. When we employ a superior HOI model, such as GEN-VLKT, it reaches 24.57 mAP in the weakly+ supervised. In this work, the GEN-VLKT model is applied as the HOI baseline model.

2) HOI text templates: We conducted additional ablation studies to investigate various components of in the FreeA framework, and the results have been tabulated in Table IV. The experimental results show that both HOI text templates T_1 and T_2 play a vital role in HOI detection, at a 1.14 mAP increase as compared to FreeA with T_1 (Row 1 and Row 2 in Table IV). This is mainly because we observe that T_1 text template ("a photo of a person verb-ing an object) does not

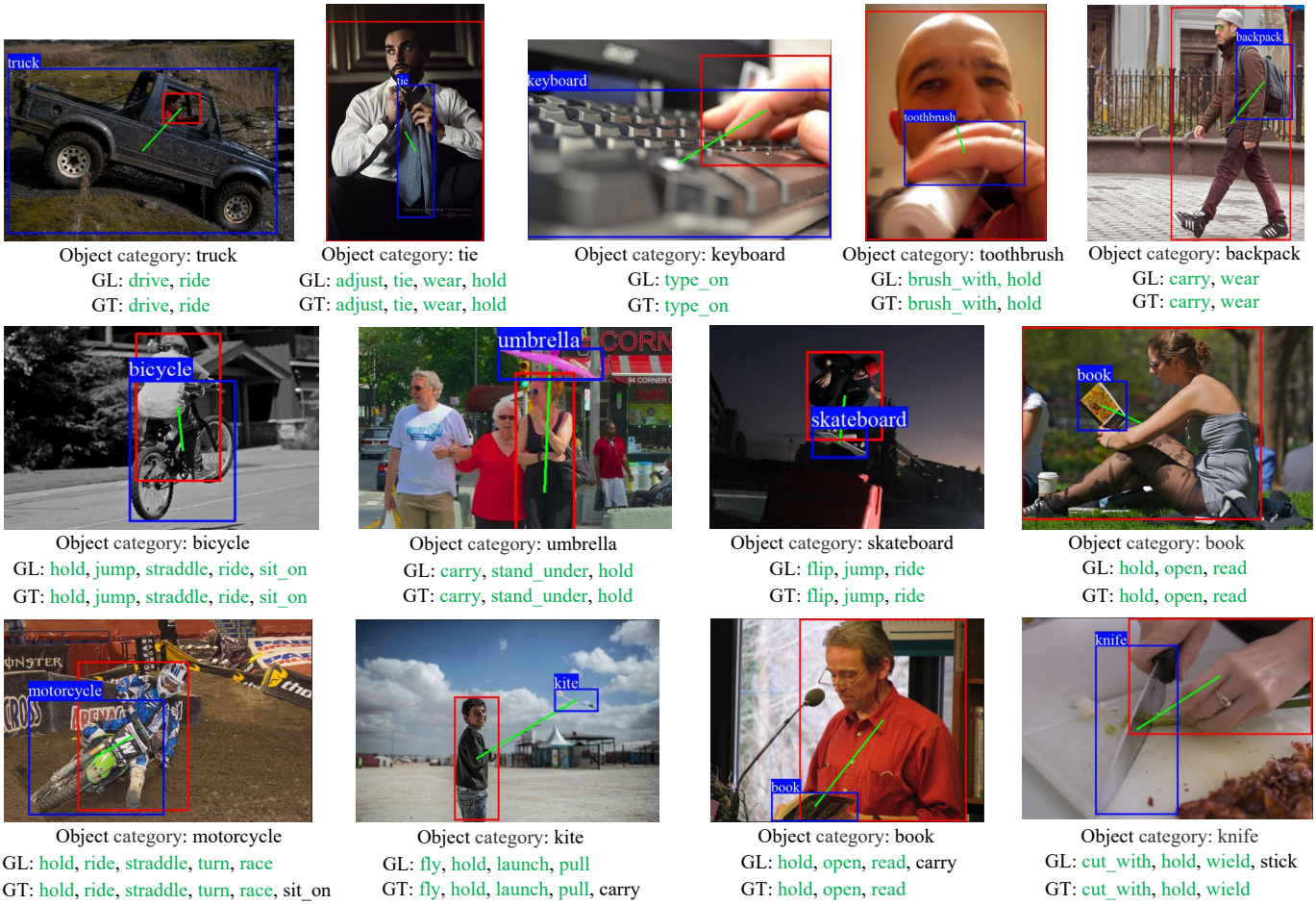

Fig. 5. Comparison of HOI labels. GL represents labels generated by our method, and GT represents ground truth. In each sub image, the red and blue rectangles are bounding boxes for the human and object, and the green lines represent the connection between their centers. Green text indicates correct interactions.

TABLE III
ABLATION STUDY USING DIFFERENT HOI MODELS IN WEAKLY+ SUPERVISED ON HICO-DET DATASETS. τ REPRESENTS PERFORMANCE UNDER FULL SUPERVISION.

Method	Source	Defalut (**mAP**↑)		
		Full	Rare	Non-Rare
QPIC [40]	CVPR 2021	20.18 29.07$^{\tau}$	15.82 21.85$^{\tau}$	21.55 31.23$^{\tau}$
STIP [55]	CVPR 2022	21.43 31.60$^{\tau}$	19.03 27.75$^{\tau}$	22.26 32.75$^{\tau}$
RCL [52]	CVPR 2023	23.24 32.87$^{\tau}$	19.74 28.67$^{\tau}$	24.38 34.12$^{\tau}$
GEN-VLKT [9]	CVPR 2022	**24.57** 33.75$^{\tau}$	**21.45** 29.25$^{\tau}$	**25.51** 35.10$^{\tau}$

capture significant differences in similarity between different actions on the same object. Therefore, we introduce the T_2 text template, which emphasizes the action ("a photo of a person verb-ing").

