

Efficient Adaptive Human-Object Interaction Detection with Concept-guided Memory

Ting Lei¹ Fabian Caba² Qingchao Chen³ Hailin Jin² Yuxin Peng¹ Yang Liu^{1*}

¹Wangxuan Institute of Computer Technology, Peking University

²Adobe Research ³National Institute of Health Data Science, Peking University

{ting_lei, qingchao.chen, pengyuxin, yangliu}@pku.edu.cn

{caba, hljin}@adobe.com

Abstract

Human Object Interaction (HOI) detection aims to localize and infer the relationships between a human and an object. Arguably, training supervised models for this task from scratch presents challenges due to the performance drop over rare classes and the high computational cost and time required to handle long-tailed distributions of HOIs in complex HOI scenes in realistic settings. This observation motivates us to design an HOI detector that can be trained even with long-tailed labeled data and can leverage existing knowledge from pre-trained models. Inspired by the powerful generalization ability of the large Vision-Language Models (VLM) on classification and retrieval tasks, we propose an efficient Adaptive HOI Detector with Concept-guided Memory (ADA-CM). ADA-CM has two operating modes. The first mode makes it tunable without learning new parameters in a training-free paradigm. Its second mode incorporates an instance-aware adapter mechanism that can further efficiently boost performance if updating a lightweight set of parameters can be afforded. Our proposed method achieves competitive results with state-of-the-art on the HICO-DET and V-COCO datasets with much less training time. Code can be found at <https://github.com/ltttku/ADA-CM>.

1. Introduction

Human-object interaction (HOI) detection is essential for comprehending human-centric scenes at a high level. Given an image, HOI detection aims to localize human and object pairs and recognize their interactions, *i.e.* a set of <human, object, action> triplets. Recently, vision Transformers [41], especially the DEtection TRansformer (DETR) [1], have started to revolutionize the HOI detection task. Two-stage methods use an off-the-shelf detector, *e.g.*

DETR, to localize humans and objects concurrently, followed by predicting interaction classes using the localized region features. One-stage methods usually leverage the pre-trained or fine-tuned weights and architecture of DETR to predict HOI triplets from a global image context in an end-to-end manner.

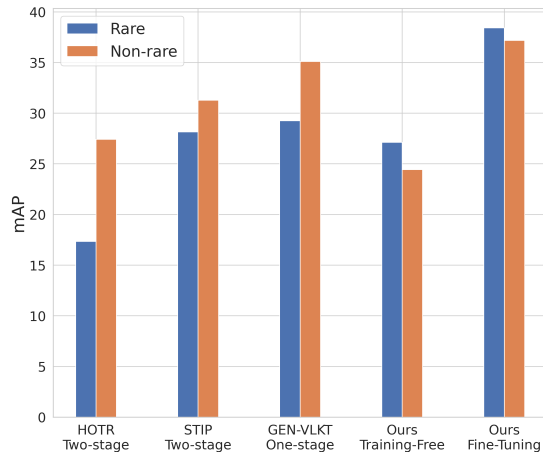
Despite the progress, most previous methods still face two major challenges in solving the HOI detection task. **First**, annotating HOI pairs requires considerable human effort. Therefore, the model may experience data scarcity when encountering a new domain. Even with the availability of ample data, the combinatorial nature of HOIs exacerbates challenging scenarios such as recognizing HOI classes from long-tailed distributions. While people can efficiently learn to recognize seen and even unseen HOI from limited samples, most previous methods [19, 39, 28] suffer a significant performance drop on rare classes as shown in Figure 1(a). **Second**, training or fine-tuning an HOI detector can contribute to high computational cost and time as shown in Figure 1(b)¹. Training two-stage methods involves exhaustively combining instance-level features to predict pairwise relationships. In contrast, training one-stage HOI detectors that adopt the architecture of DETR model can be challenging due to the heavy reliance on transformers [31].

According to the challenges mentioned above, our goal is to *build an efficient adaptive HOI detector resilient to imbalanced data, which can not only adapt to a target dataset without training but also quickly converge when fine-tuning.*

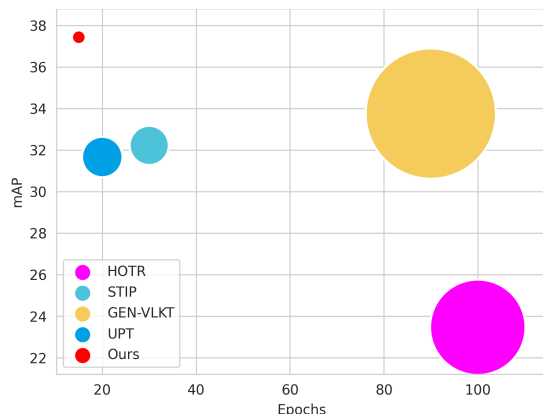
To deal with the problem of lacking labeled data in a target HOI visual domain, we propose a training-free approach with a concept-guided memory module that provides a balanced memory mechanism for all HOI classes. The concept-guided memory module leverages not only the *domain-specific visual knowledge*, but also the *domain-*

¹We exclude the training time for pre-training or fine-tuning the object detector since all methods follow the same protocol to pre-train or fine-tune the object detector.

*Corresponding author



(a) **Performance comparisons on rare and non-rare HOI classes.** ADA-CM performs well on rare and non-rare classes in both the training-free and fine-tuning settings.



