FineHOI: Part-Aware Dense Representations for Zero-Shot Human-Object Interaction Detection

Francesco Tonini*
University of Trento
Trento, Italy
Fondazione Bruno Kessler
Trento, Italy

Lorenzo Vaquero
Fondazione Bruno Kessler
Trento, Italy

Mohammad Mahdi Derakhshani
University of Amsterdam
Amsterdam, The Netherlands

Cees Snoek
University of Amsterdam
Amsterdam, The Netherlands

Elisa Ricci
University of Trento
Trento, Italy
Fondazione Bruno Kessler
Trento, Italy

Cigdem Beyan
Department of Computer Science
University of Verona
Verona, Italy

Abstract

Human-Object Interaction (HOI) detection aims to localize humans and objects in images and classify their interactions. Zero-shot HOI focuses on recognizing interactions that are not observed during training, requiring models to generalize beyond seen verb-object compositions. Recent approaches leverage Vision-Language Models (VLMs), benefiting from rich semantic representations. However, they often rely on global or detector-centric features that compress interaction cues and hinder fine-grained spatial reasoning. To overcome this limitation, we propose FineHOI, a zero-shot HOI framework that explicitly models interactions from dense patch-level features. Our approach is motivated by the observation that human-object interactions are defined by localized spatial relationships, which are not preserved by global and detector-centric representations. To this end, we introduce an Adaptive Part-Level Attention module that decomposes humans and objects into semantically coherent parts via unsupervised clustering, and re-weights them based on their interaction relevance. These representations are then integrated through a Region-Aware Interaction Transformer that integrates part-aware and global features and produces the final HOI embedding. Extensive experiments demonstrate that FineHOI consistently outperforms existing zero-shot HOI methods, achieving particularly strong gains on unseen interactions. Code is available here.

CCS Concepts

• **Computing methodologies** → **Machine learning**.

Keywords

Human-object interaction, vision-language models

*Work done while at the University of Amsterdam.

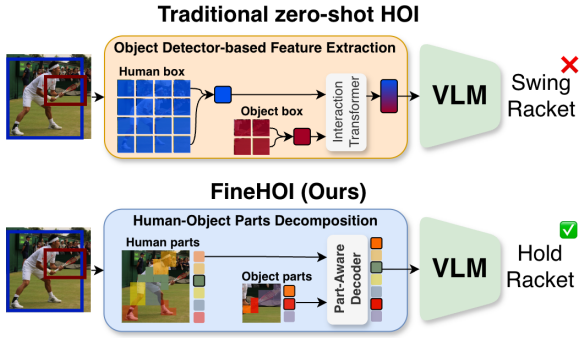


Figure 1: Comparison of zero-shot HOI approaches. Traditional methods, *e.g.* [18, 19, 21] rely on object-level features which do not capture the fine-grained details needed for interaction classification. In contrast, FineHOI explicitly models dense patch-level interactions through human/object parts decomposition, and can dynamically select which parts are most relevant for an interaction (*e.g.* hands, upper-body, object handle), enabling finer understanding of human-object interactions.

ACM Reference Format:
Francesco Tonini, Lorenzo Vaquero, Mohammad Mahdi Derakhshani, Cees Snoek, Elisa Ricci, and Cigdem Beyan. 2026. FineHOI: Part-Aware Dense Representations for Zero-Shot Human-Object Interaction Detection. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3836016

1 Introduction

Human-Object Interaction (HOI) detection aims to locate human-object pairs in images and identify their interactions as a set of ⟨*human*, *verb*, *object*⟩ triplets. HOI models are critical for a range of downstream applications, including image and video captioning [13, 30, 39], visual surveillance [22], and autonomous driving [5], as they enhance perception and scene understanding in complex environments. Existing HOI methods [35, 43] rely heavily on large-scale annotated datasets [4, 10] to learn interaction representations. However, such annotations are expensive to obtain, inherently ambiguous, and difficult to scale to new domains or rare, long-tail interactions.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
MM '26, Rio de Janeiro, Brazil
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3836016

Recently, zero-shot HOI models have emerged as a promising direction to address these limitations [1, 9, 14, 18, 19, 21, 27, 37]. Unlike conventional closed-set HOI approaches [8, 16, 33, 36, 40, 43], which are limited to recognizing interactions seen during training, zero-shot HOI models aim to generalize to unseen interactions. In this setting, models are trained on a subset of interaction categories and evaluated on disjoint unseen interactions, including novel verb-object compositions as well as unseen verbs or objects, requiring compositional generalization beyond observed data. This poses a significant challenge, as accurate recognition depends on capturing transferable and fine-grained interaction cues rather than memorizing combinations of verb-objects seen during training time.

This progress has been largely enabled by Vision-Language Models (VLMs), such as CLIP [28], which learn a shared multimodal embedding space that can be leveraged for zero-shot HOI detection. This enables zero-shot recognition of novel interactions at test time by ranking text prompts for unseen verbs and objects against the visual embedding of the detected human-object pair. Several recent works have explored this paradigm through techniques such as prompt learning [19], region-aware prompting [37], and lightweight adaptation of pretrained multimodal models [18, 21, 27]. Some approaches further incorporate auxiliary language supervision, derived from Large Language Models (LLMs) [19, 37], while others focus on adapting VLMs without additional supervision [18, 21, 27].

However, existing HOI detection methods rely on object detectors pretrained on large-scale datasets and build interaction representations from detector-centric features [18, 21, 27, 37]. As shown in Fig. 1, these features are typically instantiated as pooled region descriptors, object-centric embeddings, or detector queries, which summarize each human and object into a coarse representation. Although effective for localization and coarse semantic understanding, this early compression can discard the localized spatial cues needed to distinguish subtle interactions. This early aggregation is particularly problematic in the zero-shot setting, where generalization depends on preserving transferable local evidence rather than relying on coarse object-level patterns. For example, as shown in Fig. 1, distinguishing whether a person is *holding* a tennis racket rather than *swinging* depends on localized cues around the hand-racket contact region together with the surrounding pose. Closely related to our approach, LAIN [18] introduces feature-level adaptations in the visual encoder to better capture locality. However, it still builds each human-object representation from detector-derived object-centric features, distilling the features into a single token for classification. Thus, although LAIN enhances the visual features, its final interaction representation remains compressed at the pair level, which can weaken the fine-grained features required to distinguish subtle HOIs.

To overcome these limitations, we propose **FineHOI**, a zero-shot HOI framework that constructs interaction representations directly from dense image features instead of detector-centric embeddings.

We hypothesize that detector features, although effective for localization, are too object-centric and spatially coarse to support fine-grained HOI reasoning. Accordingly, we use the detector only to localize the regions of interest (*i.e.*, human and object bounding boxes), and derive visual features from dense image patches rather than detector features. Furthermore, motivated by the observation that HOI cues are often localized within specific contact regions (*e.g.*, hands and racket handle, as in Fig. 1), rather than distributed across the entire instances, we introduce a novel *Adaptive Part-Level Attention* module, which decomposes human and object regions into structurally coherent parts and re-weights them based on their interaction relevance. This enables the model to focus on interaction-critical regions while preserving fine-grained spatial structure.