3) Action selection approaches: Two action selection approaches, namely, "*top*-1" and "adaption", are designed to determine which action should be selected when human-object interactions are not present in an image (Row 2 and Row 3 in Table IV). The "*top*-1" refers to selecting the most salient action, whereas "adaption" (Eq. 11) retains actions within a specified threshold range. The results show that the "adaption" approach outperforms the "*top*-1" approach at +0.86 mAP. The *top*-1 only selects the action with the highest similarity, however, HOI typically involve interactions with multiple actions, and the "adaptation" method can satisfy this to provide multiple choices.

4) Dynamic threshold: We further test the effect of dynamic threshold θ (Eq. 10) to the FreeA (Row 3 and Row 4 in Table IV). It is observed that using dynamic threshold instead of fixed threshold results in a 0.96 mAP improvement. This indicates that the variability in the subtraction of the average similarity of all actions from the highest similarity action obtained through CLIP is significant, and fixed threshold cannot overcome this problem.

5) Background retention or deletion: The image background can be a double-edged sword. Sometimes it works in your favour when for simple image background, sometimes it works against you when one includes complex background details leading to different interferences. We conducted experiments to verify the effect of image background. The results indicate that retaining the background leads to performance decrease (Row 4 and Row 5 in Table IV).

TABLE IV
ABLATION STUDY USING DIFFERENT MODULES ON HICO-DET DATASETS.

Row	T_1	T_2	*top*-1	Adaption	Dynamic threshold	Segmentation (retain background)	Segmentation (delete background)	ICM	**mAP**↑ (Full)
1	✓		✓			✓			12.87
2	✓	✓	✓			✓			14.01
3	✓	✓		✓		✓			14.87
4	✓	✓		✓	✓	✓			15.83
5	✓	✓		✓	✓		✓		16.14
6	✓	✓		✓	✓		✓	✓	**16.96**