(b) **Performance vs Efficiency analysis on HICO-DET dataset.** The size of the blobs is proportional to the models' training gpu time, spanning from 14.85 to 163.7 hours.

Figure 1. **Comparisons on HICO-DET dataset.** Our model, ADA-CM, achieves an mAP of 37.47%, performing well on rare HOI classes as shown in 1(a). Compared to other HOI detectors, our method offers tremendous benefits in terms of accuracy and efficiency at training time.

agnostic semantic knowledge for HOI detection task. Specifically, we extract features from the identified regions of interest to create domain-specific visual knowledge. Inspired by the impressive zero-shot capability of large visual-language models [36], we extract the semantic embeddings of HOI categories and treat the language prior as the domain-agnostic semantic knowledge. This enables the model to leverage linguistic commonsense to capture possible co-occurrences and affordance relations between objects and interactions. To store and retrieve the knowledge, we propose to construct a key-value concept-guided memory module, which can help to mitigate the problem of data scarcity by providing additional prior information and guidance to the model. As shown in Figure 1(a), our approach

can work well in a *training-free* manner and can effectively detect rare HOI classes.

Moreover, to quickly adapt to new domains, we propose to unfreeze the properly initialized concept-guided cache memory and *inject lightweight residual instance-aware adapters* at spatial sensitive feature maps during training. Specifically, unfreezing the cache memory enables the model to select which knowledge is highlighted or suppressed dynamically for the HOI task. Since many valuable cues for the HOI detection task may appear from the early spatial-aware and fine-grained feature maps, we propose to early inject prior knowledge into low-level feature maps to capture the local geometric spatial structures required for pair-wise relationship detection. During fine-tuning, while the instance-aware adapters are tailored to facilitate HOI understanding given instance-level prior knowledge, the explicit concept-guided memory mechanism can help alleviate forgetfulness of rare HOI classes as shown in Figure 1(a). Furthermore, Figure 1(b) demonstrates that our network can achieve the best performance by training for a few epochs with fast convergence speed.

Contributions. Our key idea is to design a HOI detector that leverages knowledge from pre-trained vision-language models and can be adapted to new domains via training-free or fine-tuning. Our work brings three contributions:

- (1) To the best of our knowledge, we are the first to propose a training-free human-object interaction detection approach, by constructing a balanced concept-guided memory that leverages domain-specific visual knowledge and domain-agnostic semantic knowledge.
- (2) We demonstrate that unfreezing the properly initialized cache memory and injecting lightweight residual instance-aware adapters at spatial sensitive feature maps during training further boost the performance.
- (3) Our approach achieves competitive results on VCOCO [10] and state-of-the-art on HICO-DET [2] dataset by training for a few epochs with fast convergence speed.

2. Related Work

2.1. HOI Detection

HOI learning has been rapidly progressing in recent years with the development of large-scale datasets [10, 2, 22, 27] and vision Transformers [41, 1]. HOI detection methods can be categorized into one- or two-stage approaches. One-stage HOI detectors [4, 19, 39, 53, 16, 35, 43, 28] usually formulate HOI detection task as a set prediction problem originating from DETR [1] and perform object detection and interaction prediction in parallel [4, 19, 39, 53] or sequentially [28]. However, they tend to cost a lot of computing resources and converge slowly. Two-stage methods [2, 34, 52, 6, 46, 47] usually utilize pre-trained detectors [37, 11, 1] to detect human and object

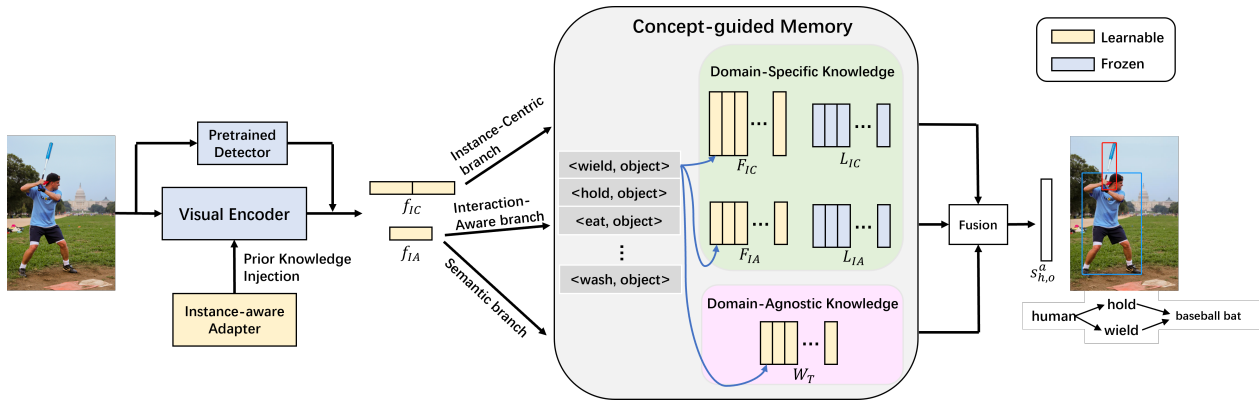


Figure 2. **The overall framework of our ADA-CM**. The proposed method supports two settings: *training-free* and *fine-tuning*. For both, human, object and union features are obtained from training samples through CLIP and converted to instance-centric features f_{IC} and interaction-aware features f_{IA} . The converted features and their corresponding label vectors are used to construct a concept-guided memory to represent the domain-specific visual knowledge. W_T represents the semantic embeddings that are extracted from CLIP’s text encoder with handcrafted prompts of HOI labels, serving as the domain-agnostic semantic knowledge. All knowledge contained in the memory can be combined at inference time in a *training-free* manner. During *fine-tuning*, we unfreeze the properly initialized concept-guided memory and inject lightweight residual instance-aware adapters. The weights of blue blocks are frozen while the yellow blocks’ are learnable. (Best viewed in color.)

proposals and exhaustively enumerate all possible human-object pairs in the first stage. Then they design an independent module to predict the multi-label interactions of each human-object pair in the second stage. Despite their improved performance, both one- and two-stage methods suffer from sheer drop on rare classes, as shown in Figure 1(a), due to the long-tailed distribution of HOI training data.