The resulting representations are processed by a novel *Region-Aware Interaction Transformer* that integrates interaction-relevant parts with aggregated patch features to form the final HOI representation, thereby preserving both the global context of the interaction and the localized evidence needed to disambiguate fine-grained visual relations in the zero-shot setting. Experiments demonstrate consistent improvements across multiple zero-shot settings, with particularly strong gains in challenging scenarios involving unseen interactions. Our contributions are fourfold:

- We introduce FineHOI, a zero-shot HOI detection method that builds interaction representations directly from dense patch-level features, avoiding early compression into object-level embeddings and preserving fine-grained spatial cues.
- We propose an Adaptive Part-Level Attention module that performs unsupervised part decomposition with cross-entity attention to identify and emphasize interaction-relevant regions, enabling explicit modeling of localized human-object contact.
- We design a Region-Aware Interaction Transformer that integrates part-aware representations with global context through dual-path reasoning, improving robustness and generalization to unseen interactions.
- We achieve consistent improvements across all zero-shot settings, outperforming prior methods while not relying on detector features or external LLM supervision.

2 Related Work

One-stage vs two-stage HOI Detection. HOI detection aims to localize human-object pairs and classify their interactions. Existing methods can be broadly categorized into one-stage and two-stage approaches. One-stage methods formulate HOI detection as a set prediction problem, typically following DETR-like architectures with learnable queries to jointly predict humans, objects, and interactions. While effective in supervised closed-set settings, these methods require extensive annotations and do not generalize to unseen interactions [16, 17, 35, 45].

Two-stage methods, on the other hand, first detect human and object instances using pretrained object detectors and then classify interactions for each candidate pair [4, 7, 11, 36, 41, 44]. These approaches leverage various cues, including spatial configurations [7, 36, 41], visual attention [42], and contextual reasoning [8, 25, 34, 41], to model interactions. Due to their modular design and flexibility, two-stage pipelines are often adopted in settings requiring generalization, such as zero-shot HOI detection. Our method follows this paradigm.

Zero-shot HOI Detection. Zero-shot HOI detection aims to recognize interactions that are not observed during training [1]. Early approaches address this challenge through compositional learning [1, 14], while more recent methods leverage VLMs, such as CLIP [28], to transfer semantic knowledge to unseen interactions [2, 9, 18, 19, 21, 27]. These approaches typically align visual features of human-object pairs with textual representations

of interaction labels through prompt learning [19], proposal-based reasoning [37], or lightweight adaptation strategies [18].

To improve generalization, recent works incorporate additional language information. Some approaches [19, 37] exploit auxiliary language supervision to enrich interaction representations, while others [18, 21, 27] focus on adapting VLMs without additional supervision. However, all these methods rely on object detectors pretrained on large-scale datasets to generate human-object proposals [18, 21, 27, 37], and build interaction representations from detector-centric features or global embeddings, which tend to summarize the human-object pair too early and can discard the localized visual cues needed to distinguish fine-grained interactions.

Despite these recent advances, these methods overlook a crucial aspect, *i.e.* the explicit modeling of structured local spatial relationships, which is critical for fine-grained detection of HOIs. In contrast, our proposed FineHOI explicitly models the interaction structure by decomposing image regions corresponding to the human and object into parts and reasoning over their relationships. Rather than relying on global embeddings or detector-centric features, we construct interaction representations directly from dense patch-level visual features, enabling structured and localized HOI modeling. Closely related to our approach, LAIN [18] improves locality in the visual encoder. However, as shown in Fig. 1, it still relies on detector features and reduces each human-object pair to a single representation at an early stage. As a result, fine-grained contact cues are mostly lost.

3 FineHOI

FineHOI builds interaction representations from the dense patch-level spatial features of a VLM visual encoder, enabling more fine-grained HOI recognition. As illustrated in Fig. 2, given an input image, FineHOI first employs an off-the-shelf object detector to localize human and object instances, then uses a VLM to extract dense patch-level spatial features and generate soft masks for the corresponding regions of interest (Sec. 3.1). Next, it partitions the human and object regions into structurally coherent parts and estimates their interaction relevance to produce refined part-aware representations (Sec. 3.2). Finally, FineHOI fuses the refined human-object representations into a joint interaction embedding, which is used for HOI classification.

Problem Formulation. Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and candidate sets of object $\mathbb{O}$ and verb $\mathbb{V}$ labels, respectively, the goal of HOI detection is to locate and classify all interactions present in the image. Formally, the i-th HOI instance is represented as $(\mathbf{b}_{h_i}, \mathbf{b}_{o_i}, o_i, v_i)$, where $\mathbf{b}_{h_i}, \mathbf{b}_{o_i} \in \mathbb{R}^4$ denote the human and object bounding boxes, while $o_i \in \mathbb{O}$ and $v_i \in \mathbb{V}$ denote the object and verb label. In the zero-shot HOI setting, training and evaluation are performed on disjoint sets of valid interaction classes, each defined by a $\langle verb, object\rangle$ combination. Specifically, the model is trained only on seen interactions and evaluated on unseen interactions that are not observed during training.

3.1 Object Detection and Soft-Mask Extraction

As seen in Fig. 2 in red, an off-the-shelf object detector $\mathcal{OD}$ processes the input image $\mathbf{I}$ to produce a set of object proposals $\{d_i\}_{i=1}^{N}$, where each proposal $d_i = (\mathbf{b}_i, s_i, c_i)$ consists of a bounding box $\mathbf{b}_i \in \mathbb{R}^4$, a detection confidence score $s_i \in [0, 1]$, and a class label $c_i \in \mathbb{O}$. After non-maximum suppression, detections are partitioned into human and object boxes, and all possible human-object pairs are built. For ease of notation, we consider a single human-object pair, where $d_h = (\mathbf{b}_h, s_h, c_h)$ and $d_o = (\mathbf{b}_o, s_o, c_o)$ denote the human and object detections, respectively.

Let $\mathcal{V}$ and $\mathcal{T}$ denote the visual and text encoders of a VLM, respectively, whose outputs lie in a shared embedding space. Given $\mathbf{I}$, the visual encoder $\mathcal{V}$ produces a spatial feature map $\mathbf{F} \in \mathbb{R}^{H_p\times W_p\times D}$, where $H_p \times W_p$ is the patch-grid resolution and D is the embedding dimension.

For each patch-grid location (x, y) of the image feature map $\mathbf{F}$, with $x \in \{1, \ldots, H_p\}$ and $y \in \{1, \ldots, W_p\}$, we define the human soft mask as:

$$\mathbf{M}_h^{(x,y)} = \mathbb{1}\left[(x, y) \in \mathbf{b}_h\right] s(\mathbf{F}^{(x,y)}, \mathbf{t}_h) \tag{1}$$

where $s(\cdot, \cdot)$ denotes cosine similarity, $\mathbf{t}_h = \mathcal{T}(c_h) \in \mathbb{R}^D$ is the text embedding of the human, $c_h =$ "person", $\mathbb{1}[\cdot]$ is the indicator function, and $(x, y) \in \mathbf{b}_h$ indicates that the patch-grid location (x, y) lies within the human bounding box $\mathbf{b}_h$. This formulation yields a mask $\mathbf{M}_h$ that suppresses patches outside the bounding box while preserving class-conditioned similarity within it. This enables the model to focus on semantically relevant regions while filtering out background noise, providing a more discriminative representation for interaction reasoning. The object soft mask $\mathbf{M}_o$ is defined analogously using the object label $\mathbf{c}_o$ and the object bounding box $\mathbf{b}_o$.

3.2 Adaptive Part-Level Attention

While the soft masks from the previous stage indicate *where* the human and the object are located within the proposed bounding boxes, they do not capture *how* the two entities interact. HOIs are inherently local (*e.g.*, a hand grasping a cup or a foot kicking a ball), and not all image regions associated with the human and the object contribute equally. Identifying the parts of the human and the object that are most relevant to the interaction is therefore crucial for extracting fine-grained features.

To address this, we introduce a novel *Adaptive Part-Level Attention* module (Fig. 2, orange block), which (a) partitions patch-level features into semantically-coherent parts, and then (b) uses said parts to adapt the corresponding visual features according to their relevance to the interaction.

