

Contextualized Representation Learning for Effective Human-Object Interaction Detection

Zhehao Li, Yucheng Qian, Chong Wang , Yinghao Lu, Zhihao Yang and Jiafei Wu

Abstract—Human-Object Interaction (HOI) detection aims to simultaneously localize human-object pairs and recognize their interactions. While recent two-stage approaches have made significant progress, they still face challenges due to incomplete context modeling. In this work, we introduce a Contextualized Representation Learning that integrates both affordance-guided reasoning and contextual prompts with visual cues to better capture complex interactions. We enhance the conventional HOI detection framework by expanding it beyond simple human-object pairs to include multivariate relationships involving auxiliary entities like tools. Specifically, we explicitly model the functional role (affordance) of these auxiliary objects through triplet structures $\langle\text{human}, \text{tool}, \text{object}\rangle$. This enables our model to identify tool-dependent interactions such as “filling”. Furthermore, the learnable prompt is enriched with instance categories and subsequently integrated with contextual visual features using an attention mechanism. This process aligns language with image content at both global and regional levels. These contextualized representations equip the model with enriched relational cues for more reliable reasoning over complex, context-dependent interactions. Our proposed method demonstrates superior performance on both the HICO-Det and V-COCO datasets in most scenarios. The source code is available at <https://github.com/lzzhhh1019/CRL>.

Index Terms—Human-Object Interaction Detection, Two-Stage Network, Prompt Learning, Attention Mechanism.

I. INTRODUCTION

HUMAN-OBJECT Interaction (HOI) detection is a fundamental and challenging task in visual recognition that aims to model complex semantic relationships between humans and surrounding objects in a visual scene, ultimately recognizing triplets of the form $\langle\text{human}, \text{action}, \text{object}\rangle$. It requires a simultaneous understanding of visual features, spatial configurations and contextual semantics to effectively identify all valid interaction patterns within given images. Therefore, other high-level semantic understanding tasks, such as activity recognition [1]–[3] and video comprehension [4]–[6] can benefit from HOI.

Manuscript received 3 October 2025; This work was supported by the Ningbo Municipal Natural Science Foundation of China (No. 2022J114), National Natural Science Foundation of China (No. 62271274), Ningbo S&T Project (No.2024Z004) and Ningbo Major Research and Development Plan Project (No.2023Z225).

Zhehao Li, Chong Wang, Yinghao Lu, and Zhihao Yang are with the Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, Zhejiang 315211, China, E-mail: llzzhhh1019@163.com, wangchong@nbu.edu.cn.

Yucheng Qian is with Nanjing University. Email: qianycqq@163.com.

Jiafei Wu is with Zhejiang Lab, Hangzhou, China. E-mail: wufjiafei@zhejianglab.com.

✉ Corresponding Author: Chong Wang.

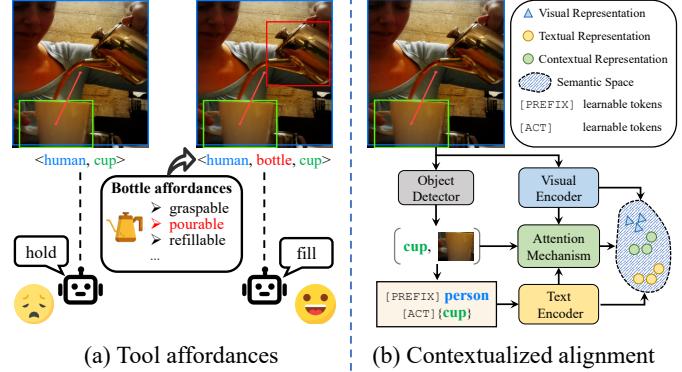


Fig. 1. Contextualized representations in HOI. (a) Tool affordances (e.g., bottle’s pourable) help distinguish complex interactions (e.g., human fill cup) from direct human-object relations (e.g., human hold cup). (b) Contextualized alignment with instance categories (e.g., cup) and corresponding visual features narrows down the potential actions to relevant ones (e.g., fill, hold).