The most relevant work to ours is GEN-VLKT [28], which proposes a Visual-Linguistic Knowledge Transfer (VLKT) module to transfer knowledge from a visual-linguistic pre-trained model CLIP. Different from previous methods, our ADA-CM utilizes a concept-guided memory module to explicitly memorize balanced prototypical feature representations for each category, which not only helps quickly retrieve knowledge in the training-free mode, but also relieves forgetfulness for rare classes during fine-tuning.

2.2. Adapting Pretrained Models

Vision and language Pretraining is a fundamental research topic in artificial intelligence. It has made rapid progress [24, 36, 23, 21] in recent years thanks to the transformer mechanism [41]. Instead of fine-tuning whole pre-trained models on downstream tasks which always costs a large of computational resources, many works study feature adapters [15, 8, 48, 38, 5] or prompt learning [51, 17] to conduct fine-tuning.

TipAdapter [48] propose to construct an adapter via a key-value cache model, where the idea of our concept-guided memory originates from. However, TipAdapter is designed for object classification and only needs to store

object-level features. In contrast, our approach differs in two key aspects: First, our approach is tailored for the more complex task of HOI understanding, and hence we store more fine-grained prototypical features including both instance-wise and pair-wise features. Second, given the combination of the long-tailed human distribution and long-tailed object distribution, we balance the cached features of different HOI concepts. By doing so, we can ensure that each HOI concept is given appropriate weight and attention, even if some concepts are less frequent in the dataset.

Some works [8, 48] utilize adapter at the last layers for classification task, however, the valuable cues for HOI detection task may appear at the start from the early spatial-aware and fine-grained feature maps, which make any adjustment purely at the tail of the network less effective. Other works [38, 5] utilize adapter at the early layers of the pretrained visual encoders, but they do not focus on the pair-wise spatial relationship and thus are not suitable for the HOI detection task. To address these we first propose to store features representing interaction concepts under training-free mode. Furthermore, we propose to not only unfreeze the global knowledge cache memory in fine-tuning, but also early injects prior knowledge to low-level feature maps to capture the local geometric spatial structures required for pair-wise relationship detection.

3. Method

3.1. Overview

The architecture of our ADA-CM is shown in Figure 2. Our approach is two-stage and consists of two main steps:

1) object detection and 2) interaction prediction. Given an image \mathcal{I} , we first use an off-the-shelf object detector \mathcal{D} , e.g. DETR, and apply appropriate filtering strategies to extract all detections and exhaustively enumerate human-object pairs. Every human-object pair could be represented by a quintuple: $(b_h, s_h, b_o, s_o, c_o)$. Here $b_h, b_o \in \mathbb{R}^4$ denote the detected bounding box of a human and object instance, respectively. $c_o \in \{1, \dots, O\}$ is the object category. s_h and s_o denote the confidence score of the detected human and object, respectively. The second stage includes a multi-branch concept-guided memory module (Figure 2(right)), to recognize the action category $a_{h,o} \in \{1, \dots, A\}$ and produce a confidence score $s_{h,o}^a$ for each human-object pair.

In the training-free setting, ADA-CM is well-suited for situations where there is limited data available. This makes it a valuable tool for addressing the challenges of data scarcity. We construct the multi-branch concept-guided memory, which consists of the instance-centric branch, the interaction-aware branch and the semantic branch, to store domain-specific visual knowledge and domain-agnostic semantic knowledge. Given human-object pairs in a training set, we first extract the fine-grained features and interaction-aware features, for the former two branches. Fine-grained features are used to capture the detailed characteristics of HOIs, such as the pose or orientation state of the detected instances. On the other hand, interaction-aware features capture interaction-relevant environmental and contextual information that can affect the interaction, such as spatial layout and social context. Then for the semantic branch, we extract the semantic embeddings of the HOI categories and incorporate them into HOI cache memory as domain-agnostic knowledge. Semantic features enable the model to leverage linguistic commonsense in order to capture potential co-occurrences and affordance relations between objects and interactions.

In the fine-tuning setting, to further improve the model’s performance, we include an instance-aware adapter, as shown in Figure 3. This adapter injects prior knowledge into the visual encoder to better encode instance-level features through an effective fine-tuning. In addition, we unfreeze the cached keys of the concept-guided memory (yellow modules in Figure 2). The logits of these different types of memories are then linearly combined to output the final score $s_{h,o}^a$. Finally, the HOI score $\hat{s}_{h,o}^a$ for each human-object pair can be written as:

$$\hat{s}_{h,o}^a = (s_h \cdot s_o)^\lambda \cdot \sigma(s_{h,o}^a) \quad (1)$$

where λ is a hyperparameter used at inference time to suppress overconfident objects[46] and σ is a sigmoid function.

