Disentangling and Reusing Interaction Cues for Zero-Shot HOI Detection

Jie Gu[1,2,*], He-Yang Xu[1,2,*], Hongxiang Gao[1,2,†], and Chengyu Liu[1,2,†]

[1] State Key Laboratory of Digital Medical Engineering, Southeast University, China
[2] School of Instrument Science and Engineering, Southeast University, China
{gu_jie, xuhy, hongxiang_seu, chengyu}@seu.edu.cn

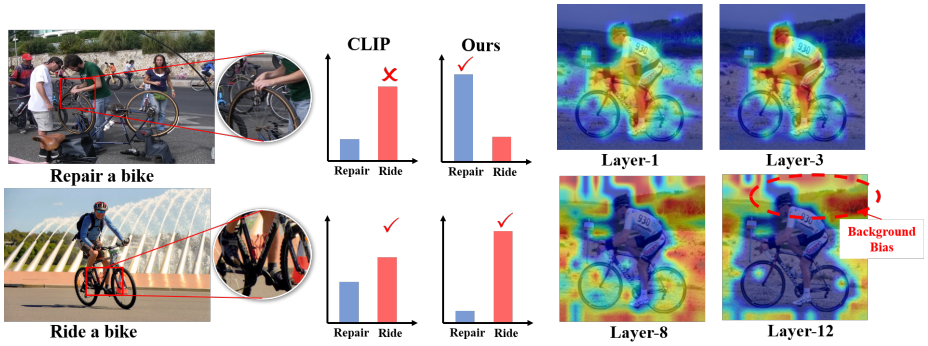
Abstract. Zero-shot Human-Object Interaction (HOI) detection has achieved significant progress by leveraging the open-vocabulary capabilities of Vision-Language Models. However, the generalization of existing methods is severely restricted by a two-fold representational bottleneck. First, at the feature level, essential fine-grained interaction cues are frequently overshadowed by high-level global semantics. Second, at the decision level, models exhibit a strong background shortcut bias, often relying on scene co-occurrences rather than the actual human-object interactions. To address these challenges, we propose a novel zero-shot HOI detection framework, **DRIC**, which is explicitly formulated for **D**isentangling and **R**eusing **I**nteraction **C**ues. Specifically, we introduce the Context-Orthogonal Residual Disentangling (CORD) module, which projects features onto a global prototype subspace to decouple fine-grained residual details from global semantic biases, subsequently reintegrating them for enhanced local discriminability. Furthermore, the proposed Stratified Local Fusion (SLF) mechanism mitigates background shortcuts by aggregating multi-depth features from the image encoder of CLIP. This recovers essential contact-level details, guiding the model toward actual interactions. Extensive experiments demonstrate that DRIC consistently outperforms state-of-the-art methods across zero-shot evaluation settings, validating the robustness of the proposed framework in handling interaction ambiguity and cross-scene transfer.

Keywords: Human-Object Interaction · Zero-Shot Learning · Vision-Language Models

1 Introduction

The primary objective of Human-Object Interaction (HOI) detection is to localize human-object pairs within an image and identify their interactions, formally defined as a triplet ⟨human, object, action⟩. Serving as a cornerstone for embodied intelligence perception, HOI detection holds significant value in applications such as Augmented/Virtual Reality (AR/VR) [16,18,43], and robotic

*Equal contribution. †Corresponding authors.



(a) Interaction Ambiguity in Similar Contexts

(b) Visual Saliency Shift in CLIP Encoder Layers

Fig. 1: Challenges in Zero-Shot HOI Detection. (a) In semantically similar contexts (e.g., riding vs. repairing), the model struggles to distinguish interactions due to the lack of fine-grained local evidence. (b) As the depth of the CLIP encoder layers increases, visual saliency shifts from local details to background shortcuts.

assistance [44, 53]. Although significant advances have been made recently, conventional HOI detection methods [4, 8, 27, 35, 38] predominantly rely on fully supervised paradigms within closed-set scenarios. These approaches depend heavily on extensive annotations of predefined interaction categories, making it difficult for them to generalize to unseen verbs or novel human-object combinations.

Recently, Vision-Language Models (VLMs), such as CLIP [36], have demonstrated exceptional zero-shot transfer capabilities through large-scale image-text contrastive learning. This alignment paradigm facilitates zero-shot HOI detection by formulating interaction categories as textual prompts and exploiting the joint embedding space of CLIP for cross-category transfer. Although existing methods for zero-shot HOI detection [22, 28, 42, 51] have achieved substantial improvements, we observe that the generalization of these approaches is severely restricted by a two-fold representational bottleneck, which undermines reliability and transferability in zero-shot settings.

(1) Fine-grained interaction cues are overshadowed by global semantics. Pre-trained on massive image-text pairs, the global representations of foundation models inherently attenuate local interaction cues. Consequently, high-level features struggle to preserve the fine-grained details (*e.g.*, contact regions and poses) necessary for distinguishing similar interactions. As illustrated in Fig. 1a, actions like "ride-bike" and "repair-bike" exhibit high global similarity; their critical distinction relies entirely on local details that existing models fail to exploit, rendering accurate discrimination highly challenging.

(2) Background shortcut learning reinforces spurious correlations over interaction-centric evidence. Models tend to utilize background or scene cues for interaction discrimination rather than relying on actual human-object contact and action details. As Fig. 1b illustrates, although intermediate layers retain some interaction focus, the final deep representations are highly

susceptible to background domination. Consequently, models establish spurious correlations (*e.g.*, predicting "ride" based solely on an outdoor setting) rather than learning generalized interaction patterns. This shortcut bias severely degrades cross-scene generalization and critically hinders transferability to unseen categories in zero-shot settings.

To address these limitations, we propose DRIC, a novel framework for zero-shot Human-Object Interaction detection designed to construct an interaction-oriented structured local evidence pathway without compromising the original global semantic alignment of CLIP. Specifically, we introduce two complementary mechanisms. First, the Context-Orthogonal Residual Disentangling (CORD) module structurally decouples global semantics, explicitly separating fine-grained residual details related to interactions from global biases. These cues are subsequently reintegrated into the representations of CLIP via common-ground modulated residuals, thereby amplifying local discriminability. Second, the Stratified Local Fusion (SLF) mechanism aggregates multi-scale features of human-object pairs across varying depths of the image encoder of CLIP, recovering critical contact-level details discarded in the final layer to mitigate background shortcuts. Furthermore, because the efficacy of hierarchical fusion inherently depends on the availability of high-quality local cues, we introduce a progressive two-stage training paradigm. This strategy initially establishes a robust foundation of fine-grained features via the disentangling module, and subsequently guides the fusion mechanism to optimally exploit these enriched details. Extensive experiments demonstrate that DRIC establishes a new state-of-the-art across zero-shot evaluation settings, yielding robust cross-scene generalization and highly interpretable, interaction-centric visual evidence.