Existing HOI detection approaches can generally be divided into one-stage and two-stage methods. One-stage methods perform object detection and interaction prediction simultaneously, but adapting their decoders for various HOIs often incurs substantial training costs, sometimes requiring hundreds of GPU hours. On the other hand, two-stage methods leverage pre-trained object detectors to identify humans and objects before explicitly establishing relationships between them for interaction prediction. Thanks to its flexibility, the two-stage approach has quickly become the preferred paradigm. However, traditional two-stage HOI approaches have primarily relied on features derived from human-object pairs, such as human pose [7] or spatial information [8]. In real-world interactions, many scenarios are not solely defined by these pairs. They may also involve auxiliary tools or objects that play a crucial role in the interaction. While prior HOI studies [9] have leveraged object affordances, such modeling usually remains tied to the acted-on object and does not explicitly account for tool affordances. In contrast, we model the functional role of tools within the $\langle\text{human}, \text{tool}, \text{object}\rangle$ ternary. For example, as illustrated in Fig. 1(a), when a person fills a cup using a bottle, understanding the bottle’s affordance (i.e., its potential function like being pourable) offers significant insights into deciphering the nature of the interaction.

Meanwhile, recent advancements in pre-trained language models like BERT [10] and GPT [11] offer rich semantic insights for understanding HOIs. Prompt Learning, as an efficient method for finetuning language models, has been extended to visual tasks by CoOp [12]. Many studies have

as integrated Prompt Learning into the HOI task [13]–[15]. However, these approaches rely exclusively on textual data and do not integrate image information from specific samples. As a result, crucial contextual visual cues are left untapped, which are essential for accurately capturing interactions.

It's worth noting that both the affordances and the text-related appearance of a specific object act as vital forms of contextualized representation, essential for identifying complex interactions. Despite the importance, their applications in HOI has not been thoroughly explored. This gap in research greatly motivates us to delve deeper into this area. In this work, we propose a new two-stage HOI detection framework, named Contextualized Representation Learning (CRL). To be specific, unlike traditional human-object pairs, multivariate relationships (unary, binary and ternary ones) are modeled to capture tool-mediated interactions through affordance-guided triplets $\langle \text{human}, \text{tool}, \text{object} \rangle$. In parallel, the contextualized representations of detected instances are injected into the process of prompt learning. Specifically, as illustrated in Fig. 1(b), we extract localized visual features from the corresponding instance regions and refine the prompt representation through a cross-attention mechanism with VLM's global image embedding. This design allows the prompt to be conditioned on both textual semantics and instance-level visual cues, improving its capacity to capture context-dependent interactions.

To sum up, our contributions are two-fold:

- We are the first to emphasize the role of tools in improving HOI understanding. Guided by the concept of affordances, we propose a multivariate relationship modeling framework that introduces $\langle \text{human}, \text{tool}, \text{object} \rangle$ triplets to capture complex tool-mediated interactions beyond conventional human-object pairs.
- We introduce a contextualized prompt learning module that incorporates detected object categories and their corresponding visual features into the prompt, enabling contextual alignment between textual and visual modalities at both semantic and instance levels.

II. RELATED WORK

A. HOI Detection

Existing Human-Object Interaction (HOI) detection approaches can be divided into one-stage and two-stage methods. One-stage methods [16]–[20] simultaneously predict human and object boxes, categories, and interaction classes. Recent advancements in one-stage detectors, particularly those utilizing transformer architectures [21]–[24], have shown promising performance. This end-to-end approach simplifies the inference process of HOI. However, it imposes heavy computational costs during training, which can limit its practicality.

To alleviate this issue, two-stage methods [25]–[30] undergone significant development recently. They typically rely on pre-trained detectors (e.g., DETR) to generate object proposals, focusing more on extracting interaction context from candidate human-object pairs. For example, UPT [8] first detects all humans and objects (i.e., unary features) and then applies self-attention to unary features and human-object pairs, effectively enhancing the confidence of positive samples. PViC

[31] incorporates image features back into the representation of human-object pairs via cross-attention to compensate for missing contextual information. Nonetheless, traditional methods often treat all objects equally, failing to account for the role that tools play in interactions. In this work, we explore the concept of tool affordances to more effectively capture these relational cues.