Unsupervised Part Decomposition. For each human and object, we first extract the patch features from $\mathbf{F}$ that lie within their respective bounding boxes, yielding the crop-level features $\tilde{\mathbf{F}}_h \in \mathbb{R}^{H_h\times W_h\times D}$ and $\tilde{\mathbf{F}}_o \in \mathbb{R}^{H_o\times W_o\times D}$, where $H_h \times W_h$ and $H_o \times W_o$ corresponds to the size of the human and object bounding boxes on the patch grid. By isolating these features before decomposition, we ensure that we focus exclusively on the human/object structure, avoiding interference from the surrounding context.

We then apply a modified version of Normalized Cut [31] (NCut) that acts on feature patches rather than pixels. This partitions $\tilde{\mathbf{F}}_h$ and $\tilde{\mathbf{F}}_o$ into K parts, producing structurally coherent regions that capture meaningful object components, enabling more fine-grained interaction modeling than holistic region representations. We derive the visual representation of each human part from its corresponding spatial region in the human soft mask $\mathbf{M}_h$. Specifically, for each part $k \in \{1, \ldots, K\}$, we obtain its features via mask-weighted pooling

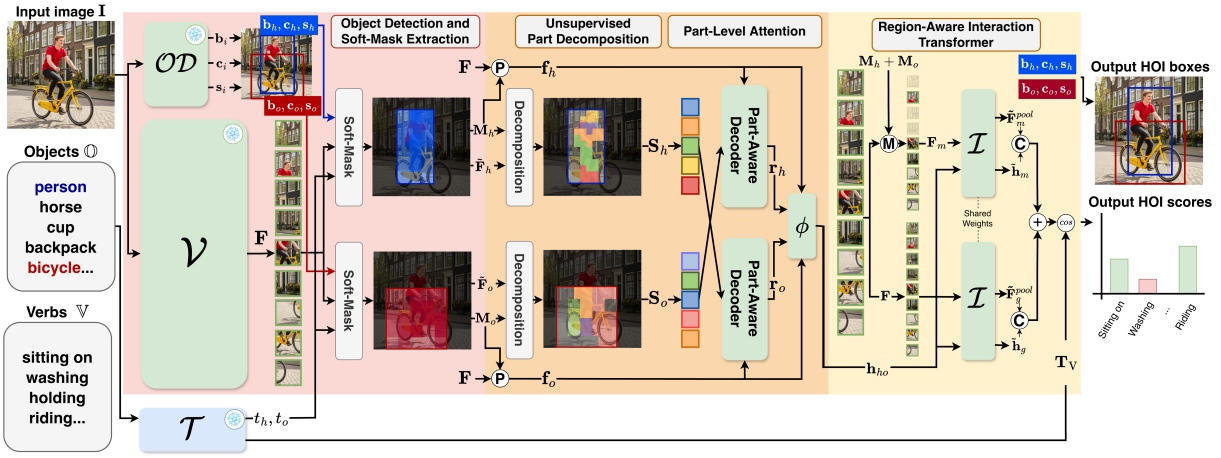


Figure 2: FineHOI consists of three stages: Object Detection and Soft-Mask Extraction, which uses an object detector $\mathcal{OD}$ to localize the human and object instances, while a visual encoder $\mathcal{V}$ extracts dense spatial feature map used for human and object soft masks construction; Adaptive Part-Level Attention, which decomposes the human and object patches into structurally coherent parts and refines them based on interaction relevance; and Region-Aware Interaction Transformer, which fuses the refined human-object representations with contextual patch features for HOI classification. Ⓟ stands for feature pooling (Eq. (3)), Ⓜ for mask-guided feature pooling (Eq. (9)), and Ⓒ for concatenation.

over the crop features:

$$\mathbf{s}_{h,k} = \frac{\sum_{(x,y)\in \mathbf{P}_{h,k}} \mathbf{M}_h^{(x,y)} \tilde{\mathbf{F}}_h^{(x,y)}}{\sum_{(x,y)\in \mathbf{P}_{h,k}} \mathbf{M}_h^{(x,y)}}. \tag{2}$$

Here, $\mathbf{P}_{h,k}$ denotes the set of patch-grid locations belonging to the k-th human part. The object part features $\mathbf{s}_{o,k}$ are defined analogously using $\mathbf{M}_o$, $\tilde{\mathbf{F}}_o$, and $\mathbf{P}_{o,k}$. The use of $\mathbf{M}_h$, $\mathbf{M}_o$ ensures that background patches receive lower weight even within each part, keeping the representation focused on semantically relevant regions. This reduces noise from irrelevant context and improves the discriminability of interaction features. We apply this procedure to all human and object parts, which yields the corresponding sets of part features, $\mathbf{S}_h = \{\mathbf{s}_{h,k}\}_{k=1}^K$ and $\mathbf{S}_o = \{\mathbf{s}_{o,k}\}_{k=1}^K$, for the human and object, respectively.

Part-Aware Decoder. We argue that not all parts contribute equally to interaction prediction. For example, when predicting ⟨*human, riding, bike*⟩, the parts corresponding to the hands and legs are more informative than those corresponding to the torso. Accordingly, after decomposing humans and objects into parts, we use the cross-entity attention mechanism described below to adaptively modulate the contribution of each part.

First, we compute a pooled feature vector for the human branch by performing mask-weighted average pooling over all patch features:

$$\mathbf{f}_h = \frac{\sum_{x,y} \mathbf{M}_h^{(x,y)} \mathbf{F}^{(x,y)}}{\sum_{x,y} \mathbf{M}_h^{(x,y)}}. \tag{3}$$

The object feature vector $\mathbf{f}_o$ is defined analogously using the object soft mask $\mathbf{M}_o$. These feature vectors will act as queries and attend to the part features of the other entity, enabling the model to adaptively identify interaction-relevant parts based on the pooled representation of each entity. Specifically, the object feature vector $\mathbf{f}_o$ will attend over $\mathbf{S}_h$ to identify the interacting parts (*e.g.*, feet), while the human feature vector $\mathbf{f}_h$ will attend over object parts $\mathbf{S}_o$ to identify the corresponding contact parts (*e.g.*, pedals).

To this end, given the human feature vector $\mathbf{f}_h$, we compute an interaction-aware representation $\mathbf{r}_h \in \mathbb{R}^{d_v}$ via cross-attention:

$$\mathbf{Q} = \mathbf{f}_h \mathbf{W}_q, \qquad \mathbf{K} = \mathbf{S}_o \mathbf{W}_k, \qquad \mathbf{V} = \mathbf{S}_o \mathbf{W}_v, \tag{4}$$

$$\mathbf{r}_h = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_v}}\right)\mathbf{V}, \tag{5}$$

where $\mathbf{W}_q \in \mathbb{R}^{D\times D}$, $\mathbf{W}_k \in \mathbb{R}^{D\times D}$, and $\mathbf{W}_v \in \mathbb{R}^{D\times D}$ are learned projection matrices. This formulation allows the model to attend to interaction-relevant parts. By conditioning each entity on the parts of the other, it explicitly models the contact regions that define the interaction. We also apply the above process to the object feature vector $\mathbf{f}_o$ using the human part features $\mathbf{S}_h$, yielding $\mathbf{r}_o$.

Finally, we combine the pooled and interaction-aware representations to obtain a unified human-object representation:

$$\mathbf{h}_{ho} = \phi(\mathbf{f}_h \,\|\, \mathbf{f}_o \,\|\, \mathbf{r}_h \,\|\, \mathbf{r}_o), \tag{6}$$

where $\phi : \mathbb{R}^{4D} \to \mathbb{R}^D$ is a learned linear projection, and $\|$ denotes concatenation. This vector combines global semantic information with localized interaction-aware cues from the attended parts.