In summary, our main contributions are threefold:

- **Explicit Global-Local Disentanglement.** We propose the CORD module to explicitly decouple fine-grained interaction cues from global semantic biases, fundamentally enhancing the local perception and zero-shot transfer capabilities of CLIP.
- **Mitigation of Background Shortcuts.** We introduce the SLF mechanism to aggregate multi-scale spatial features, effectively suppressing the inherent background shortcut bias in Vision-Language Models to leverage actual interaction details.
- **State-of-the-Art Performance.** Extensive experiments demonstrate that DRIC establishes a new state-of-the-art on the HICO-DET benchmark, yielding superior generalization across all zero-shot evaluation settings.

2 Related Work

2.1 Human-Object Interaction (HOI) Detection

Conventional Human-Object Interaction (HOI) detection falls into one-stage and two-stage paradigms. One-stage approaches [10, 20, 31, 40, 45, 58] employ end-to-end frameworks to directly predict HOI triplets, concurrently generating bounding boxes, object categories, and interaction labels. Transformers have advanced

HOI detection [5–7, 12, 19, 20, 28, 46], especially DETR-inspired models [2] using queries to directly regress HOI triplets [4, 21, 41, 50]. However, relying on fully annotated data limits their generalization to unseen interactions [47, 48]. Conversely, two-stage approaches [23–25,33,56,57] use pre-trained detectors [2,11,37] to independently localize humans and objects, subsequently constructing candidate pairs to classify their interactions. To extract discriminative features, these methods incorporate auxiliary priors like spatial geometries [59], structural priors [14, 27], and human poses [26]. Furthermore, they often employ graph-based relational reasoning to model high-order dependencies [17, 35, 55].

2.2 Zero-Shot HOI Detection

Zero-shot HOI detection aims to recognize interaction relationships absent from the training set. Early approaches [13–15, 32] largely adopt compositional learning strategies, decomposing HOI representations into separate human-object pairs and verbs. To achieve generalization to novel combinations, these models recombine components utilizing mechanisms like graph propagation [32], representation recombination [13, 15], and external dataset mining [14]. However, relying on training semantics limits their capacity to address interactions involving unseen verbs. Vision-Language Models (VLMs) offer an effective solution for unseen category generalization by introducing open-vocabulary knowledge. To enhance transferability, methods generally integrate language priors through two main strategies: distilling CLIP distributions to align visual features with frozen text embeddings [1, 33, 34, 49], or utilizing CLIP representations and text prompts to initialize classifiers and guide representation learning [28,34,54]. Nevertheless, distillation primarily occurs on seen classes, introducing bias. Furthermore, CLIP's object-recognition focus creates a strong global bias that overlooks fine-grained details, such as clothing textures and local contact regions. Consequently, HOIs are often clustered by objects rather than actions, resulting in insufficient discrimination between distinct verbs for identical human-object pairs. Conversely, integrating CLIP into two-stage frameworks [33] facilitates more stable transfer and mitigates end-to-end decoder overfitting by decoupling object detection and interaction classification. Building on this, subsequent methods enhance CLIP's locality awareness by aggregating adjacent image patches and spatial priors [22], and optimize action discrimination by injecting grouped visual variance and Gaussian perturbations into context embeddings [52].

3 Method

3.1 Disentangling and Reusing Interaction Cues

Given an input image I, the primary objective of zero-shot HOI detection is to simultaneously localize all human-object pairs and predict their interaction categories, including those unseen during the training phase. An HOI instance is formally defined as a quadruple $(\mathbf{b}_h, \mathbf{b}_o, v, o)$, where $\mathbf{b}_h, \mathbf{b}_o \in \mathbb{R}^4$ denote the

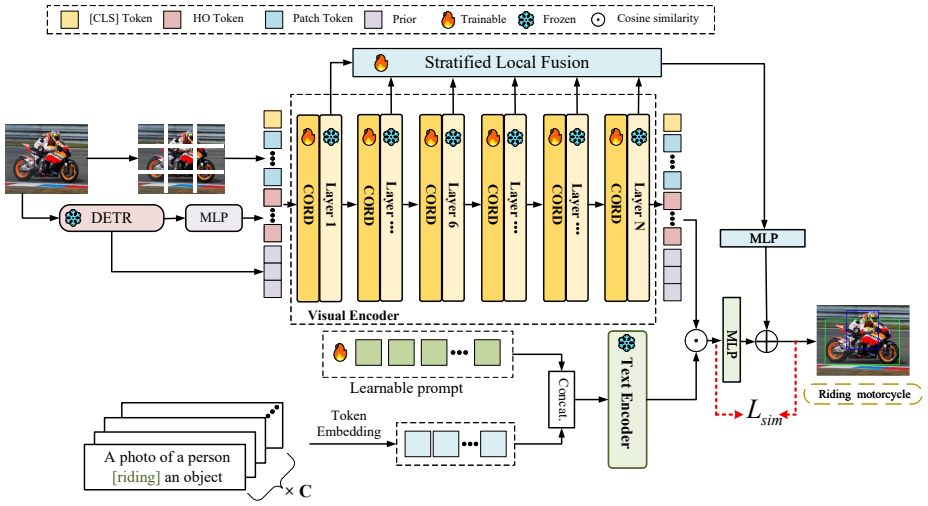


Fig. 2: Overall Framework of DRIC. We adopt a two-stage paradigm. In the first stage, humans and objects are detected. In the second stage, we extract fine-grained interaction cues via the Context-Orthogonal Residual Disentangling (CORD) module and fuse multi-scale features through Stratified Local Fusion (SLF) to suppress background shortcut learning for robust zero-shot interaction classification.

bounding boxes for the human and the object, respectively; $v \in \mathcal{V} = \{1, \ldots, |\mathcal{V}|\}$ represents the verb (or interaction) category; and $o \in \mathcal{O} = \{1, \ldots, |\mathcal{O}|\}$ signifies the object category. In the standard zero-shot setting, annotations are provided exclusively for *seen* interaction categories during training, while textual descriptions are available for both *seen* and *unseen* categories. Consequently, the model is required to generalize the learned interaction knowledge to unseen categories without relying on spurious correlations or dataset co-occurrence biases.