3.2. Concept-guided Memory

Most conventional methods directly apply a multi-label classifier fitted from the dataset to recognize the HOIs for

interaction understanding. However, due to the complicated human-centric scenes with various interactive objects, such paradigms suffer from a long-tailed distribution which is the combination of the long-tailed human distribution and the long-tailed object distribution. To alleviate the problem, we introduce a multi-branch concept-guided memory to explicitly store different types of balanced concept features for all HOI classes. The concept-guided memory consists of three memory branches that are leveraged at the interaction prediction stage: 1) the instance-centric branch; 2) the interaction-aware branch; 3) the semantic branch. The three branches store interactive instances’ appearance, interaction-relevant context, and linguistic common sense of each HOI class. These branches complement each other, enhancing the model’s HOI understanding. Among them, the instance-centric branch and the interaction-aware branch store the domain-specific visual knowledge while the semantic branch stores the domain-agnostic semantic knowledge.

Instance-Centric Branch We use a key-value structure (F_{IC}, L_{IC}) for the Instance-Centric Branch as shown in Figure 2. We use f_h, f_o to represent human features and object features, respectively. Given human-object pairs (b_h, b_o) in the training set, we use the CLIP to encode the cropped regions as f_h, f_o for b_h, b_o , respectively. Then the concatenated features $f_{IC} = \text{Concat}(f_h, f_o)$ are stored in F_{IC} as a key and the corresponding labels are transformed as multi-hot vectors and stored in L_{IC} as values.

Interaction-Aware Branch Similar to the instance-centric branch, the interaction-aware branch is also modeled by a concept-feature dictionary (F_{IA}, L_{IA}) . Given human-object pairs (b_h, b_o) , we compute the union regions of the human-object pairs b_u and extract the interaction-aware features f_{IA} from the cropped union regions. Then f_{IA} and its corresponding label vectors are stored in (F_{IA}, L_{IA}) . The interaction-relevant information concentrates more on the background semantics and the spatial configuration of human-object pairs. This design choice further help to make more informed interaction predictions.

Semantic Branch While the prior two branches are rich in domain-specific knowledge, the semantic branch focuses on domain-agnostic knowledge, leveraging linguistic common sense to boost interaction prediction. To construct W_T for the semantic branch, we first use handcrafted prompts (i.e., A photo of a person is <ACTION> an object) to generate the raw text description of interactions. Then we pass such query through the CLIP text encoder and obtain the weights for the semantic classifier. Note that the semantic branch utilizes the alignment of visual and textual feature generated by CLIP, and thus can easily extend to a novel HOI by adding its semantic feature to the classifier.

Given the above branches of concept-guided memory,

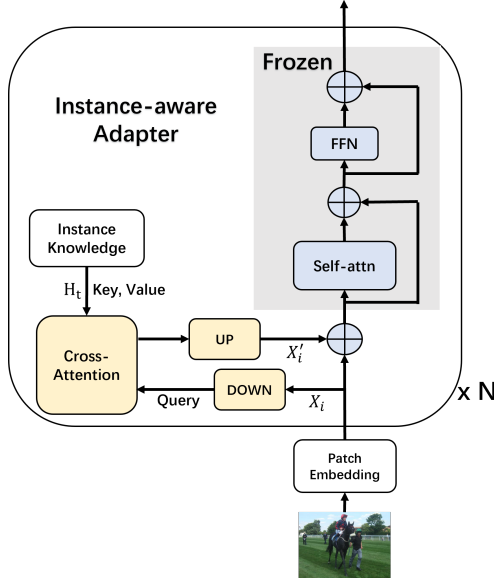


Figure 3. **Instance-aware Adapter.** Architecture of our modified CLIP visual encoder which is equipped with instance-aware adapter modules in each transformer block. The instance-aware adapters aid in the early injection of instance-level prior knowledge into spatial-aware and fine-grained feature maps, thereby assisting the pair-wise relationship detection task.

the interaction scores could be estimated as follows:

$$\begin{aligned}
 s_{vis} &= \gamma_{IC} \cdot (f_{IC} F_{IC}^T) L_{IC} + \gamma_{IA} \cdot (f_{IA} F_{IA}^T) L_{IA} \\
 s_{h,o}^a &= s_{vis} + \gamma_T \cdot f_U W_T^T
 \end{aligned} \quad (2)$$

where γ_{IC} , γ_{IA} and γ_T controls the balanced weights of the three branches, and f_U represents the feature of union regions of human-object pairs, identical to f_{IA} during implementation.

3.3. Instance-aware Adapter

ADA-CM could utilize the knowledge from CLIP and a few-shot training set to achieve competitive performance for HOI detection tasks in the training-free setting. However, the performance improvement upper-bounds as more and more samples are cached as shown in Figure 4. The information stored in memory features is a holistic representation that lacks detailed spatial information, which is critical for HOI understanding. As a result, we design an instance adapter that could be inserted into each block of the CLIP visual encoder as shown in Figure 3. By incorporating the prior knowledge of the feature maps rich in spatial configurations, we equip CLIP with instance-aware knowledge, making it better understand actions to boost the performance of interaction prediction.