3.3 Region-Aware Interaction Transformer

Each human-object token $\mathbf{h}_{ho}$ captures pair-specific interaction cues. Nevertheless, a reliable prediction of the HOI also requires contextual information from the surrounding image. To integrate these complementary sources of evidence, we introduce a dual-path architecture that explicitly models global and local information separately (Fig. 2, yellow block). It leverages a region-aware interaction transformer $\mathcal{I}$: an encoder-only Vision Transformer [6] model with shared parameters but different visual inputs, combining scene-level context with interaction-relevant parts.

In the global path, the human-object pair token $\mathbf{h}_{ho}$ is processed together with the image patch features $\mathbf{F}$ through the interaction transformer $\mathcal{I}$:

$$\tilde{\mathbf{h}}_g, \tilde{\mathbf{F}}_g = \mathcal{I}(\mathbf{h}_{ho} \,\|\, \mathbf{F}), \tag{7}$$

where $\tilde{\mathbf{h}}_g \in \mathbb{R}^D$ denotes the updated pair token and $\tilde{\mathbf{F}}_g \in \mathbb{R}^{H_p \times W_p \times D}$ denotes the refined patch features. We then construct the global HOI embedding $\mathbf{e}_g \in \mathbb{R}^{2D}$ by concatenating the updated pair token with the spatial average of the refined patch features $\mathbf{F}_g^{pool}$:

$$\mathbf{e}_g = \tilde{\mathbf{h}}_g \,\|\, \underbrace{\frac{1}{H_p W_p} \sum_{x=1}^{H_p} \sum_{y=1}^{W_p} \tilde{\mathbf{F}}_g^{(x,y)}}_{\hat{\mathbf{F}}_g^{pool}}. \tag{8}$$

In the mask-guided path, the human-object pair token $\mathbf{h}_{ho}$ is jointly encoded with mask-guided patch features through the same interaction transformer $\mathcal{I}$. We first define the mask-guided patch features as:

$$\mathbf{F}_m^{(x,y)} = \left(\mathbf{M}_h^{(x,y)} + \mathbf{M}_o^{(x,y)}\right) \mathbf{F}^{(x,y)}, \tag{9}$$

and then apply the interaction transformer to the pair token and the mask-guided patch features:

$$\tilde{\mathbf{h}}_m, \tilde{\mathbf{F}}_m = \mathcal{I}(\mathbf{h}_{ho} \,\|\, \mathbf{F}_m), \tag{10}$$

where $\tilde{\mathbf{h}}_m \in \mathbb{R}^D$ denotes the updated pair token and $\tilde{\mathbf{F}}_m \in \mathbb{R}^{H_p \times W_p \times D}$ represents the refined mask-guided patch features. We finally construct the mask-guided HOI embedding $\mathbf{e}_m \in \mathbb{R}^{2D}$ combining both the pair token and the average masked feature patches $\mathbf{F}_m^{pool}$, to emphasize localized interaction cues within the human-object pair:

$$\mathbf{e}_m = \tilde{\mathbf{h}}_m \,\|\, \underbrace{\frac{\sum_{x=1}^{H_p} \sum_{y=1}^{W_p} \left(\mathbf{M}_h^{(x,y)} + \mathbf{M}_o^{(x,y)}\right) \tilde{\mathbf{F}}_m^{(x,y)}}{\sum_{x=1}^{H_p} \sum_{y=1}^{W_p} \left(\mathbf{M}_h^{(x,y)} + \mathbf{M}_o^{(x,y)}\right)}}_{\tilde{\mathbf{F}}_m^{pool}}. \tag{11}$$

The two paths capture complementary cues: the global path preserves scene-level context, whereas the mask-guided path emphasizes features associated with high human-object mask responses. Lastly, we ℓ_2-normalize the two embeddings and combine them into the final HOI representation:

$$\mathbf{e}_{hoi} = \alpha \bar{\mathbf{e}}_g + \bar{\mathbf{e}}_m, \tag{12}$$

where α is a learnable scalar that balances the contribution of the global path.

3.4 Training and Inference

Following prior zero-shot HOI methods [18, 19, 21], we cast HOI as a multi-label classification task. Let $\mathbb{V}$ denote the set of verb classes, and let $\mathbf{T}_V \in \mathbb{R}^{|\mathbb{V}| \times 2D}$ be the corresponding text embeddings produced by text encoder $\mathcal{T}$. We compute scores for each verb by comparing the HOI embedding $\mathbf{e}_{hoi}$ with all verb embeddings in the shared space:

$$\mathrm{HOI}_{\mathrm{pred}} = \sigma\left(s(\mathbf{e}_{hoi}, \mathbf{T}_V)/\tau\right), \tag{13}$$

where $s(\cdot,\cdot)$ denotes cosine similarity, σ is the sigmoid function, and τ is a learnable temperature.

During training, for each candidate pair, we construct a binary label vector indicating the valid verb classes and optimize the predicted scores with binary cross-entropy:

$$\mathcal{L}_{\mathrm{hoi}} = \mathrm{FL}(\mathrm{HOI}_{\mathrm{pred}}, \mathrm{HOI}_{\mathrm{gt}}), \tag{14}$$

where $\mathrm{HOI}_{\mathrm{gt}} \in \{0,1\}^{N_{\mathrm{pair}} \times |\mathbb{V}|}$ contains the target verb labels for all N_{pair} human-object pairs, and FL is the binary focal loss [23].

At inference, following prior methods [18, 19, 27] we define the HOI scores as:

$$\mathrm{HOI}_{\mathrm{test}} = \mathrm{HOI}_{\mathrm{pred}} \cdot \mathbf{s}_h^{\lambda} \cdot \mathbf{s}_o^{\lambda}, \tag{15}$$

where $\mathbf{s}_h, \mathbf{s}_o$ are the detection scores of the paired human and object instances, and λ is a fixed hyper-parameter.

4 Experiments

Datasets. We evaluate FineHOI on standard HOI detection benchmarks, HICO-DET [4] and V-COCO [10]. HICO-DET contains 47,776 images, with 38,118 for training and 9,658 for testing, covering 600 HOI categories defined over 117 verbs and 80 objects. V-COCO is built on the COCO [24] dataset and consists of 10,346 images, split into 2,533 training, 2,867 validation, and 4,946 test images, with annotations spanning 29 action categories across 80 object classes.

Evaluation Metrics. Following the standard evaluation protocols used by state-of-the-art zero-shot HOI methods [2, 9, 18, 19, 37], we evaluate FineHOI using mean average precision (mAP) under five zero-shot settings [9, 15]: Unseen Verb (UV), where the 117 verbs are split into 97 seen and 20 unseen classes; Unseen Object (UO), where the 80 objects are split into 68 seen and 12 unseen classes; and three interaction-level settings, Unseen Combination Rare-First (UC-RF), Unseen Combination Non-Rare-First (UC-NF), and Unseen Combination (UC), where the 600 interactions are split into 480 seen and 120 unseen pairs, with the unseen set consisting of either the rarest interactions (RF), the most frequent interactions (NF), or a predefined subset, respectively.

We report performance on Seen and Unseen HOI classes, along with the overall average (Full) and harmonic mean (HM), which provides a balanced measure of generalization across seen and unseen categories. Furthermore, we also evaluate against zero-shot HOI methods in the closed-set setting, and report performance on the Full set, as well as on the Rare and Non-Rare splits, where the 600 HOI categories are divided into 138 Rare classes with fewer than 10 training instances and 462 Non-Rare classes [4]. In V-COCO, we report average precision under the AP^{S2} protocol which evaluates only interactions involving an explicit semantic role object, such as $\langle$*human*, *holding*, *phone*$\rangle$ or $\langle$*human*, *sitting on*, *horse*$\rangle$ [10].