The overall framework of DRIC is illustrated in Fig. 2. We adhere to the established two-stage paradigm for HOI detection: the first stage performs the detection of humans and objects, while the second stage conducts the classification of interactions. Specifically, we employ a pre-trained object detector, DETR [2], to identify objects within the image I, applying standard filtering strategies to obtain a set of detections $\{(\mathbf{b}_i, c_i, s_i, \mathbf{g}_i)\}_{i=1}^{N}$, where $\mathbf{b}_i \in \mathbb{R}^4$, $c_i \in \mathcal{O}$, $s_i \in \mathbb{R}^1$, and $\mathbf{g}_i \in \mathbb{R}^{D_{\text{det}}}$ represent the bounding box, object category, confidence score, and feature vector of the i-th instance, respectively. N and D_{det} denote the number of detected objects and the dimension of the object feature, respectively. The features of these pairs are subsequently fed into a Multilayer Perceptron (MLP) to generate human-object (HO) tokens $\mathbf{T} \in \mathbb{R}^{N_{\text{pair}} \times D}$, where N_{pair} and D denote the total number of valid candidate human-object pairs and the feature dimension of the CLIP model, respectively. These tokens serve to predict the interaction verb $v \in \mathcal{V}$ for each pair. Concurrently, we construct a prior em-

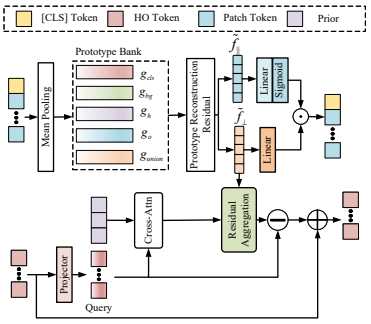


Fig. 3: Framework of CORD. CORD projects features onto a global prototype subspace to disentangle interaction-specific residuals.

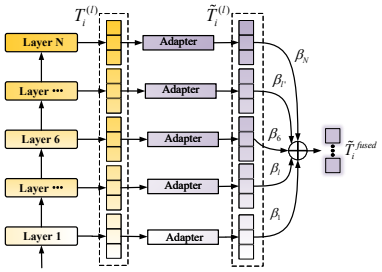


Fig. 4: Framework of SLF. SLF aggregates features from multiple layers to recover fine-grained details lost in deep layers.

bedding $\mathbf{p} = \text{priors\_proj}([\mathbf{b}_i; c_i; s_i])$ utilizing the outputs of DETR to provide task-adaptive guidance.

Subsequently, the HO tokens, in conjunction with the `[CLS]` token and image patches, are input into the CLIP image encoder. Prior to each layer of the CLIP image encoder, we integrate the proposed Context-Orthogonal Residual Disentangling (CORD) module. This integration ensures that fine-grained interaction cues are explicitly decoupled from global biases and effectively reintegrated into the visual representations. Furthermore, we design a Stratified Local Fusion (SLF) module to aggregate the features of human-object pairs from all hierarchical layers of the CLIP image encoder, thereby enhancing the capacity of the model to leverage interaction cues while mitigating background shortcuts. Regarding the textual component, we concatenate learnable prompt tokens with the handcrafted text template "a photo of a person [verb-ing] an object" and feed them into the CLIP text encoder to obtain the text features $\mathbf{T}_{txt} \in \mathbb{R}^{N_{|\mathcal{V}|} \times D}$. The HOI score is computed via the cosine similarity between the HO tokens and the text features: $s^{\text{hoi}} = \cos(\mathbf{T}, \mathbf{T}_{txt})$.

3.2 Context-Orthogonal Residual Disentangling

Representations derived from the image encoder of CLIP tend to exhibit a global bias. This bias often overshadows interaction-specific distinctions and makes it difficult to discern subtle nuances between different verbs associated with the same human-object pair.

To address this limitation, we propose the Context-Orthogonal Residual Disentangling (CORD) module. As depicted in Fig. 3, we initially construct a global prototype bank to characterize key semantic components within the image: $\mathbf{G} \in \mathbb{R}^{K \times d}, \quad d \ll D$, where K denotes the number of prototypes and d represents the feature dimension of the prototypes. We derive five global prototypes by mapping the `[CLS]` token and performing average pooling over four

distinct regions: the background, the human region, the object region, and the union region. Concatenating these elements yields the global prototype bank: $\mathbf{G} = [\mathbf{g}_{\text{cls}}; \mathbf{g}_{\text{bg}}; \mathbf{g}_h; \mathbf{g}_o; \mathbf{g}_{\text{union}}]$.

We further decompose the image representation into a *global common component* and an *interaction-related residual component*. First, we project $\mathbf{F}$ into a low-dimensional space consistent with the prototypes: $\tilde{\mathbf{F}} = \phi(\mathbf{F}) \in \mathbb{R}^{HW \times d}$. Next, we define an orthogonal projection operator on the subspace spanned by the global prototypes:

$$\mathbf{P} = \mathbf{G}(\mathbf{G}^\top \mathbf{G} + \varepsilon \mathbf{I})^{-1} \mathbf{G}^\top, \tag{1}$$

where ε is a term introduced for numerical stability. For any given patch feature $\tilde{\mathbf{f}} \in \mathbb{R}^d$, its projection onto the prototype subspace and the corresponding orthogonal residual are formulated as:

$$\tilde{\mathbf{f}}_{\parallel} = \mathbf{P}\tilde{\mathbf{f}}, \qquad \tilde{\mathbf{f}}_{\perp} = \tilde{\mathbf{f}} - \tilde{\mathbf{f}}_{\parallel}, \tag{2}$$

where $\tilde{\mathbf{f}}_{\parallel}$ represents the global common component, which captures the shared semantics at the scene layout and category levels encoded by the global prototypes. Conversely, $\tilde{\mathbf{f}}_{\perp}$ serves as the residual component orthogonal to the global semantics, which is more likely to preserve local discriminative cues related to human-object interactions. Through explicit subspace constraints, this orthogonal decomposition compels the model to decouple the local variances essential for the accurate discrimination of interactions from strong global context biases.