Prior Knowledge The prior knowledge consists of three components. (1) Spatial configuration, which captures the

geometric information of objects and enable discrimination between fine-grained categories. (2) Semantic information of extracted objects: we use CLIP language encoder to obtain visual embeddings for given detected objects, which enables us to leverage the language priors to capture which objects can be interacted with. (3) A confidence score that reflects the quality/uncertainty of the candidate instance. The prior knowledge is unique to each image and contains information about all the detected instances in it. The process of the instance-aware adapter, which fuses the CLIP’s visual features with these components, will be introduced below.

Instance-aware Adapter The instance-aware adapters are sub-networks with a small number of parameters and replaced at the beginning of every block in the transformer encoder. With the adapters, ADA-CM could learn prior knowledge from extracted objects which provides better instance representation for downstream tasks with fine-tuning. We use cross-attention module to inject the prior knowledge to CLIP’s visual encoder as shown in Figure 3.

To be more specific, we denote $X_i \in \mathbb{R}^{H'W' \times d}$ as the input feature map of i -th block of the visual encoder, where $H'W'$ is the shape of the feature map. To reduce the computational cost, we first reduce the dimension of features through down-projection. The weight matrices of down-projection and up-projection blocks are denoted as $W_{down} \in \mathbb{R}^{d \times d'}$ and $W_{up} \in \mathbb{R}^{d' \times d}$, where d and d' satisfy $d' \ll d$. Additionally, we append a multi-head cross attention (MHCA) module after the first projection layer to incorporate the outputs of the down-projection layer and the prior knowledge from DETR:

$$\begin{aligned}
 H_t &= MLP_p(P_t) \\
 X'_i &= MHCA((X_i \cdot W_{down}), H_t, H_t) \cdot W_{up}
 \end{aligned} \quad (3)$$

where $P_t = \{p_t\}_{t=1}^{N_t}$ represents the prior knowledge of extracted instances, $p_t = \{b_t, c_t, e_t\}$ consists of box coordinates b_t , the confident score c_t , the object text embedding e_t extracted from CLIP text encoder and N_t denotes the number of extracted instances in a given image. $H_t = \{h_t | h_t \in \mathbb{R}^{d'}\}_{t=1}^{N_t}$ represents hidden features obtained after the prior features pass through down-projection MLPs.

3.4. Training and Inference

Training-Free Setting In the training-free setting, given an image at inference time, for every detected human-object pair, we extract its features f_h, f_o, f_u as presented in Section 3.2. Then the interaction scores $\hat{s}_{h,o}^a$ could be calculated as shown in Equation 1 and 2.

Fine-Tuning Setting During the fine-tuning phase, the blue components shown in Figure 2 are frozen and kept the same as those in the training-free phase while the weights of yellow components are trainable. The process of initializing

Method	Backbone	TP	Full	Rare	Non-rare
One-stage Methods					
GPNN [34]	ResNet-101	-	13.11	9.41	14.23
UnionDet [18]	ResNet-50-FPN	-	17.58	11.72	19.33
IP-Net [44]	Hourglass-104	-	19.56	12.79	21.58
PPDM [27]	Hourglass-104	-	21.94	13.97	24.32
HOI-Trans [53]	ResNet-50	-	23.46	16.91	25.41
HOTR [19]	ResNet-50	9.90	25.10	17.34	27.42
AS-Net [4]	ResNet-50	-	28.87	24.25	30.25
QAHOI [3]	Swin-Base	-	29.47	22.24	31.63
QPIC [39]	ResNet-101	41.46	29.90	23.92	31.69
Iwin [40]	ResNet-50-FPN	-	32.03	27.62	34.14
CDN [45]	ResNet-101	-	32.07	27.19	33.53
GEN-VLKT [28]	ResNet-50+ViT-B	42.05	33.75	29.25	35.10
Two-stage Methods					
InteractNet [9]	ResNet-50-FPN	-	9.94	7.16	10.77
iCAN [7]	ResNet-50	-	14.84	10.45	16.15
TIN [26]	ResNet-50	-	17.03	13.42	18.11
PMFNet* [42]	ResNet-50-FPN	-	17.46	15.65	18.00
DRG [6]	ResNet-50-FPN	-	19.26	17.74	19.71
VCL [12]	ResNet-50	-	19.43	16.55	20.29
FCMNet* [32]	ResNet-50	-	20.41	17.34	21.56
ACP [20]	ResNet-152	-	20.59	15.92	21.98
IDN* [25]	ResNet-50	-	23.36	22.47	23.63
STIP [49]	ResNet-50	-	30.56	28.15	31.28
UPT (TF) [47] †	ResNet-50	0	13.64	9.66	14.85
UPT [47]	ResNet-50	13.24	31.66	25.94	33.36
ADA-CM (TF)	ResNet-50+ViT-B	0	25.19	27.24	24.58
ADA-CM (FT)	ResNet-50+ViT-B	3.12	33.80	31.72	34.42
ADA-CM (FT)	ResNet-50+ViT-L	6.62	38.40	37.52	38.66

Table 1. **State-of-the-art Comparison on HICO-DET.** The table compares the HOI detection performance (mAP×100) on the HICO-DET test set. TP: Number of Trainable Params(M), TF: Training-Free, FT: Fine-Tuning. "*" indicates the method uses additional features from pose estimation of body-parts. † indicates we apply our generic memory design to a representative approach. Our method equipped with fine-tuned adapter achieves a new state-of-the-art.

the cache memory is identical to that used in the training-free mode. We use the original image as the input of visual encoder and adopt ROI-Align[11] to extract f_{IC} , f_{IA} defined in Section 3.2. The whole model is trained on focal loss [29] \mathcal{L} . We denote θ as the learnable parameters of our model. The optimization procedure could be written as follows:

$$\theta^* = \arg \min_{\theta} \mathcal{L}(g(f(\mathcal{D}(\mathcal{I}), \mathcal{I}), M), s_{GT}) \quad (4)$$

where \mathcal{I} , \mathcal{D} represent the input image and pre-trained detector as defined in Section 3.1, f is the feature extraction procedure presented in Section 3.2, M represents the concept-guided memory, g is a similarity function defined in Equation 2, and s_{GT} represents the ground-truth label.