4.1 Implementation Details

Following established practice in zero-shot HOI detection [2, 18, 19], FineHOI employs a frozen object detector, DETR [3], to identify human and object instances. However, unlike state-of-the-art zero-shot HOI methods, we use it only as a region proposal network and do not rely on its features at any stage of the model. We discard detections with confidence scores below 0.2 and retain up to 15 human and object instances per image. Unless otherwise specified, we instantiate the vision $\mathcal{V}$ and text backbones $\mathcal{T}$ using DINOtxt [32]. Consistent with recent zero-shot HOI methods [18, 19], all backbone parameters are kept frozen and adapted only through residual adapters. For unsupervised part decomposition, we adopt the NCut implementation of [38] and set the maximum number of parts to $K = 10$. We train FineHOI for 20 epochs using the AdamW optimizer [26] with a batch size of 4 on a single RTX A6000 GPU. The initial learning rate is set to 10^{-3} and is reduced to 10^{-4} after epoch 10. At inference time,

Table 1: State-of-the-art comparison of zero-shot HOI detection methods on the HICO-DET [4] dataset. Results for prior methods are taken from [18, 37], while results marked with † are reproduced using the official code. Δ indicates the use of detector-centric features. Best results are highlighted in bold, and second-best results are underlined.

Model		Δ	UV				UO				UC-RF				UC-NF				UC			
			Full	Seen	Unseen	HM	Full	Seen	Unseen	HM	Full	Seen	Unseen	HM	Full	Seen	Unseen	HM	Full	Seen	Unseen	HM
w/ LLM supervision																						
ADA-CM [20]	CLIP ViT-B/16	✓	-	-	-	-	-	-	-	-	33.01	34.35	27.36	30.86	31.39	31.13	32.41	31.77	-	-	-	-
EZ-HOI [19]	CLIP ViT-L/14	✓	32.32	33.49	25.10	29.30	32.27	32.06	33.28	32.67	33.13	34.15	29.02	31.59	31.17	30.55	33.66	32.11	-	-	-	-
DYSCO [34]	CLIP ViT-L/14	✓	-	-	-	-	-	-	-	-	25.24	23.96	30.36	27.16	25.18	24.58	27.56	26.07	-	-	-	-
VDRP [37]	CLIP ViT-L/14	✓	37.18	38.16	31.16	34.31	**37.81**	**37.50**	39.39	38.41	38.13	38.48	36.72	37.58	36.46	36.21	37.48	36.83	-	-	-	-
w/o LLM supervision																						
CLIP4HOI [27]	CLIP ViT-B/16	✓	26.02	31.14	30.42	30.78	27.71	33.25	32.11	32.68	28.47	35.48	34.08	34.78	31.44	28.26	28.90	28.58	32.11	33.25	27.71	30.48
HOIGen [9]	CLIP ViT-B/16	✓	32.34	34.31	20.27	33.32	33.48	32.90	36.35	32.63	33.86	34.57	31.01	32.79	33.08	32.86	33.98	33.42	33.44	34.23	30.26	32.25
SSGR [2]	CLIP ViT-B/16	✓	34.37	36.60	20.67	28.49	35.41	35.14	36.73	35.93	35.71	36.98	30.60	33.79	33.80	33.52	34.89	34.20	-	-	-	-
CMMP [21]	CLIP ViT-L/14	✓	36.38	37.28	30.84	33.75	36.74	36.15	39.67	37.83	37.13	37.42	35.98	36.69	35.13	33.53	33.52	34.50	36.56	37.15	34.46	35.75
LAIN [18]	CLIP ViT-L/14	✓	37.20	38.04	32.05	35.05	37.60	36.96	**40.78**	**38.87**	38.13	38.54	36.57	37.55	36.22	35.90	37.52	36.71	36.81	37.95	32.25	35.10
LAIN [18]†	DINOtxt ViT-L/16	✓	35.43	36.09	31.34	33.72	36.01	34.72	38.11	36.41	36.82	37.55	33.90	35.82	33.89	34.09	33.09	33.59	32.97	33.53	30.71	32.12
LAIN [18]†	DINOtxt ViT-L/16		34.50	34.98	31.57	33.27	34.85	34.38	37.18	35.78	34.79	35.56	31.70	33.63	34.80	34.60	35.58	35.09	34.80	35.11	33.59	34.35
FineHOI	DINOtxt ViT-L/16		**38.10**	**38.51**	**35.58**	**37.04**	36.62	35.92	40.13	38.03	**38.39**	**38.55**	**37.79**	**38.17**	**36.58**	**36.30**	**38.11**	**37.15**	**36.88**	**38.00**	**34.79**	**36.10**

Table 2: Evaluation of zero-shot methods on closed-set HOI detection on HICO-DET [4] and V-COCO [10] datasets. Results for prior methods are taken from [18, 37], while results marked with † are reproduced using the official code. Δ indicates the use of detector-centric features. Best results are in bold, second-best are underlined.

Model		Δ	HICO-DET				V-COCO
			Full	Rare	Non-rare	HM	AP^{S2}
w/ LLM supervision							
ADA-CM [21]	CLIP ViT-L/14	✓	38.40	37.52	38.66	38.09	63.9
EZ-HOI [19]	CLIP ViT-L/14	✓	38.61	37.70	38.89	38.30	66.1
DYSCO [34]	CLIP ViT-L/14	✓	28.24	34.22	26.46	30.34	47.8
VDRP [37]†	CLIP ViT-L/14	✓	38.90	39.51	38.71	39.12	66.2
w/o LLM supervision							
CLIP4HOI [27]	CLIP ViT-L/14	✓	35.33	33.95	35.74	34.85	66.3
HOIGen [9]	CLIP ViT-B/16	✓	34.84	34.52	34.94	34.74	-
SSGR [2]	CLIP ViT-B/16	✓	36.86	35.74	37.20	36.47	64.2
CMMP [21]	CLIP ViT-L/14	✓	38.14	37.75	38.25	38.00	64.0
LAIN [18]†	CLIP ViT-L/14	✓	37.30	38.69	36.88	37.79	65.1
LAIN [18]†	DINOtxt ViT-L/16	✓	36.73	35.96	36.95	36.46	63.2
LAIN [18]†	DINOtxt ViT-L/16		34.37	35.19	34.13	34.66	62.3
FineHOI	DINOtxt ViT-L/16		**39.13**	**39.91**	**38.89**	**39.24**	**66.5**

we set $\lambda = 2.8$ in Eq. (15) as in prior works [2, 18, 19]. Additional implementation details are provided in the Supp. Mat..

4.2 Comparison with the State-of-the-art

Zero-shot results. FineHOI achieves state-of-the-art performance across most zero-shot settings on HICO-DET (Tab. 1), particularly on UV, UC-RF, UC-NF, and UC, with consistent gains across these splits. The largest improvements are observed in the UV and UC settings, where, for example, FineHOI reaches **35.58%** unseen mAP, outperforming the second-best method LAIN by **+3.53** points in UV and **+0.33** in UC. Other strong gains are also obtained on UC-RF and UC-NF settings, achieving **37.79%** and **38.11%**, respectively, demonstrating robust generalization to unseen interaction compositions. On the UO setting, FineHOI remains competitive but does not surpass methods that leverage additional sources of object-centric features (**Δ**): LAIN [18] and VDRP [37]. Such features can provide useful cues due to their pretraining on overlapping object classes. We analyze this effect below, showing empirically that incorporating detector-centric features yields noticeable gains for LAIN, while FineHOI already captures object-agnostic interaction cues effectively without relying on such features. Furthermore, the other best performer on UO, VDRP, benefits from external LLM supervision that enriches object semantics. In contrast, FineHOI does not rely on additional LLM-derived supervision. To ensure a fair comparison, we additionally evaluate LAIN under the same backbone and feature setting (*i.e.*, without **Δ**, see Supp. Mat. for the illustration of the usage of **Δ** by the prior methods). Under this setup, FineHOI consistently outperforms all methods across all settings, confirming its effectiveness.