Following the orthogonal decomposition, we utilize the global component to adaptively modulate the local representation, selectively enhancing local interaction differences while preserving global semantic stability. Specifically, we first generate gating coefficients based on the global common component: $\mathbf{m} = \sigma(\mathbf{W}_m \tilde{\mathbf{f}}_{\parallel})$, where $\sigma(\cdot)$ denotes the Sigmoid activation function and $\mathbf{W}_m$ represents learnable parameters. Subsequently, under the gating constraint, we reintegrate the orthogonal residual into the patch representation:

$$\tilde{\mathbf{f}}' = \tilde{\mathbf{f}} + \alpha(\mathbf{m} \odot \mathbf{W}_{\perp} \tilde{\mathbf{f}}_{\perp}), \tag{3}$$

where $\odot$ denotes element-wise multiplication and α is a learnable residual scaling coefficient. Finally, we map the modulated features back to the original dimension and add them to the original patch tokens via a residual connection:

$$\mathbf{F}' = \mathbf{F} + \mathbf{W}_o \tilde{\mathbf{F}}'. \tag{4}$$

Furthermore, we facilitate the sparse routing of local interaction evidence to the corresponding HO tokens to enable interaction reasoning conditioned on the human-object pair. For each human-object pair, we utilize the corresponding HO token $\mathbf{T}_i$ as a query vector and project it into the attention query space via a linear mapping: $\mathbf{Q}_i = \mathbf{W}_q \mathbf{T}_i$. First, we enable the HO token to perform cross-attention updates on the prior feature $\mathbf{p}$:

$$\mathbf{T}_i^{(1)} = \text{CrossAttn}(\mathbf{Q}_i, \mathbf{p}). \tag{5}$$

Subsequently, using the HO token as a query, we perform cross-attention aggregation on the orthogonal residual patch features $\tilde{\mathbf{F}}_\perp$:

$$\mathbf{T}_i^{(2)} = \text{CrossAttn}(\mathbf{T}_i^{(1)}, \tilde{\mathbf{F}}_\perp). \tag{6}$$

By exclusively utilizing the orthogonal residual component as memory, we guide the HO token to focus on local differences related to interactions. Finally, we update the HO token in a residual manner:

$$\mathbf{T}'_i = \mathbf{T}_i + \gamma \mathbf{W}_t \text{Norm}(\mathbf{T}_i^{(2)} - \mathbf{Q}_i), \tag{7}$$

where γ is a learnable scaling coefficient and $\text{Norm}(\cdot)$ denotes normalization. The updated HO tokens and patch tokens are then fed back into the CLIP image encoder:

$$[\mathbf{T}^{(l)}, \mathbf{cls}^{(l)}, \mathbf{F}^{(l)}] = \mathcal{V}^{(l)}([\mathbf{T}'^{(l-1)}; \mathbf{cls}^{(l-1)}; \mathbf{F}'^{(l-1)}]), \tag{8}$$

where $\mathcal{V}^{(l)}(\cdot)$ represents the l-th Transformer block of the CLIP image encoder.

3.3 Stratified Local Fusion

As identified in our analysis, current zero-shot HOI methods predominantly rely on the final layer of the CLIP image encoder, which is highly susceptible to background shortcut learning. As demonstrated in Fig. 1b, while the final layer tends to drift toward scene-level heuristics, essential interaction-centric details are often preserved in the early and intermediate layers.

To mitigate this representational bias and fully exploit the fine-grained cues decoupled by CORD, we propose the **Stratified Local Fusion (SLF)** mechanism. This component is specifically designed to suppress background shortcuts by aggregating multi-depth features, thereby recovering the contact-level evidence discarded in the deepest layers. As illustrated in Fig. 4, SLF comprises multiple side-path adapter branches and a learnable hierarchical weighted fusion module to integrate global context with localized discriminative cues.

Specifically, we extract HO token representations corresponding to the same human-object pair across all hierarchical layers of the CLIP image encoder. Let $\mathbf{T}_i^{(l)}$ denote the HO token at the l-th layer, where $\mathcal{L}$ represents the set of selected layer indices and $l \in \mathcal{L}$. Given the significant divergence in feature distributions across these layers, we introduce a lightweight adapter for each layer to perform intra-layer alignment:

$$\tilde{\mathbf{T}}_i^{(l)} = \text{Adapter}^{(l)}(\mathbf{T}_i^{(l)}), \tag{9}$$

where $\text{Adapter}^{(l)}(\cdot)$ is sequentially composed of a linear layer, an activation function, and another linear layer.

Subsequently, we employ a learnable hierarchical weighted fusion strategy to aggregate the HO tokens across the selected layers:

$$\tilde{\mathbf{T}}_i^{\text{fused}} = \sum_{l \in \mathcal{L}} \beta_l \tilde{\mathbf{T}}_i^{(l)}, \tag{10}$$

where $\{\beta_l\}$ denotes the set of learnable fusion weights. To preserve the established vision-language alignment of the pre-trained model, we initialize the weight corresponding to the final layer with a larger value.

We then pass the fused HO token $\tilde{\mathbf{T}}^{\text{fused}}$ through a lightweight MLP to obtain a fused score s^{fused}, which is subsequently combined with the original zero-shot score s^{hoi}:

$$s = s^{\text{hoi}} + s^{\text{fused}}. \tag{11}$$

To prevent the fused information from deviating excessively from the global similarity constraints, we incorporate a similarity consistency loss utilizing L_1 regularization:

$$\mathcal{L}_{\text{sim}}(s^{\text{hoi}}, s^{\text{fused}}) = \|s^{\text{hoi}} - s^{\text{fused}}\|_1. \tag{12}$$

3.4 Training and Inference

Two-Stage Training Strategy. Since the efficacy of hierarchical fusion inherently depends on the availability of high-quality local cues, we adopt a deliberate two-stage training paradigm.

Stage 1: CORD Optimization. In the first stage, we exclusively activate the CORD module while keeping the SLF mechanism deactivated to prioritize the construction of a reliable feature foundation. The network is optimized using the Focal Loss [29] applied solely to the zero-shot HOI score s^{hoi}:

$$\mathcal{L}_{\text{stage1}} = \text{FocalBCE}(s^{\text{hoi}}, y), \tag{13}$$

where y denotes the binary target label indicating the presence of a specific interaction.

Stage 2: SLF Integration. Once the CORD module has learned to robustly extract fine-grained residual features, we freeze its parameters. Subsequently, we activate and train the SLF module to aggregate features across varying depths. In this stage, the overall objective function incorporates the fused score s and the similarity regularization term:

$$\mathcal{L}_{\text{stage2}} = \text{FocalBCE}(s, y) + \lambda_{\text{sim}}\mathcal{L}_{\text{sim}}, \tag{14}$$

where we set the hyperparameter λ_{sim} to 1×10^{-4}. This decoupled training strategy prevents the gradients of the hierarchical fusion from interfering with the orthogonal disentanglement, thereby ensuring optimal convergence.