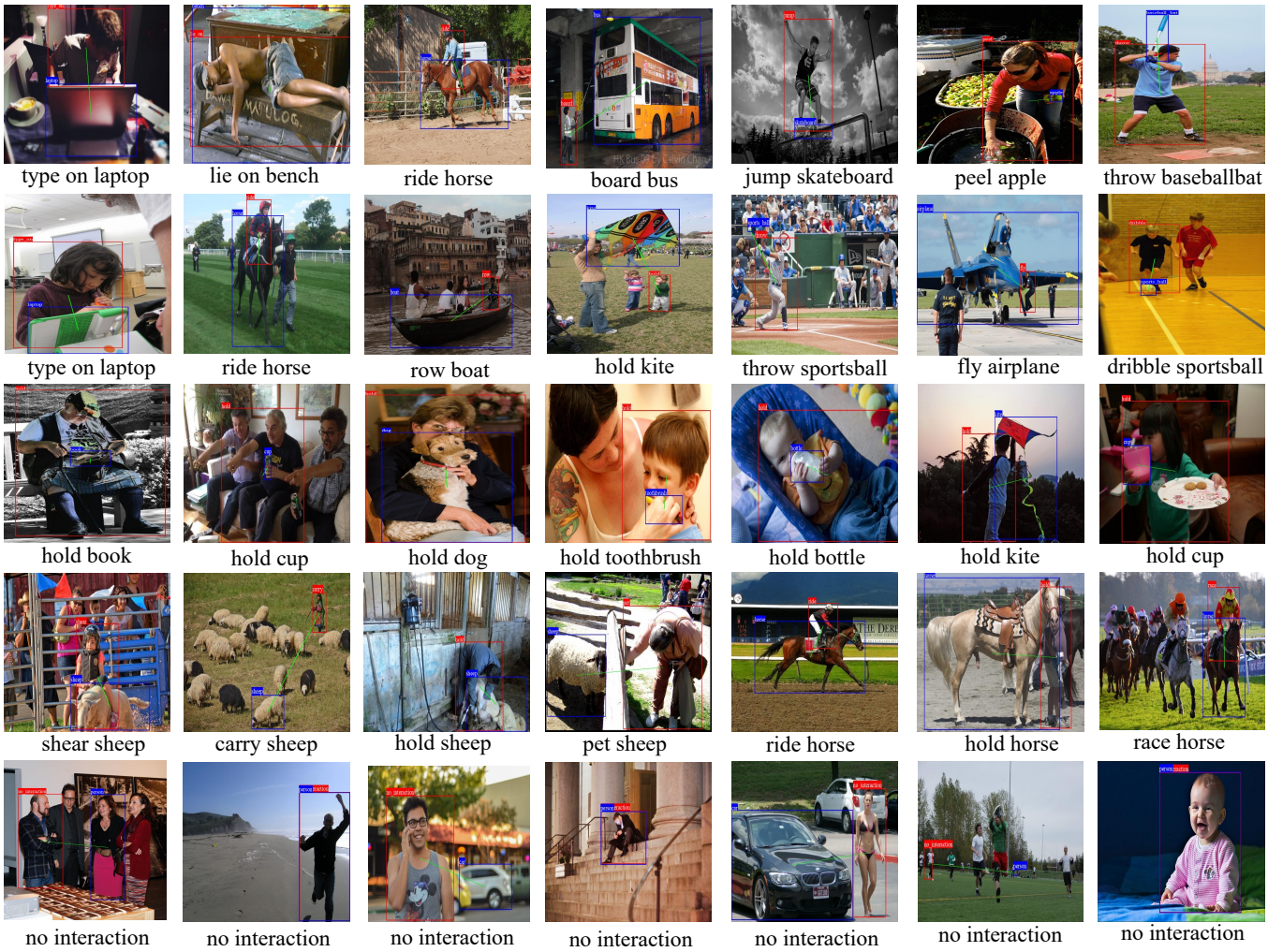

Fig. 6. HOI detection samples in HICO-Det. We retain only the human-object interaction pairs with the highest probability of interaction in the images. The first two rows showcase samples of HOI for different actions. The third row displays HOI samples for the same action with different objects. The fourth row illustrates HOI samples for different actions with the same object. The last row demonstrates that our method can also identify examples with no interactions.