Table 3: Ablation of FineHOI on HICO-DET dataset under the UV setting. SM stands for soft-masks (Sec. 3.1), APA for Adaptive Part-Level Attention (Sec. 3.2), and $\mathcal{I}$ for the Region-Aware Interaction Transformer (Sec. 3.3). Best results in bold.

SM	APA	$\mathcal{I}$	Full	Seen	Unseen	HM
✗	✗	✗	33.39	34.53	26.42	30.48
✓	✗	✗	35.36	36.05	31.13	33.59
✓	✓	✗	35.36	35.55	34.21	34.88
✓	✓	✓	**38.10**	**38.51**	**35.58**	**37.04**

Table 4: Effectiveness of soft-mask on the HICO-DET dataset under the UV setting. Best results in bold.

	Full	Seen	Unseen	HM
RoIAlign	35.61	36.34	31.13	33.73
Uniform	35.41	35.68	33.80	34.74
Soft-mask	**38.10**	**38.51**	**35.58**	**37.04**

Closed-set results. Although FineHOI is designed for zero-shot HOI detection, we also evaluate it in the closed-set setting. As shown in Tab. 2, FineHOI achieves the best performance among all compared methods, reaching **39.13** mAP on HICO-DET, including **39.91** mAP on the Rare split, and **66.5** AP^{S2} on V-COCO. In particular, FineHOI outperforms the strongest competing methods that rely on additional supervision. Compared to VDRP [37], which leverages LLM guidance and detector-centric features, FineHOI improves performance by **+0.40** mAP on the Rare split. Furthermore, compared

to LAIN [18] under the same DINOtxt-based setting (without detector-centric features Δ), FineHOI achieves gains of **+4.76** mAP on Full, **+4.72** mAP on Rare, and **+4.58** HM. This shows that improved visual grounding remains beneficial even when all interaction classes are observed during training. For completeness, we also report an adapted LAIN baseline under the same DINOtxt-based setting. The performance gap remains consistent, with FineHOI outperforming LAIN across all metrics, confirming that our part-aware interaction representation provides advantages beyond the zero-shot regime.

4.3 Ablation Studies

Component analysis. We progressively enable each component of FineHOI to analyze its individual contributions (Tab. 3). The baseline (first row) corresponds to a setting where the proposed soft masks, Adaptive Part-Level Attention, and Region-Aware Interaction Transformer are replaced with standard RoIAlign features and a vanilla transformer. As components are added, both Unseen mAP and HM improve consistently, indicating improved zero-shot generalization. Notably, incorporating the soft-mask module (SM) yields a noticeable improvement of **+4.71** Unseen mAP and **+3.11** HM, while also improving Seen performance, indicating that patch-level grounding provides stronger spatial alignment than standard RoI features. Introducing Adaptive Part-Level Attention (APA) further boosts Unseen mAP by **+3.08** and HM by **+1.29**, with negligible impact on Seen performance. This suggests that part decomposition primarily enhances compositional generalization rather than fitting seen interactions. Finally, incorporating the Region-Aware Interaction Transformer ($\mathcal{I}$) leads to the strongest overall performance, improving Full mAP by **+2.74**, Seen mAP by **+2.96**, and Unseen mAP by **+1.37**, highlighting the importance of integrating localized interaction cues with global context. Overall, the three components are complementary: soft masks improve spatial grounding, Adaptive Part-Level Attention strengthens unseen generalization, and the Region-Aware Interaction Transformer provides contextual refinement. We conduct these ablations on the UV setting, which directly evaluates generalization to novel verbs, following prior work [18]. A similar trend is observed across zero-shot settings: UO, UC-RF, UC-NF, and UC. Corresponding results are provided in Supp. Mat..

Soft-mask feature extraction formulation. Tab. 4 evaluates the impact of the proposed soft-mask by comparing it with alternative feature pooling strategies. We consider three variants: (1) RoIAlign [12], (2) uniform average pooling over patch features, and (3) the proposed semantically weighted soft-mask (Sec. 3.1). Replacing RoIAlign with a uniform mask already improves zero-shot generalization, yielding a gain of **+2.67%** Unseen mAP and **+1.01%** HM. However, this comes at the cost of a **-0.66%** drop in Seen performance, indicating that uniform aggregation alone is insufficient to preserve the most informative features. In contrast, the proposed soft-mask achieves the best performance across all metrics, improving over RoIAlign by **+2.49%** Full, **+2.17%** Seen, **+4.45%** Unseen, and **+3.31%** HM. This demonstrates that semantically weighting patch features is crucial for interaction modeling, as it emphasizes interaction-relevant regions while suppressing less informative areas. Overall, these results highlight a clear progression: while coarse spatial selection (uniform) already benefits generalization, incorporating semantic weighting is essential to fully exploit fine-grained interaction cues.

Table 5: Effectiveness of multimodal embeddings on HICO-DET under the UV setting. Best results in bold.

	Full	Seen	Unseen	HM
HO token ($\tilde{\mathbf{h}}_{ho}$)	35.49	36.21	31.07	33.64
+ global ($\bar{\mathbf{e}}_g$)	36.18	37.05	30.79	33.92
+ global ($\bar{\mathbf{e}}_g$) + mask-guided ($\bar{\mathbf{e}}_m$)	**38.10**	**38.51**	**35.58**	**37.04**

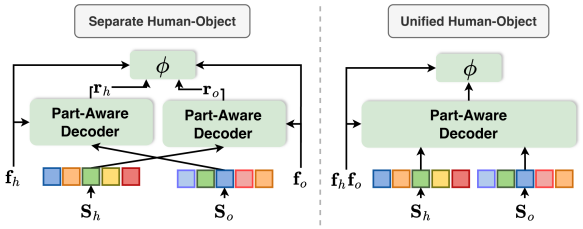


Figure 3: Visualization of the part-level attention module with separate or unified human and object parts decoder.

Table 6: Comparison between unified and separate part-level attention on the HICO-DET dataset under the UV setting.

	Full	Seen	Unseen	HM
Unified human/object	36.36	37.01	32.34	34.66
Separate human/object	**38.10**	**38.51**	**35.58**	**37.04**

On the effectiveness of masked embeddings. We analyze the contribution of the visual components used to construct the final multimodal embedding within the Region-Aware Interaction Transformer (Sec. 3.3). As shown in Tab. 5, we consider three configurations: (a) using only the human-object token ($\tilde{\mathbf{h}}_{ho}$), (b) augmenting it with the global branch ($\bar{\mathbf{e}}_g$), and (c) the full model, with the global ($\bar{\mathbf{e}}_g$) and mask-guided ($\bar{\mathbf{e}}_m$) branches. Adding the global branch improves Full and Seen performance by **+0.69%** and **+0.84%**, respectively, but slightly reduces Unseen mAP by **-0.28%**. This suggests that global context provides useful scene-level cues, but may also introduce irrelevant information that hinders generalization to unseen interactions. In contrast, incorporating the mask-guided branch (on top of the global branch) yields substantial improvements across all metrics, boosting Full by **+1.92%**, Seen by **+1.46%**, Unseen by **+4.79%**, and HM by **+3.12%**. This highlights the importance of localized, interaction-aware features for zero-shot HOI recognition. Overall, the two branches are complementary: the global branch captures coarse contextual information, while the mask-guided branch focuses on interaction-relevant regions, and their combination enables robust and fine-grained interaction modeling.