4. Experiments

4.1. Experimental Settings

Datasets: We conducted extensive experiments on both the HICO-DET [2] and V-COCO [10] datasets. HICO-DET consists of 47,776 images (38,118 training images and 9,658 test images). It has 600 HOI categories constructed of 117 action classes and 80 object classes. In addition, we simulate a zero-shot detection setting by holding out 120 rare interactions following previous settings [12, 13, 43]. V-COCO is subset of COCO and has 10,396 images (5,400 trainval images and 4,946 test images). It has 24 different types of actions and 80 types of objects.

Evaluation Metric: Following the standard evaluation, we use mean average precision (mAP) to examine the model performance. For HICO-DET, we report the mAP over three different category sets: all 600 HOI categories (Full),

138 HOI categories with less than 10 training instances (Rare), and the remaining 462 HOI categories (Non-rare). For V-COCO, we report the average precision (AP) under two scenarios, AP_{role}^{S1} and AP_{role}^{S2} , which represent different scoring ways for object occlusion cases.

4.2. Implementation Details

We fine-tune the detector DETR prior to training and then freeze its weights. Specifically, for HICO-DET, we fine-tune DETR on HICO-DET with its weights initialized from the publicly available model pre-trained on MS COCO [30]. For V-COCO, we pre-train DETR from scratch on MS COCO, excluding those images in the test set of V-COCO. We employ two ViT variants as our backbone architectures: ViT-B/16 and ViT-L/14, where "B" and "L" refer to base and large, respectively. The input resolution for ViT-B and ViT-L is 224 pixels and 336 pixels, respectively. γ_{IC} , γ_{IA} and γ_T are set to be 0.5, 0.5 and 1.0, respectively. λ is set to 1 during training and 2.8 during inference [46, 47]. We use AdamW [33] as the optimizer with an initial learning rate of 1e-3 and train ADA-CM for 15 epochs. The model is trained on a single NVIDIA A100 device with an efficient batch size of 8.

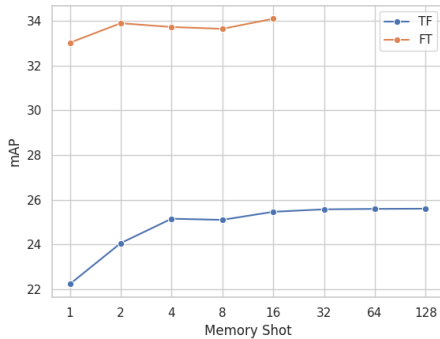


Figure 4. **Memory size ablation.** Memory shot indicates the number of samples in the memory per HOI. TF: Training-Free, FT: Fine-Tuning. We observe two behaviors: (i) FT performs well independently as to the memory size, and (ii) our TF benefits from larger memory sizes but does pretty well with as low as 16-32 samples.

4.3. HOI State-of-the-art Comparison

We compare the performance of our model with existing methods on HICO-DET and V-COCO datasets. For HICO-DET, we evaluate our model on both the default setting and the zero-shot setting as described in Section 4.1. As shown in Figure 1, our model under training-free setting outperforms many models on the HICO-DET dataset, e.g. HOI-Trans, IDN, which require training. Note that we are the first to propose a training-free HOID approach. Ex-

Method	AP_{role}^{S1}	AP_{role}^{S2}
One-stage Methods		
UnionDet [18]	47.5	56.2
IP-Net [44]	51.0	-
HOI-Trans [53]	52.9	-
GG-Net [50]	54.7	-
HOTR [19]	55.2	64.4
AS-Net [4]	53.90	-
QPIC [39]	58.8	61.0
Iwin [40]	60.47	-
CDN† [45]	61.68	63.77
GEN-VLKT† [28]	62.41	64.46
Two-stage Methods		
InteractNet [9]	40.0	-
iCAN [7]	45.3	52.4
PMFNet* [42]	52.0	-
VCL [12]	48.3	-
DRG [6]	51.0	-
FCMNet* [32]	53.1	-
ACP* [20]	53.2	-
IDN*† [25]	53.3	60.3
UPT [47]	59.0	64.5
ADA-CM (TF, ViT-B)	39.09	43.93
ADA-CM (FT, ViT-B)	56.12	61.45
ADA-CM (FT, ViT-L)	58.57	63.97

Table 2. **State-of-the-art comparison on V-COCO.** We report HOI detection performance (mAP×100) on the V-COCO test set. "*" indicates the method uses additional features from pose estimation. † indicates the method uses bounding box annotations of test set to pre-train the object detector. Note that our method achieves competitive performance *wrt* the state-of-the-art. It is also worth highlighting that some of these methods require extensive training and might be over-specialized for this target dataset.