6) The ratio of background: The effects of retention and deletion background information are presented in Table V. The ratio of 1.0:1.0 represents feeding two types of images separately into CLIP and then combine the final similarity results in a 1:1 ratio. Through experiments, it is found that as the proportion of deletion background image increased, the mAP results of FreeA also increased continuously. The last row in Table V indicates using only deletion background image, and it can be observed that in this case, FreeA achieves the the best performance.

7) ICM: Regarding the proposed ICM component, the experimental results demonstrate an improvement of 0.82 mAP when ICM is utilized compared to when it is not (Row 5 and Row 6 in Table IV). The improvement is foreseeable because the ICM module emphasizes the correlation between actions.

8) The effectiveness for various componets of FreeA in V-COCO: We also conduct ablation experiments of each component of FreeA on the V-COCO, as shown in Table VI. Similar to the results on the HICO-Det, on the V-COCO dataset, using two text templates (T_1 and T_2) performs better than using only

Fig. 7. HOI detection samples in V-COCO. The first two rows show HOI samples for different actions involving different objects. The third and fourth rows depict HOI samples for the same action with different objects and different actions with the same object, respectively.

TABLE V
ABLATION STUDIES USING DIFFERENT SEGMENTATION IMAGES ON HICO-DET DATASETS. THE "()" REPRESENTS THE SUM RATIO OF THE RESULTS. FOR EXAMPLE, "1.0:1.0" MEANS THAT TWO TYPES OF IMAGES ARE INPUT INTO CLIP, AND THE RESULTS ARE COMBINED IN A 1:1 RATIO.

Segmentation (retrain background)	Segmentation (delete background)	**mAP↑**
✓(1.0)	✓(1.0)	16.07
✓(0.5)	✓(1.5)	16.37
✓(0.25)	✓(1.75)	16.41
×(0.0)	✓(2.0)	**16.96**

the T_1 text template (Row 1 and Row 2), and the "adaption" strategy is more effective than "*top*-1" (Row 2 and Row 3). The dynamic threshold method is also proven effective on the V-COCO (Row 3 and Row 4). Regarding the issue of whether to retain the background, the results on the V-COCO also indicate better performance when deleting background information (Row 4 and Row 5). Finally, by introducing the ICM module, the experimental results are further improved (Row 5 and Row 6).

V. VISUALIZATION

We visualize some results of generated labels compared with ground truth HOI labels, as shown in Figure 5. Most of the time the generated labels match the ground truth labels perfectly (row1 and row2).

We visualize the proposed FreeA method on two datasets, HICO-det and V-COCO, without using any manually annotated labels. For clarity, we retain only the human-object interaction pairs with the highest probability of interaction in the images. Figure 6 shows the visualizations of FreeA on HICO-Det, and Figure 7 presents the visualizations of FreeA on V-COCO. As shown in Figure 6, the first two rows display the detection results of FreeA on different actions. The third row lists cases where the same action is performed with different objects. The fourth row visualizes different actions involving the same object, and the last row illustrates the detection performance of FreeA in the absence of interacting human-object pairs. This demonstrates that FreeA can effectively detect various types of interactions.

For the V-COCO, the first two rows of Figure 7 visualize the detection results for different actions. The third and fourth rows show cases of the same action with different objects and the same object with different actions, respectively.

VI. CONCLUSION

We propose a novel weakly-- supervised HOI detection method, termed as FreeA. Weakly-- supervised FreeA means the training labels are not manually annotated from the raw datasets, but automatically generated from CLIP with the combination of candidate image construction, human-object potential interaction mining and human-object interaction inference modules. As compared to those weakly, weakly+, and

TABLE VI
ABLATION STUDY USING DIFFERENT MODULES ON V-COCO DATASETS.

Row	T_1	T_2	top-1	Adaption	Dynamic threshold	Segmentation (retain background)	Segmentation (delete background)	ICM	mAP↑ AP^{S1}_{role}	(AP^{S2}_{role})
1	✓		✓			✓			25.06	(23.77)
2	✓	✓	✓			✓			25.14	(24.67)
3	✓	✓		✓		✓			27.01	(28.18)
4	✓	✓		✓	✓	✓			27.30	(28.19)
5	✓	✓		✓	✓		✓		30.11	(31.55)
6	✓	✓		✓	✓		✓	✓	**30.82**	(**32.60**)

fully supervised HOI methods, extensive experiments have demonstrated the effectiveness and advantages of the proposed FreeA. Our contributions to the field include presenting a new problem of weakly supervised HOI detection, and showing the utilization of CLIP model for generating HOI labels.