Inference. During the inference phase, the final confidence score s^{final} is computed by integrating the predicted interaction score with the bounding box detection priors, following standard practices [56]:

$$s^{\text{final}} = s \cdot s_h^{\lambda} \cdot s_o^{\lambda}, \tag{15}$$

where λ acts as a weighting hyperparameter to balance the contributions, while s_h and s_o represent the confidence scores generated by the object detector for the human and the object, respectively.

4 Experiments

4.1 Experimental Settings

Dataset. We conduct experiments on the two public benchmark datasets: HICO-DET [3] and V-COCO [9]. HICO-DET comprises a total of 47,776 images, with 38,118 allocated to the training set and 9,658 to the test set. The dataset encompasses 600 HOI categories, constructed from 80 object categories and 117 verb categories. V-COCO is a subset of the MS-COCO [30] dataset. It consists of 5,400 and 4,946 images for training and testing. V-COCO consists of 80 object classes and 29 action classes.

Zero-Shot Settings. We evaluate our method under five zero-shot settings: 1) Unseen Composition (UC): Withholds specific HOI triplets while keeping all objects and verbs in training. 2) RF-UC and 3) NF-UC: Select rare and non-rare HOI categories as unseen classes, respectively. 4) Unseen Verb (UV) and 5) Unseen Object (UO): Exclude specific verbs or objects during training, making all their associated triplets unseen.

Evaluation Metrics. Following standard evaluation protocols, we employ the mean Average Precision (mAP) as our primary metric. Specifically, a detected human-object pair is considered a true positive if it satisfies the following criteria: 1) the Intersection over Union (IoU) between the predicted bounding boxes for both the human and the object and the ground truth boxes exceeds 0.5; and 2) the predicted interaction category is correct. To provide a more balanced assessment, we also report the harmonic mean (HM) metric, which measures performance across both seen and unseen HOI classes.

Implementation Details. Following the standard training setup utilized in previous two-stage zero-shot HOI detection methods, such as LAIN [22], we fine-tune DETR on the HICO-DET dataset prior to training DRIC. We set the feature dimension d to 128 and initialize both α and γ to 1×10^{-3}. We employ ViT-B/16 as our backbone architecture. Further details are provided in the Appendix.

4.2 Comparison with State-of-the-Art Methods

We conduct a comprehensive comparative evaluation of DRIC against existing state-of-the-art methods on the HICO-DET benchmark. This evaluation encompasses five distinct zero-shot settings: UC, RF-UC, NF-UC, UO, and UV.

As shown in Tabs. 1 and 3, DRIC demonstrates exceptional zero-shot transfer capabilities, significantly outperforming existing state-of-the-art methods across all evaluation settings. Compared to the strong baseline LAIN [22], DRIC achieves substantial improvements on unseen categories. Specifically, it yields absolute mAP gains of 2.18%, 0.94%, 0.63%, 0.53%, and 0.36% under the NF-UC, UC, UO, UV, and RF-UC settings, respectively. Furthermore, the Harmonic Mean

Table 1: Comparative performance analysis across zero-shot settings. We report the Harmonic Mean (HM), alongside the mAP for Unseen, Seen, and Full categories under four distinct settings (RF-UC, NF-UC, UV, and UO). For fair comparison, all methods adopt a unified backbone composed of a ResNet-50 object detector and a ViT-B/16 CLIP visual encoder. The best results are highlighted in **bold**.

	RF-UC (Rare First)				**NF-UC (Non-Rare First)**			
Method	HM	Unseen	Seen	Full	HM	Unseen	Seen	Full
CMMP [25]	31.07	29.45	32.87	32.18	30.85	32.09	29.71	30.18
EZ-HOI [23]	31.37	29.02	34.15	33.13	32.03	33.66	30.55	31.17
LAIN [22]	33.36	31.86	35.06	34.41	34.31	36.41	32.44	33.23
VDRP [52]	32.77	31.29	34.41	33.78	33.85	36.45	31.60	32.57
DRIC (Ours)	**33.74**	**32.22**	**35.41**	**34.77**	**35.28**	**38.59**	**32.49**	**33.71**
	UV (Unseen Verb)				**UO (Unseen Object)**			
Method	HM	Unseen	Seen	Full	HM	Unseen	Seen	Full
CMMP [25]	29.13	26.23	32.75	31.84	32.40	33.76	31.15	31.59
EZ-HOI [23]	28.69	25.10	33.49	32.32	32.66	33.28	32.06	32.27
LAIN [22]	31.19	28.96	33.80	33.12	35.58	37.88	33.55	34.27
VDRP [52]	29.80	26.69	33.72	32.73	34.41	36.13	32.84	33.39
DRIC (Ours)	**31.87**	**29.49**	**34.67**	**33.95**	**36.04**	**38.51**	**33.86**	**34.63**

(HM), which measures the overall performance balance, consistently improves, with notable increases of 0.97%, 0.68%, and 0.46% under the NF-UC, UV, and UO settings.

Notably, previous methods [22,23,25,52] are highly susceptible to background shortcuts and global semantic biases, suffering severe performance degradation on unseen interactions. Benefiting from the fine-grained residual disentanglement of the CORD module and the spatial detail aggregation of the SLF mechanism, DRIC effectively overcomes this bottleneck. It maintains high discriminative power even for completely unseen actions (UV) or rare combinations (RF-UC).

Beyond zero-shot scenarios, DRIC inherently enhances fully supervised detection. As detailed in Tab. 2, our method achieves a leading Full mAP of 36.30 and Non-rare mAP of 36.56 on HICO-DET, alongside a peak AP^{S2}_{role} of 65.7 on V-COCO. While LAIN marginally leads the Rare category, these comprehensive results demonstrate that explicitly disentangling interaction cues universally strengthens visual representations and interaction discrimination, even given complete annotations.

4.3 Ablation Study

We perform ablation studies under the **HICO-DET UV (Unseen Verb)** setting to evaluate the zero-shot generalization capability of our method when encountering entirely novel actions.

Table 2: Performance under fully supervised settings. Evaluation is conducted on the HICO-DET and V-COCO datasets. For V-COCO, we report AP_{role} under Scenario 2. The best results are highlighted in **bold**.

Method	HICO-DET			V-COCO
	Full	Rare	Non-rare	AP^{S2}_{role}
CMMP [25]	32.26	33.53	33.24	61.2
LAIN [22]	36.02	**35.70**	36.11	65.1
DRIC (Ours)	**36.30**	35.44	**36.56**	**65.7**

Table 3: Performance under the UC setting. Comparison with state-of-the-art methods on Unseen Composition.