Separate versus unified part-level attention. We compare our proposed part-level attention (Sec. 3.2) with an alternative "unified" approach where the human $\mathbf{f}_h$ and object $\mathbf{f}_o$ features are concatenated into a single query sequence (Fig. 3, right), and the human $\mathbf{S}_h$ and object $\mathbf{S}_o$ sub-parts are concatenated into a single key/value sequence. As shown in Tab. 6, modeling the human and object separately yields consistently better results, improving Full mAP by **+1.74%**, Seen mAP by **+1.50%**, Unseen mAP by **+3.24%**, and HM by **+2.38%**. The gains are especially pronounced on the Unseen split, with an improvement of **+3.24%**, which is more than twice

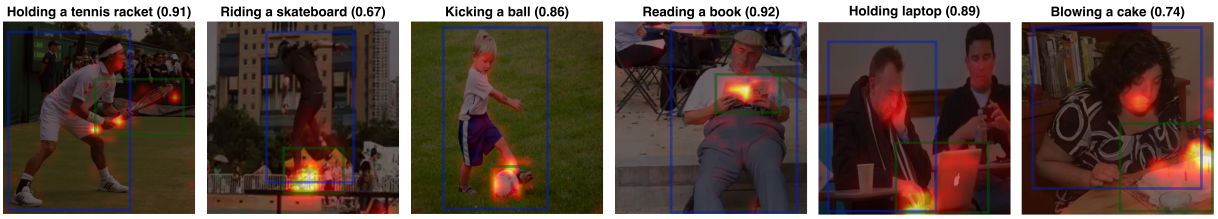


Figure 4: Attention maps of the interaction transformer $\mathcal{I}$ in FineHOI. Warmer colors indicate higher attention to visual patches. Blue and green boxes denote the human and object detections, respectively.

Table 7: Performance of FineHOI and LAIN with DINOtxt and detector features (Δ) on HICO-DET under the UO setting.

	Δ	Full	Seen	Unseen	HM
LAIN [18]	✓	36.01	34.72	38.11	36.41
		34.85	34.38	37.18	35.78
FineHOI (Ours)	✓	**36.84**	**35.93**	**41.42**	**38.67**
		36.62	35.92	40.13	38.03

Table 8: Comparison between our proposed soft-mask and SAM2 on the HICO-DET dataset under all zero-shot settings.

		Full	Seen	Unseen	HM
UV	SAM2	37.57	38.20	33.68	35.94
	Soft-mask	**38.10**	**38.51**	**35.58**	**37.04**
UO	SAM2	36.61	**36.38**	37.70	37.04
	Soft-mask	**36.62**	35.92	**40.13**	**38.03**
UC-RF	SAM2	37.90	37.95	37.69	37.82
	Soft-mask	**38.39**	**38.54**	**37.79**	**38.17**
UC-NF	SAM2	36.12	35.93	36.84	36.39
	Soft-mask	**36.58**	**36.20**	**38.11**	**37.15**
UC	SAM2	35.13	35.25	34.25	34.54
	Soft-mask	**36.88**	**37.41**	**34.79**	**36.10**

the gain observed on Seen (**+1.50%**). This suggests that separate attention better supports generalization to novel interactions. We attribute this behavior to the asymmetric nature of HOIs: the human should attend to object parts, while the object should attend to human parts. In contrast, the unified formulation mixes the two roles into a single attention space, which weakens entity-specific conditioning and makes it harder to localize the interaction-relevant regions.

Effect of detector priors in the UO setting. We analyze the role of object detector features (Δ) in the UO setting, where generalization depends on recognizing object categories not observed during training. In this regime, detector features provide auxiliary cues due to pretraining on overlapping object classes. As shown in Tab. 7, incorporating Δ leads to moderate improvements in Unseen performance for both methods. However, FineHOI consistently outperforms LAIN both with and without Δ. Notably, even without Δ, FineHOI surpasses LAIN equipped with Δ by **+2.02%** Unseen mAP and **+1.62%** HM. These results indicate that, while Δ can provide complementary information, the proposed part-aware interaction representation already captures transferable interaction cues effectively, enabling stronger generalization to unseen object categories without relying on detector-centric features.

Comparison between soft masks and SAM2. We further compare our proposed soft-mask formulation (Sec. 3.1) against binary masks generated by SAM2 [29], a strong segmentation model, using DETR bounding boxes as prompts (see Supp. Mat. for additional details on the implementation of SAM2). Despite SAM2 providing high-quality binary masks, our soft-mask consistently performs better across all zero-shot settings. In particular, it improves Full mAP in UV, UC-RF, UC-NF, UC, and yields especially large gains on the Unseen split (UO **+2.43%**, UV **+1.90%**, UC-NF **+1.27%**). Moreover, soft-mask achieves a higher harmonic mean (HM) in all settings (UO **+0.99%**, UV **+1.10%**, and UC **+1.56%**), indicating a better balance between Seen and Unseen performance. We argue that, in our setting, binary masks (like the ones generated by SAM2) impose an overly strict separation between foreground and background. In contrast, soft masks preserve evidence around boundaries and contact regions, which proves beneficial for modeling fine-grained interactions.

4.4 Qualitative Results

We present the attention maps of the interaction transformer $\mathcal{I}$ in Fig. 4 to analyze how FineHOI localizes interaction cues. Across diverse HOI categories, the attention consistently focuses on localized regions that are directly relevant to the predicted interaction, such as hand-racket contact area in ⟨*human*, *holding*, *tennis racket*⟩, ball-foot in ⟨*human*, *kicking*, *ball*⟩, hands in ⟨*human*, *holding*, *laptop*⟩, or mouth in ⟨*human*, *blowing*, *cake*⟩. These visualizations indicate that FineHOI effectively identifies interaction-critical regions, rather than relying on coarse object-level features. This behavior supports our design choice of combining soft-mask grounding and part-level attention, which enables fine-grained reasoning over HOIs.

5 Conclusion

We present FineHOI, a zero-shot HOI detection framework that models interactions from dense patch-level representations. Moving beyond detector-centric features, our approach captures fine-grained spatial relationships through soft-mask grounding, adaptive part-level attention, and a region-aware interaction transformer, enabling the model to focus on interaction-critical regions while preserving contextual information. Extensive experiments show that FineHOI achieves state-of-the-art performance across standard zero-shot settings, with particularly strong results on unseen interactions. Ablation studies further demonstrate that each component contributes to improved generalization, with the largest gains stemming from semantically grounded patch representations and part-aware reasoning. Future work will explore richer interaction reasoning in the zero-shot setting, including training-free approaches that leverage broader semantic context beyond pairwise human-object regions. We also plan to extend the framework to video-based HOI understanding.

Acknowledgements

We thank the EuroHPC initiative for the availability of high performance computing resources. This work was supported by the EU Horizon projects ELIAS (No. 101120237), ELLIOT (No. 101214398), TURING (No. 101215032), IAMI (No. 101168272), and by the Autonomous Province of Trento under Provincial Law No. 6/2023, as part of the VisRec project (CUP C69J25000880001). FT acknowledges travel support from ELIAS (GA No. 101120237).

References

[1] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. 2020. Detecting Human-Object Interactions via Functional Generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 34. 10460–10469.