REFERENCES

[1] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 381–389.

[2] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. 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*, 2020, pp. 482–490.

[3] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, "Learning human-object interaction detection using interaction points," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4116–4125.

[4] X. Zhong, X. Qu, C. Ding, and D. 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*, 2021, pp. 13 234–13 243.

[5] B. Kim, T. Choi, J. Kang, and H. J. Kim, "UnionDet: Union-level detector towards real-time human-object interaction detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 498–514.

[6] H.-S. Fang, Y. Xie, D. Shao, and C. Lu, "DIRV: Dense interaction region voting for end-to-end human-object interaction detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1291–1299.

[7] M. Chen, Y. Liao, S. Liu, Z. Chen, F. Wang, and C. Qian, "Reformulating HOI detection as adaptive set prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9004–9013.

[8] C. Zou, B. Wang, Y. Hu, J. Liu, Q. Wu, Y. Zhao, B. Li, C. Zhang, C. Zhang, Y. Wei *et al.*, "End-to-end human object interaction detection with HOI transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 825–11 834.

[9] Y. Liao, A. Zhang, M. Lu, Y. Wang, X. Li, and S. 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*, 2022, pp. 20 123–20 132.

[10] Y. Wang, Q. Liu, and Y. Lei, "Ted-net: Dispersal attention for perceiving interaction region in indirectly-contact hoi detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[11] W.-K. Lin, H.-B. Zhang, Z. Fan, J.-H. Liu, L.-J. Yang, Q. Lei, and J. Du, "Point-based learnable query generator for human–object interaction detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 6469–6484, 2023.

[12] T. He, L. Gao, J. Song, and Y.-F. Li, "Toward a unified transformer-based framework for scene graph generation and human-object interaction detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 6274–6288, 2023.

[13] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8359–8367.

[14] Y. Liu, Q. Chen, and A. Zisserman, "Amplifying key cues for human-object-interaction detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 248–265.

[15] A. Bansal, S. S. Rambhatla, A. Shrivastava, and R. Chellappa, "Detecting human-object interactions via functional generalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 10 460–10 469.

[16] A. Iftekhar, S. Kumar, R. A. McEver, S. You, and B. Manjunath, "GTNet: Guided transformer network for detecting human-object interactions," *arXiv preprint arXiv:2108.00596*, 2021.

[17] X. Zhong, C. Ding, X. Qu, and D. Tao, "Polysemy deciphering network for human-object interaction detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 69–85.

[18] Y. Liu, J. Yuan, and C. W. Chen, "ConsNet: Learning consistency graph for zero-shot human-object interaction detection," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4235–4243.

[19] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 401–417.

[20] Y. Gao, Z. Kuang, G. Li, W. Zhang, and L. Lin, "Hierarchical reasoning network for human-object interaction detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 8306–8317, 2021.

[21] D.-J. Kim, X. Sun, J. Choi, S. Lin, and I. S. Kweon, "Acp++: Action co-occurrence priors for human-object interaction detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 9150–9163, 2021.

[22] H. Wang, L. Jiao, F. Liu, L. Li, X. Liu, D. Ji, and W. Gan, "Ipgn: Interactiveness proposal graph network for human-object interaction detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 6583–6593, 2021.

[23] M. E. Unal and A. Kovashka, "Vlms and llms can help detect human-object interactions with weak supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, url: https://asu-apg.github.io/odrum/posters_2023/poster_6.pdf*, 2023.

[24] S. K. Kumaraswamy, M. Shi, and E. Kijak, "Detecting human-object interaction with mixed supervision," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1228–1237.

[25] H. Zhang, Z. Kyaw, J. Yu, and S.-F. Chang, "Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 4233–4241.