Method	**Unseen**	**Seen**	**Full**	**HM**
CMMP [25]	29.60	32.39	31.84	30.93
LAIN [22]	31.64	35.04	34.36	33.25
DRIC	**32.58**	**35.73**	**34.70**	**34.08**

Table 4: Effectiveness of components. Ablation studies evaluated under the **UV** setting.

CORD	**SLF**	**Unseen**	**Seen**	**Full**
-	-	24.88	31.06	30.19
✓	-	28.58	33.95	33.20
-	✓	27.19	32.87	32.07
✓	✓	**29.49**	**34.67**	**33.95**

Effectiveness of Components. As shown in Tab. 4, the baseline model yields limited performance on unseen verbs (an mAP of 24.88). Incorporating CORD significantly improves the mAP on unseen categories to 28.58, validating its capability to explicitly decouple fine-grained interaction cues from global semantic biases. Adding SLF separately also mitigates background shortcuts, raising the mAP on unseen categories to 27.19. Notably, the full DRIC framework (CORD + SLF) achieves the best results across all metrics. This demonstrates that the residual disentanglement of CORD and the multi-scale evidence recovery of SLF are highly complementary.

Analysis of CORD Mechanism. We evaluate different projection strategies in the CORD module. As shown in Tab. 5, heuristic operations like feature concatenation or subtraction yield sub-optimal unseen mAPs of 28.41 and 28.05, respectively, under the UV setting. Even dynamic cross-attention falls short. In contrast, our orthogonal projection achieves the highest unseen mAP of 29.49. This confirms that geometrically projecting features onto the null space of global prototypes is essential to explicitly decouple fine-grained residual details from global semantic biases, effectively overcoming the feature-level representational bottleneck.

Impact of SLF Strategy. Table 6 shows the impact of fusing features from different layers. Relying solely on the final layer (L12) yields a limited unseen mAP of 28.57, validating our observation that deep representations are highly susceptible to background domination. By aggregating features across all layers (L1-L12), the unseen mAP peaks at 29.49. This progression demonstrates that stratified fusion successfully recovers critical contact-level details preserved in

Table 5: CORD projection strategy. Comparing the orthogonal strategy with heuristic strategies.

Strategy	Unseen	Seen	Full
Concat.	28.41	33.46	32.75
Subtract.	28.05	32.91	32.23
CrossAttn.	28.86	34.09	33.36
Orthogonal	**29.49**	**34.67**	**33.95**

Table 6: SLF layer strategy. Evaluated under the **UV** setting. L: Layer index of ViT-B/16.

Fused Layers	Unseen	Seen	Full
L12	28.58	33.95	33.20
L1-L6 + L12	28.85	34.16	33.42
L6-L12	29.03	34.31	33.57
L1-L12	**29.49**	**34.67**	**33.95**

Table 7: Ablation on training strategies. Evaluated under the UV setting on HICO-DET.

Strategy	Unseen	Seen	Full
One-Stage Joint	28.78	33.9	33.18
two-stage	**29.49**	**34.67**	**33.95**

Table 8: Robustness to background perturbations. Measured in Full mAP. $\Delta\downarrow$ denotes the drop rate.

Model	Original	Perturbed	$\Delta\downarrow$
Baseline	30.19	26.49	12.26%
LAIN [22]	33.12	31.94	3.56%
DRIC (Ours)	**33.95**	**33.21**	**2.18%**

shallow and intermediate layers, guiding the focus of the model toward actual interactions rather than spurious scene co-occurrences.

Necessity of the Two-Stage Training Paradigm. As stated in the introduction, the efficacy of hierarchical fusion inherently depends on a robust foundation of high-quality local cues. Table 7 quantitatively validates this progressive paradigm. A naive one-stage joint training approach yields a limited unseen mAP of 28.78 and a Full mAP of 33.18. By explicitly decoupling the optimization process, our two-stage strategy elevates the unseen mAP to 29.49 and Full mAP to 33.95. This performance gap confirms that establishing fine-grained features via the disentangling module first is critical to prevent fusion gradients from prematurely interfering with the orthogonal projection subspace.

Robustness Against Scene Co-Occurrence Bias. One of the goals of DRIC is to mitigate background shortcuts and ensure robust cross-scene transfer. We explicitly quantify this in Tab. 8 by replacing original test backgrounds with task-irrelevant scenes. The baseline model suffers a severe 12.26% performance drop, verifying its habitual reliance on background shortcuts. While LAIN reduces this drop to 3.56%, DRIC demonstrates superior robustness with a minimal drop rate of 2.18% and a perturbed Full mAP of 33.21. This provides compelling evidence that our framework successfully suppresses spurious correlations, ensuring that the final decisions rely strictly on actual interaction-centric visual evidence.

4.4 Qualitative Results

To intuitively understand how DRIC improves zero-shot HOI detection, we visualize the activation maps derived from the final interaction score via Grad-CAM [39] in Fig. 5.

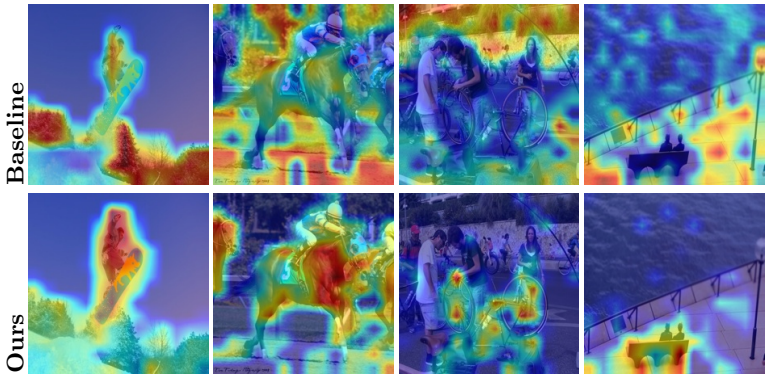


Fig. 5: Qualitative comparison of class activation maps using Grad-CAM [39]. We visualize the activation maps from the final interaction score.

As observed in the **Top Row**, the baseline model exhibits a strong global bias and relies heavily on background shortcuts. It is consistently distracted by the salient environmental context rather than the actual interaction. For instance, the baseline incorrectly anchors its decisions on the trees in the snowboarding scene (first column), the race track (second column), and the water or railings in the sitting scene (fourth column). Furthermore, in cluttered multi-person scenarios (third column), the baseline fails to isolate the active agents, with its activation spreading diffusely across background pedestrians.