B. Vision-Language Models in HOI

To obtain more effective HOI representations, several studies [32] explore knowledge transfer from vision-language pre-trained models (e.g., CLIP [33]). This approach not only enriches the learned representations but also improves the model's capability in HOI recognition. GEN-VLKT [32] utilizes CLIP knowledge for interaction classification and the distillation of visual features. HOICLIP [34] uses the features obtained by the VLM visual encoder and proposes a new transfer strategy that uses visual semantic algorithms to represent action. ViPLO [16] adopts the Vision Transformer (ViT) from CLIP as its backbone and introduces a pose-conditioned graph to capture the local features of human joints. ADA-CM [35] develops a concept-guided memory mechanism to represent visual embeddings and semantic knowledge simultaneously. The success of VLMs opens up new avenues for HOI detection. CLIP4HOI [36] utilizes CLIP's vision-language knowledge to enhance human-object interaction detection by decoupling human and object detection and adapting CLIP into a fine-grained classifier for better interaction discrimination.

C. Prompt Learning in HOI

Prompt learning has become very popular for fine-tuning VLMs on downstream tasks. Context Optimization (CoOp) [12] encodes prompt context as learnable vectors, achieving strong performance with only a few labeled samples. To better adapt to downstream tasks, Conditional Context Optimization (CoCoOp) [37] builds upon CoOp by employing a lightweight network to generate input-conditional context tokens for each image. THID [13] is the first to propose the use of learnable language prompts for the HOI detection task. CMMP [14] introduces decoupled multi-modal prompts for spatial-aware HOI detection, separating visual feature extraction and interaction classification to reduce error propagation. EZ-HOI [15] leverages LLM-generated class descriptions to guide prompt learning.

However, most existing prompt learning methods for HOI tasks neglect instance-specific visual details and fail to adapt contextually. To overcome these limitations, we integrate contextualized visual representations to create entity-aware prompts, thereby enhancing the model's understanding of visual content.

III. METHOD

A. Preliminary

Our proposed model builds upon the traditional two-stage methodology for Human-Object Interaction (HOI) detection, which first focuses on detecting instances and subsequently