Method	Type	Unseen	Seen	Full
VCL [12]	RF	10.06	24.28	21.43
ATL [13]	RF	9.18	24.67	21.57
FCL [14]	RF	13.16	24.23	22.01
THID [36]	RF	15.53	24.32	22.96
GEN-VLKT [28]	RF	21.36	32.91	30.56
Ours (TF, ViT-B)	RF	26.83	24.54	25.00
Ours (FT, ViT-B)	RF	27.63	34.35	33.01
VCL [12]	NF	16.22	18.52	18.06
ATL [13]	NF	18.25	18.78	18.67
FCL [14]	NF	18.66	19.55	19.37
GEN-VLKT [28]	NF	25.05	23.38	23.71
ADA-CM (TF, ViT-B)	NF	30.11	23.16	24.55
ADA-CM (FT, ViT-B)	NF	32.41	31.13	31.39

Table 3. **Zero-shot comparison on HICO-DET.** This table compares the performance of our model with state-of-the-art methods on the Zero-shot setting of HICO-DET. RF: Rare First. NF: Non-rare First. TF: Training-Free. FT: Fine-Tuning.

IC	IA	S	Full	Rare	Non-rare
✓			22.36	21.38	22.65
	✓		21.74	20.40	22.15
		✓	23.36	27.81	22.03
	✓	✓	25.02	27.28	24.35
✓	✓	✓	25.19	27.24	24.58

Table 4. **Concept-guided Memory ablation (Training-Free).** This table studies the effect of each branch on the training-free setting. IC: Instance-Centric branch, IA: Interaction-Aware branch, S: Semantic branch. Results are on HICO-DET.

Adapter	IC	IA	S	Full	Rare	Non-rare
	✓	✓	✓	27.63	25.40	28.30
✓	✓			32.95	32.32	33.13
✓		✓		31.26	30.77	31.40
✓			✓	30.48	29.12	30.89
✓	✓	✓	✓	33.80	31.72	34.42

Table 5. **Concept-guided Memory ablation (Fine-Tuning).** This table studies the effect of each branch on the fine-tuning setting. IC: Instance-Centric branch, IA: Interaction-Aware branch, S: Semantic branch, Adapter: Instance-aware Adapter. Results are on HICO-DET.

isting one-stage/two-stage methods require training the decoder/classifier respectively. Thus none of them can directly provide results for comparison without training. To further verify our method has advantages against others, we apply our generic memory design to a representative approach (UPT). Empirically, as shown in Table 1, we observe that UPT achieves a mAP of 13.64%, while our method reaches 25.19% on HICO-DET full split, demonstrating the effectiveness of our approach. Notably, transformer-based one-stage HOI detectors require training for instance grouping and pair-wise box prediction, which impedes the acquisition of reasonable training-free results.

Furthermore, our model(ViT-B) after fine-tuning outperforms all other methods, which shows our adapter’s strong capability of modeling pair-wise human-object relationships. Our model(ViT-L) can further leverage high-resolution input images due to its computational efficiency, which also facilitates feature extraction for small objects and boosts its performance. To be specific, our model achieves 4.65% and 6.74% mAP gains compared to the state-of-the-art of one- and two-stage methods, respectively. It’s worth noting that existing models tend to forget rare HOI classes while ours achieve a significant improvement especially on the Rare HOI categories, which confirms the effectiveness of our concept-guided memory for memorizing rare HOI concepts.

We choose ViT as our backbone to leverage the aligned image-text feature space provided by CLIP. In Table 1, we provide a fair comparison with GEN-VLKT, which uses

the same pre-trained weights as our model. Our approach achieves better performance on rare classes, demonstrating the efficiency and effectiveness of our method. It is noteworthy that our model requires less training time($0.1\times$), as illustrated in Figure 1(b). Besides, the "Backbone" column in Table 1 stands for which part of the model is pre-trained. Most methods with the backbone of ResNet still have transformer encoder layers followed, and thus need to optimize a large amount of parameters during training. Though a stronger backbone (ViT-B) with a frozen detector, ADA-CM has a small number of trainable parameters. As the main architecture of our backbone is frozen, it can easily be upgraded to more powerful one without struggling and have better scalability. The last row in Table 1 shows that our model’s performance can boost a lot through an easy switch, where the overall trainable parameters (tailored for this dataset) are still relatively small.

The results presented in Table 3 demonstrate the effectiveness of our proposed method for zero-shot HOI detection on HICO-DET. Our model achieves a relative gain of 29.35% and 29.38% mAP on the two zero-shot settings, respectively, compared to the best performing approach, GEN-VLKT. This improvement shows the great generalizability of our model for detecting HOIs belonging to unseen combinations and the ability to disentangle the concept of action and object. It is also worth noting that under the Non-rare First setting, specifically the systematic-hard split which contains fewer training instances and is thereby more challenging, our model shows great advantages over the seen categories compared to GEN-VLKT. This demonstrates the effectiveness of domain-specific visual knowledge in our memory. Overall, the performance of our model has significantly improved, demonstrating the effectiveness of the aligned image-text feature space.

For V-COCO, as shown in Table 2, our model achieves competitive results compared with all previous methods which also freeze the weights of detector during training. The improvement is not that significant compared to HICO-DET, which might be caused by the insufficient training samples in V-COCO.

4.4. Ablation Study

In this subsection, we conduct a series of experiments to study the effectiveness of our proposed modules. All experiments are conducted on the HICO-DET dataset.