In contrast, as shown in the **Bottom Row**, our DRIC framework generates sharp, instance-specific activation maps. The model precisely attends to the critical regions responsible for the actual interactions, such as the contact points between the human and the snowboard, the hands manipulating the bicycle wheels, and the specific location of the bench. This qualitative visualization provides compelling evidence that our method successfully learns fine-grained, discriminative interaction cues, thereby ensuring robust zero-shot transferability without relying on spurious context shortcuts.

5 Conclusion

In this work, we present DRIC, a novel framework designed to enhance the fine-grained representation capability of CLIP for zero-shot HOI detection. To overcome the prevalent limitations of global semantic bias and background shortcut learning, we introduce two complementary components. First, the Context-Orthogonal Residual Disentangling (CORD) module explicitly decouples and reintegrates fine-grained interaction cues via orthogonal projection. Second, the Stratified Local Fusion (SLF) mechanism aggregates multi-layer features to recover essential spatial details and suppress background reliance. Extensive experiments on the HICO-DET benchmark demonstrate that DRIC establishes a new state-of-the-art across all zero-shot evaluation settings.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China (2023YFC3603600), in part by the National Natural Science Foundation of China (62171123), in part by the China Postdoctoral Science Foundation under Grant 2024M750444, in part by the Jiangsu Funding Program for Excellent Postdoctoral Talent under Grant 2024ZB701 and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20241304 and the Big Data Computing Center of Southeast University.

References

1. Cao, Y., Tang, Q., Su, X., Chen, S., You, S., Lu, X., Xu, C.: Detecting any Human-Object Interaction relationship: Universal HOI detector with spatial prompt learning on foundation models. In: Adv. Neural Inform. Process. Syst. vol. 36, pp. 739–751 (2023)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End object detection with transformers. In: Eur. Conf. Comput. Vis. pp. 213–229 (2020)
3. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect Human-Object Interactions. In: WACV (2018)
4. Chen, M., Liao, Y., Liu, S., Chen, Z., Wang, F., Qian, C.: Reformulating HOI detection as adaptive set prediction. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9004–9013 (2021)
5. Chen, M., Chen, M., Yang, Y.: UAHOI: Uncertainty-aware robust interaction learning for HOI detection. Computer Vision and Image Understanding **247**, 104091 (2024)
6. Fang, S., Lin, Z., Yan, K., Li, J., Lin, X., Ji, R.: HODN: Disentangling human-object feature for HOI detection. IEEE Transactions on Multimedia **26**, 3125–3136 (2023)
7. Faure, G.J., Chen, M.H., Lai, S.H.: Holistic interaction transformer network for action detection. In: WACV. pp. 3340–3350 (2023)
8. Gao, C., Zou, Y., Huang, J.B.: iCAN: Instance-centric attention network for Human-Object Interaction detection. In: Brit. Mach. Vis. Conf. (2018)
9. Gupta, S., Malik, J.: Visual semantic role labeling. arXiv preprint arXiv:1505.04474 (2015)
10. Gupta, T., Schwing, A., Hoiem, D.: No-frills Human-Object Interaction detection: Factorization, layout encodings, and training techniques. In: Int. Conf. Comput. Vis. pp. 9677–9685 (2019)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Int. Conf. Comput. Vis. pp. 2961–2969 (2017)
12. Hong, J., Wei, J., Wang, W.: Learning Human-Object Interaction as groups. arXiv preprint arXiv:2510.18357 (2025)
13. Hou, Z., Peng, X., Qiao, Y., Tao, D.: Visual Compositional Learning for Human-Object Interaction detection. In: Eur. Conf. Comput. Vis. pp. 584–600 (2020)
14. Hou, Z., Yu, B., Qiao, Y., Peng, X., Tao, D.: Affordance transfer learning for Human-Object Interaction detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 495–504 (2021)

15. Hou, Z., Yu, B., Qiao, Y., Peng, X., Tao, D.: Detecting Human-Object Interaction via Fabricated Compositional Learning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14646–14655 (2021)
16. Hou, Z., Yu, B., Tao, D.: Compositional 3D human-object neural animation. arXiv preprint arXiv:2304.14070 (2023)
17. Ji, W., Shi, H., Zhang, X.Y.: Explicit Multi-Modal graph modeling for Human-Object Interaction detection. arXiv preprint arXiv:2509.12554 (2025)
18. Ji, Y., Liu, Y., Zhuo, Y., Yu, W., Ma, F., Huang, J., Yu, F.: OnlineHOI: Towards online Human-Object Interaction generation and perception. arXiv preprint arXiv:2509.12250 (2025)
19. Kang, D., Jeong, D., Lee, H., Park, S., Park, H., Kwon, S., Kim, Y., Paik, J.: VLM-HOI: Vision-Language Models for interpretable Human-Object Interaction analysis. In: Eur. Conf. Comput. Vis. pp. 218–235 (2024)
20. Kim, B., Choi, T., Kang, J., Kim, H.J.: UnionDet: Union-level detector towards real-time Human-Object Interaction detection. In: Eur. Conf. Comput. Vis. pp. 498–514 (2020)
21. Kim, B., Lee, J., Kang, J., Kim, E.S., Kim, H.J.: HOTR: End-to-End Human-Object Interaction detection with transformers. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 74–83 (2021)
22. Kim, S., Jung, D., Cho, M.: Locality-aware zero-shot Human-Object Interaction detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 20190–20200 (2025)
23. Lei, Q., Wang, B., Tan, R.: EZ-HOI: VLM adaptation via guided prompt learning for Zero-Shot HOI detection. In: Adv. Neural Inform. Process. Syst. vol. 37, pp. 55831–55857 (2024)
24. Lei, T., Caba, F., Chen, Q., Jin, H., Peng, Y., Liu, Y.: Efficient adaptive Human-Object Interaction detection with Concept-Guided memory. In: Int. Conf. Comput. Vis. pp. 6480–6490 (2023)
25. Lei, T., Yin, S., Peng, Y., Liu, Y.: Exploring conditional Multi-Modal prompts for Zero-Shot HOI detection. In: Eur. Conf. Comput. Vis. pp. 1–19 (2024)
26. Li, Y.L., Liu, X., Lu, H., Wang, S., Liu, J., Li, J., Lu, C.: Detailed 2D-3D joint representation for Human-Object Interaction. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10166–10175 (2020)
27. Li, Y.L., Zhou, S., Huang, X., Xu, L., Ma, Z., Fang, H.S., Wang, Y., Lu, C.: Transferable interactiveness knowledge for Human-Object Interaction detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3585–3594 (2019)
28. Liao, Y., Zhang, A., Lu, M., Wang, Y., Li, X., Liu, S.: GEN-VLKT: Simplify association and enhance interaction understanding for HOI detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 20123–20132 (2022)
29. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Int. Conf. Comput. Vis. pp. 2980–2988 (2017)
30. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Eur. Conf. Comput. Vis. pp. 740–755. Springer (2014)
31. Lin, X., Shi, C., Yang, Z., Tang, H., Zhou, Z.: Sgc-net: stratified granular comparison network for open-vocabulary hoi detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4539–4549 (2025)
32. Liu, Y., Yuan, J., Chen, C.W.: ConsNet: Learning consistency graph for Zero-Shot Human-Object Interaction detection. In: ACM Int. Conf. Multimedia. pp. 4235–4243 (2020)