[2] Ping Cao, Chunjie Zhang, Xiaolong Zheng, and Yao Zhao. 2026. Semantic-Spatial Guided Reasoning for Human-Object Interaction Detection. *IEEE Signal Processing Letters* 33 (2026), 551–555.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. In *European Conf. Comput. Vis. (ECCV)*, Vol. 12346. Springer, Glasgow, UK, 213–229.

[4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. 2018. Learning to Detect Human-Object Interactions. In *IEEE Winter Conf. Appl. Comp. Vis. (WACV)*. IEEE Computer Society, Lake Tahoe, NV, USA, 381–389.

[5] Yuan Chen, Zi-han Ding, Ziqin Wang, Yan Wang, Lijun Zhang, and Si Liu. 2024. Asynchronous large language model enhanced planner for autonomous driving. In *European Conf. Comput. Vis. (ECCV)*, Vol. 15094. Springer, Milan, Italy, 22–38.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[7] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. 2020. DRG: Dual Relation Graph for Human-Object Interaction Detection. In *European Conf. Comput. Vis. (ECCV)*, Vol. 12357. Springer, Glasgow, UK, 696–712.

[8] Pei Geng, Jian Yang, and Shanshan Zhang. 2025. Horp: Human-object relation priors guided hoi detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 25325–25335.

[9] Yixin Guo, Yu Liu, Jianghao Li, Weimin Wang, and Qi Jia. 2024. Unseen no more: Unlocking the potential of clip for generative zero-shot hoi detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 1711–1720.

[10] Saurabh Gupta and Jitendra Malik. 2015. Visual Semantic Role Labeling. *CoRR* abs/1505.04474 (2015), 1–11.

[11] Tanmay Gupta, Alexander G. Schwing, and Derek Hoiem. 2019. No-Frills Human-Object Interaction Detection: Factorization, Layout Encodings, and Training Techniques. In *IEEE Int. Conf. Comput. Vis. (ICCV)*. 9676–9684.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[13] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. *Advances in neural information processing systems* 32 (2019).

[14] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. 2020. Visual Compositional Learning for Human-Object Interaction Detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 584–600.

[15] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. 2021. Detecting human-object interaction via fabricated compositional learning. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*. Computer Vision Foundation / IEEE, Virtual, 14646–14655.

[16] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. 2021. HOTR: End-to-End Human-Object Interaction Detection With Transformers. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*. Computer Vision Foundation / IEEE, Virtual, 74–83.

[17] Bumsoo Kim, Jonghwan Mun, Kyoung-Woon On, Minchul Shin, Junhyun Lee, and Eun-Sol Kim. 2022. MSTR: Multi-Scale Transformer for End-to-End Human-Object Interaction Detection. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*. IEEE, New Orleans, LA, USA, 19556–19565.

[18] Sanghyun Kim, Deunsol Jung, and Minsu Cho. 2025. Locality-aware zero-shot human-object interaction detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 20190–20200.

[19] Qinqian Lei, Bo Wang, and Robby Tan. 2024. Ez-hoi: Vlm adaptation via guided prompt learning for zero-shot hoi detection. *Advances in Neural Information Processing Systems* 37 (2024), 55831–55857.

[20] Ting Lei, Fabian Caba, Qingchao Chen, Hailin Jin, Yuxin Peng, and Yang Liu. 2023. Efficient Adaptive Human-Object Interaction Detection with Concept-guided Memory. In *IEEE Int. Conf. Comput. Vis. (ICCV)*. IEEE, Paris, France, 6457–6467.

[21] Ting Lei, Shaofeng Yin, Yuxin Peng, and Yang Liu. 2024. Exploring conditional multi-modal prompts for zero-shot hoi detection. In *European Conference on Computer Vision*. Springer, 1–19.

[22] Huadong Li, Ying Wei, Shuailei Ma, Mingyu Chen, and Ge Li. 2024. Ripple Transformer: A Human-Object Interaction Backbone and a New Prediction Strategy for Smart Surveillance Devices. *IEEE Trans. Consumer Electron.* 70, 1 (2024), 2257–2268.

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[25] Ye Liu, Junsong Yuan, and Chang Wen Chen. 2020. ConsNet: Learning Consistency Graph for Zero-Shot Human-Object Interaction Detection. In *ACM Multimedia (ACMMM)*. ACM, Seattle, WA, USA, 4235–4243.

[26] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[27] Yunyao Mao, Jiajun Deng, Wengang Zhou, Li Li, Yao Fang, and Houqiang Li. 2023. Clip4hoi: towards adapting clip for practical zero-shot hoi detection. *Advances in Neural Information Processing Systems* 36 (2023), 45895–45906.

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

[29] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024).

[30] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. 2022. End-to-End Generative Pretraining for Multimodal Video Captioning. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*. IEEE, New Orleans, LA, USA, 17959–17968.

[31] Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence* 22, 8 (2000), 888–905.

[32] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. 2025. Dinov3. *arXiv preprint arXiv:2508.10104* (2025).

[33] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. 2021. QPIC: Query-Based Pairwise Human-Object Interaction Detection With Image-Wide Contextual Information. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*. Computer Vision Foundation / IEEE, Virtual, 10410–10419.

[34] Francesco Tonini, Lorenzo Vaquero, Alessandro Conti, Cigdem Beyan, and Elisa Ricci. 2025. Dynamic Scoring with Enhanced Semantics for Training-Free Human-Object Interaction Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 2801–2810.

[35] Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. 2022. Iwin: Human-Object Interaction Detection via Transformer with Irregular Windows. In *European Conf. Comput. Vis. (ECCV)*. Springer, Tel Aviv, Israel, 87–103.

[36] Oytun Ulutan, A. S. M. Iftekhar, and B. S. Manjunath. 2020. VSGNet: Spatial Attention Network for Detecting Human Object Interactions Using Graph Convolutions. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*. Computer Vision Foundation / IEEE, Seattle, WA, USA, 13614–13623.

[37] Chanhyeong Yang, Taehoon Song, Jihwan Park, and Hyunwoo J Kim. 2025. Visual Diversity and Region-aware Prompt Learning for Zero-shot HOI Detection. *arXiv preprint arXiv:2510.25094* (2025).

[38] Huzheng Yang, James Gee, and Jianbo Shi. 2024. AlignedCut: Visual Concepts Discovery on Brain-Guided Universal Feature Space. arXiv:2406.18344 [cs.CV]

[39] Xu Yang, Hanwang Zhang, and Jianfei Cai. 2023. Deconfounded Image Captioning: A Causal Retrospect. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 11 (2023), 12996–13010.

[40] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. 2021. Mining the Benefits of Two-stage and One-stage HOI Detection. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, Vol. 34. Curran Associates, Inc., Virtual, 17209–17220.

[41] Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. 2021. Spatially Conditioned Graphs for Detecting Human-Object Interactions. In *IEEE Int. Conf. Comput. Vis. (ICCV)*. IEEE, Montreal, QC, Canada, 13299–13307.

[42] Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. 2022. Efficient Two-Stage Detection of Human-Object Interactions with a Novel Unary-Pairwise Transformer. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*. IEEE, New Orleans, LA, USA, 20072–20080.

[43] Frederic Z Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. 2023. Exploring predicate visual context in detecting of human-object interactions. In *IEEE Int. Conf. Comput. Vis. (ICCV)*. IEEE, Paris, France, 10411–10421.

[44] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. 2022. 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*. 19548–19557.

[45] Xubin Zhong, Changxing Ding, Zijian Li, and Shaoli Huang. 2022. Towards Hard-Positive Query Mining for DETR-Based Human-Object Interaction Detection. In *European Conf. Comput. Vis. (ECCV)*. Springer, Tel Aviv, Israel, 444–460.