Memory size. Here we study the impact of the size of our cache model. As shown in Figure 4, in the training-free setting, as the memory shot increases from 1 to 16, the performance consistently improves. However, when we further expand the cache model size, the performance achieves an upper-bound around 26%. For the fine-tuning setting, the performance is not sensitive to the memory size as we can see from the orange line in Figure 4, which shows the

DS	DA	Full	Rare	Non-rare
✓		32.92	31.36	33.39
	✓	30.48	29.12	30.89
✓	✓	33.80	31.72	34.42

Table 6. **Ablation on different types of knowledge.** This table studies the effect of domain-agnostic and domain-specific knowledge on the fine-tuning setting. DS: Domain-Specific. DA: Domain-Agnostic. Results are on HICO-DET.

model’s capability of efficiently utilizing the memory.

Network architecture. This ablation studies the effectiveness of different modules of our network. In the training-free setting, as shown in Table 4, by keeping one branch at a time, we could observe that the semantic branch contributes the most to the performance gains. We suppose the reason is that the domain-agnostic linguistic knowledge is well-aligned with the region features thanks to the zero-shot capability of large visual-language model. The Instance-centric branch and the Interaction-Aware branch further boost the performance, which demonstrates the complementarity of domain-specific visual knowledge and domain-specific linguistic language. In the lightweight tuning setting, we show the effectiveness of different components in Table 5. Compared to the performance in training-free setting, simply unfreezing the keys of memory boosts the performance from 25.19% to 27.63%. By further utilizing our designed instance-aware adapter for better pair-wise relationship modeling, we achieve the best performance of mAP 33.80%.

Different types of knowledge. As shown in Table 6, we also group the branches by domain-agnostic and domain-specific knowledge to see the effect of the different types of knowledge on the performance of the method in the Fine-Tuning mode. Experimental results demonstrate that domain-specific knowledge customized for the specific dataset outperforms domain-agnostic knowledge by an obvious margin of 2.24% and 2.5% mAP on rare and non-rare splits, respectively. By further taking advantage of their complementary knowledge, the combination achieves the best overall performance on all the split of HICO-DET.

5. Conclusion

We proposed ADA-CM, an efficient adaptive model for Human-Object Interaction Detection. Our method showed its effectiveness in both training-free and fine-tuning operation modes. In the training-free setting, our ADA-CM achieves a competitive performance by constructing a concept-guided cache memory that leverages domain-specific visual knowledge and domain-agnostic semantic knowledge. Furthermore, in the fine-tuning setting, with the designed instance-aware adapter, our model achieves competitive results with sota on both the zero-shot and default

setting.

6. Acknowledgements

This work was supported by National Natural Science Foundation of China (61925201,62132001), Zhejiang Lab (NO.2022NB0AB05), Adobe and CAAI-Huawei MindSpore Open Fund. We thank MindSpore² for the partial support of this work, which is a new deep learning computing framework.

References

- [1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.
- [2] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 IEEE winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018.
- [3] Junwen Chen and Keiji Yanai. Qahoi: Query-based anchors for human-object interaction detection. *arXiv preprint arXiv:2112.08647*, 2021.
- [4] 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.
- [5] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *arXiv preprint arXiv:2205.13535*, 2022.
- [6] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *European Conference on Computer Vision*, pages 696–712. Springer, 2020.
- [7] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018.
- [8] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.
- [9] 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.
- [10] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.
- [11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [12] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *European Conference on Computer Vision*, pages 584–600. Springer, 2020.

²<https://www.mindspore.cn/>

- [13] 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.
- [14] 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.
- [15] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.
- [16] ASM Iftekhar, Hao Chen, Kaustav Kundu, Xinyu Li, Joseph Tighe, and Davide Modolo. What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5353–5363, 2022.
- [17] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, 2022.
- [18] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *European Conference on Computer Vision*, pages 498–514. Springer, 2020.
- [19] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo 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*, pages 74–83, 2021.
- [20] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *European Conference on Computer Vision*, pages 718–736. Springer, 2020.
- [21] Wonjae Kim, Bokyoung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.
- [22] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.
- [23] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.
- [24] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.
- [25] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. Hoi analysis: Integrating and decomposing human-object interaction. *Advances in Neural Information Processing Systems*, 33:5011–5022, 2020.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [31] Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. *arXiv preprint arXiv:2004.08249*, 2020.
- [32] Yang Liu, Qingchao Chen, and Andrew Zisserman. Amplifying key cues for human-object-interaction detection. In *European Conference on Computer Vision*, pages 248–265. Springer, 2020.
- [33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [34] 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.
- [35] Xian Qu, Changxing Ding, Xingao Li, Xubin Zhong, and Dacheng Tao. Distillation using oracle queries for transformer-based human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19558–19567, 2022.
- [36] 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.
- [37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

- [38] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. VI-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022.
- [39] 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.
- [40] Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. Iwin: Human-object interaction detection via transformer with irregular windows. *arXiv preprint arXiv:2203.10537*, 2022.
- [41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [42] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9469–9478, 2019.
- [43] 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.
- [44] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4116–4125, 2020.
- [45] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. *Advances in Neural Information Processing Systems*, 34:17209–17220, 2021.
- [46] 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.
- [47] 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.
- [48] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.
- [49] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen 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*, pages 19548–19557, 2022.
- [50] 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.
- [51] 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.
- [52] 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.
- [53] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen 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*, pages 11825–11834, 2021.