33. Mao, Y., Deng, J., Zhou, W., Li, L., Fang, Y., Li, H.: CLIP4HOI: Towards adapting CLIP for practical Zero-Shot HOI detection. In: Adv. Neural Inform. Process. Syst. vol. 36, pp. 45895–45906 (2023)
34. Ning, S., Qiu, L., Liu, Y., He, X.: HOICLIP: Efficient knowledge transfer for HOI detection with Vision-Language Models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 23507–23517 (2023)
35. Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.C.: Learning Human-Object Interactions by graph parsing neural networks. In: Eur. Conf. Comput. Vis. pp. 401–417 (2018)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. pp. 8748–8763. PMLR (2021)
37. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with Region Proposal Networks. In: Adv. Neural Inform. Process. Syst. vol. 28 (2015)
38. Sadhu, A., Gupta, T., Yatskar, M., Nevatia, R., Kembhavi, A.: Visual semantic role labeling for video understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5589–5600 (2021)
39. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Int. Conf. Comput. Vis. pp. 618–626 (2017)
40. Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M.J., Mei, T.: Monocular, One-Stage, regression of multiple 3D people. In: Int. Conf. Comput. Vis. pp. 11179–11188 (2021)
41. Tamura, M., Ohashi, H., Yoshinaga, T.: QPIC: Query-Based pairwise Human-Object Interaction detection with Image-Wide contextual information. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10410–10419 (2021)
42. Wan, B., Tuytelaars, T.: Exploiting CLIP for zero-shot HOI detection requires knowledge distillation at multiple levels. In: WACV. pp. 1805–1815 (2024)
43. Wang, H., Zhang, Z., Hu, K., Liu, J., Zeng, X., Ji, K., Pan, M., Fan, Y., Gui, T., et al.: HOID-R1: Reinforcement learning for open-world Human-Object Interaction detection reasoning with Multimodal Large Language Model. arXiv preprint arXiv:2508.11350 (2025)
44. Wang, Q., Song, L., Xie, R., Zuo, L., Cheng, Z., Ling, J., Wang, R., Lin, Z.: PA-HOI: A physics-aware human and object interaction dataset. arXiv preprint arXiv:2508.06205 (2025)
45. Wang, Y., Lei, Y., Cui, L., Xue, W., Liu, Q., Wei, Z.: A review of Human-Object Interaction detection. In: International Conference on Computer, Vision and Intelligent Technology (ICCVIT). pp. 1–6 (2024)
46. Wang, Z., Zheng, Q., Ma, S., Ye, M., Zhan, Y., Li, D.: End-to-end HOI reconstruction transformer with Graph-Based encoding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 27706–27715 (2025)
47. Wei, X.S., Xu, H.Y., Yang, Z., Duan, C.L., Peng, Y.: Negatives make a positive: An embarrassingly simple approach to semi-supervised few-shot learning. IEEE Trans. Pattern Anal. Mach. Intell. **46**(4), 2091–2103 (2023)
48. Wei, X.S., Xu, H.Y., Zhang, F., Peng, Y., Zhou, W.: An embarrassingly simple approach to semi-supervised few-shot learning. In: Adv. Neural Inform. Process. Syst. vol. 35, pp. 14489–14500 (2022)
49. Wu, M., Gu, J., Shen, Y., Lin, M., Chen, C., Sun, X.: End-to-End Zero-Shot HOI detection via vision and language knowledge distillation. In: AAAI. vol. 37, pp. 2839–2846 (2023)

50. Xie, C., Zeng, F., Hu, Y., Liang, S., Wei, Y.: Category query learning for Human-Object Interaction classification. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 15275–15284 (2023)
51. Xue, W., Liu, Q., Wang, Y., Wei, Z., Xing, X., Xu, X.: Towards zero-shot Human-Object Interaction detection via Vision-Language integration. Neural Networks **187**, 107348 (2025)
52. Yang, C., Song, T., Park, J., Oh, C.: Visual diversity and Region-Aware prompt learning for Zero-Shot HOI detection. In: Adv. Neural Inform. Process. Syst. (2025)
53. Yang, K., Liu, R., Stiefelhagen, R., Peng, K., Chen, Y., Wen, D., Roitberg, A., Zheng, J.: RoHOI: Robustness benchmark for Human-Object Interaction detection. arXiv preprint arXiv:2507.09111 (2025)
54. Yuan, H., Jiang, J., Albanie, S., Feng, T., Huang, Z., Ni, D., Tang, M.: RLIP: Relational Language-Image Pre-Training for Human-Object Interaction detection. In: Adv. Neural Inform. Process. Syst. vol. 35, pp. 37416–37431 (2022)
55. Zhang, F.Z., Campbell, D., Gould, S.: Spatially conditioned graphs for detecting Human-Object Interactions. In: Int. Conf. Comput. Vis. pp. 13319–13327 (2021)
56. Zhang, F.Z., Campbell, D., Gould, S.: Efficient Two-Stage detection of Human-Object Interactions with a novel Unary-Pairwise transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 20104–20112 (2022)
57. Zhang, F.Z., Yuan, Y., Campbell, D., Zhong, Z., Gould, S.: Exploring predicate visual context in detecting of Human-Object Interactions. In: Int. Conf. Comput. Vis. pp. 10411–10421 (2023)
58. Zhong, X., Ding, C., Hu, Y., Tao, D.: Disentangled interaction representation for One-Stage Human-Object Interaction detection. arXiv preprint arXiv:2312.01713 (2023)
59. Zhu, M., Ho, E.S., Chen, S., Yang, L., Shum, H.P.: Geometric features enhanced Human-Object Interaction detection. IEEE Transactions on Instrumentation and Measurement (2024)