[26] B. Wan, Y. Liu, D. Zhou, T. Tuytelaars, and X. He, "Weakly-supervised hoi detection via prior-guided bi-level representation learning," *International Conference on Learning Representations*, 2023.

[27] T. Gupta, A. Schwing, and D. Hoiem, "No-frills human-object interaction detection: Factorization, layout encodings, and training techniques," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9677–9685.

[28] Z. Hou, X. Peng, Y. Qiao, and D. Tao, "Visual compositional learning for human-object interaction detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 584–600.

[29] Z. Hou, B. Yu, Y. Qiao, X. Peng, and D. Tao, "Affordance transfer learning for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 495–504.

[30] ——, "Detecting human-object interaction via fabricated compositional learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 646–14 655.

[31] J. Peyre, I. Laptev, C. Schmid, and J. Sivic, "Detecting unseen visual relations using analogies," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1981–1990.

[32] L. Shen, S. Yeung, J. Hoffman, G. Mori, and L. Fei-Fei, "Scaling human-object interaction recognition through zero-shot learning," in *2018 IEEE*

Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018, pp. 1568–1576.
[33] O. Ulutan, A. Iftekhar, and B. S. Manjunath, "VSGNet: Spatial attention network for detecting human object interactions using graph convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 617–13 626.
[34] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli, "Learning to detect human-object interactions with knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
[35] S. Ning, L. Qiu, Y. Liu, and X. 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*, 2023, pp. 23 507–23 517.
[36] S. Eum and H. Kwon, "Semantics to space (s2s): Embedding semantics into spatial space for zero-shot verb-object query inferencing," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 1384–1391.
[37] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.
[38] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.
[39] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.
[40] M. Tamura, H. Ohashi, and T. 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*, 2021, pp. 10 410–10 419.
[41] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 213–229.
[42] Z. Li, C. Zou, Y. Zhao, B. Li, and S. Zhong, "Improving human-object interaction detection via phrase learning and label composition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 1509–1517.
[43] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.
[44] M. Kilickaya and A. Smeulders, "Human-object interaction detection via weak supervision," *arXiv preprint arXiv:2112.00492*, 2021.
[45] C. Gao, Y. Zou, and J.-B. Huang, "ICAN: Instance-centric attention network for human-object interaction detection," *arXiv preprint arXiv:1808.10437*, 2018.
[46] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "Pose-aware multi-level feature network for human object interaction detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9469–9478.
[47] Y.-L. Li, X. Liu, H. Lu, S. Wang, J. Liu, J. Li, and C. Lu, "Detailed 2d-3d joint representation for human-object interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 166–10 175.
[48] Y.-L. Li, X. Liu, X. Wu, Y. Li, and C. Lu, "HOI analysis: Integrating and decomposing human-object interaction," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5011–5022, 2020.
[49] B. Kim, J. Lee, J. Kang, E.-S. Kim, and H. J. Kim, "HOTR: End-to-end human-object interaction detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 74–83.
[50] B. Kim, J. Mun, K.-W. On, M. Shin, J. Lee, and E.-S. Kim, "Mstr: Multi-scale transformer for end-to-end human-object interaction detection," in *International Conference on Learning Representations*, 2022, pp. 19 578–19 587.
[51] D. Zhou, Z. Liu, J. Wang, L. Wang, T. Hu, E. Ding, and J. Wang, "Human-object interaction detection via disentangled transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 568–19 577.
[52] S. Kim, D. Jung, and M. Cho, "Relational context learning for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2925–2934.
[53] F. Baldassarre, K. Smith, J. Sullivan, and H. Azizpour, "Explanation-based weakly-supervised learning of visual relations with graph networks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*. Springer, 2020, pp. 612–630.
[54] F. Z. Zhang, D. Campbell, and S. Gould, "Spatially conditioned graphs for detecting human-object interactions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 319–13 327.
[55] Y. Zhang, Y. Pan, T. Yao, R. Huang, T. Mei, and C.-W. Chen, "Exploring structure-aware transformer over interaction proposals for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 548–19 557.