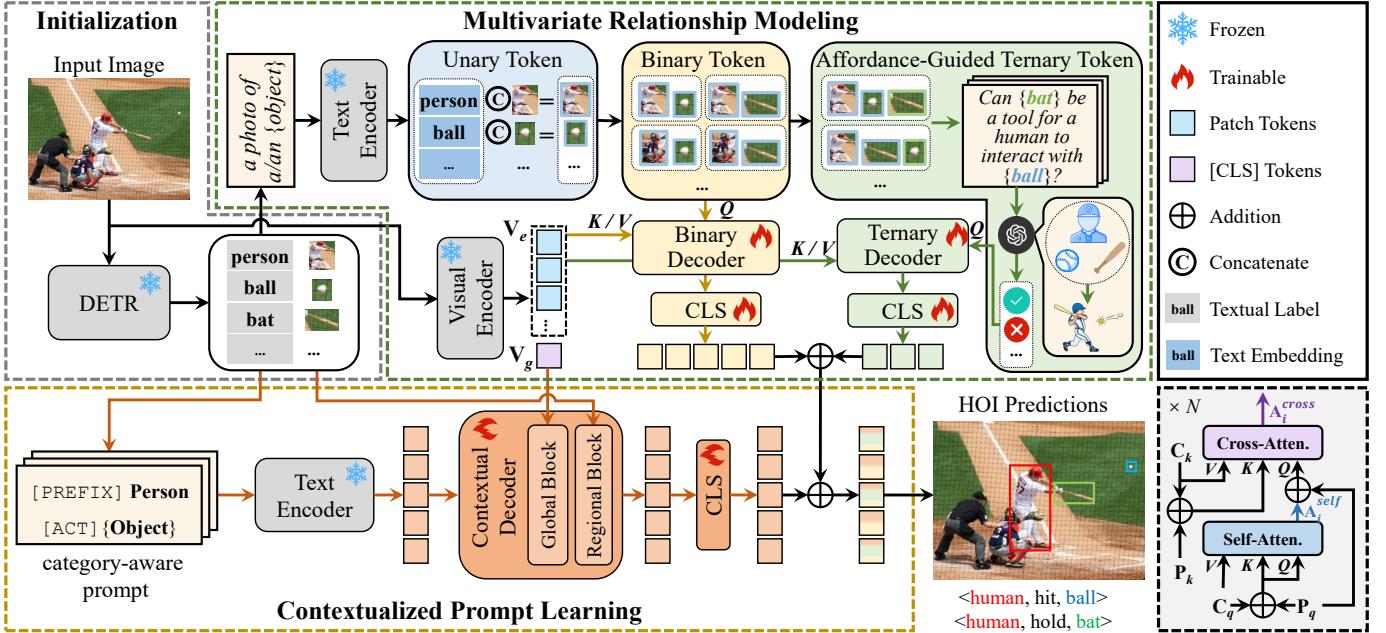


Fig. 2. Overall architecture of our Contextualized Representation Learning Network, consisting of Multivariate Relationship Modeling (MRM) and Contextualized Prompt Learning (CPL). MRM constructs unary, binary and ternary token sets from regional features to model HOIs. CPL builds a category-aware learnable prompt, fused with diverse contextual visual features. Their combined outputs are utilized for interaction prediction. The structure of the binary/ternary/contextual decoder is shown in the bottom right.

classifies the interactions between them. In the first stage, as shown in the upper-left part of Fig. 2, we employ an off-the-shelf object detector, such as DETR [38], to extract all instances, including humans and objects. As a result, a set of instances $\mathcal{Z} = \{z_i\}_{i=1}^n$ can be constructed with each detection $z_i = (\mathbf{b}_i, s_i, c_i, \mathbf{u}_i)$, consisting of the box coordinates $\mathbf{b}_i \in \mathbb{R}^4$, the confidence score $s_i \in [0, 1]$, the detected instance $c_i \in \mathcal{O}$ for the category set \mathcal{O} , and the unary instance feature $\mathbf{u}_i \in \mathbb{R}^C$. C is the embedding dimension, n is the number of detected instances.

In the second stage, the visual and semantic features of the detected instance set \mathcal{Z} are utilized to exploit the interaction information. In this work, as shown in Fig. 2, we focus on modeling relationships at three distinct levels: unary (involving individual entities), binary (encompassing human-object interactions) and ternary (covering human-tool-object dynamics) ones. This enables affordance-guided interaction modeling, which allows the model to effectively capture the progressively complex patterns inherent in HOIs. Moreover, we further extract the contextualized representation from \mathcal{Z} to improve prompt learning. Specifically, a category-aware prompt is fused with contextual unary features and CLIP's global visual features via an attention mechanism.

B. Multivariate Relationship Modeling

Unary Association. To establish the multivariate relationships among the detected instances in \mathcal{Z} , we start with the most straightforward scenario, i.e., examining the visual and textual representation of each instance (whether human or object) individually. In this unary association, its visual feature is denoted as a unary token \mathbf{u}_i , forming the set $\mathcal{U} = \{\mathbf{u}_i\}_{i=1}^n$. Meanwhile, the corresponding semantic feature \mathbf{e}_i for each

instance is obtained through the text encoder of a vision-language model (VLM). This process employs the prompt template “*a photo of a/an {object}*”, where $\{object\}$ is the instance category c_i detected in the first stage. Then, such category-level semantic information is combined with visual features to obtain an enhanced unary token \mathbf{u}'_i as,

$$\mathbf{u}'_i = \text{MLP}(\text{Concat}(\mathbf{u}_i, \mathbf{e}_i)), \quad (1)$$

where $\text{MLP}(*)$ is a multi-layer perceptron used to project the concatenated features into the target embedding space.

Binary Connection. Based on the unary association, we continue to investigate the binary connection between different instances. Inspired by UPT [8], a binary token set $\mathcal{G} = \{\mathbf{g}_l\}_{l=1}^m$ can be constructed from the unary one \mathcal{U} as follows,

$$\begin{aligned} \mathbf{g}_l &= \text{MLP}(\text{Concat}(\mathbf{u}'_i, \mathbf{u}'_j)), \\ \text{s.t. } &\left\{ \begin{array}{l} i \neq j, \\ c_i = \text{"human"}. \end{array} \right. \end{aligned} \quad (2)$$

As the fusion of human token \mathbf{u}'_i and object token \mathbf{u}'_j , $\mathbf{g}_l \in \mathbb{R}^D$ is fed into a binary decoder for interaction classification.

Following the setup in DETR [38], the structure of our binary decoder is presented in the right lower corner of Fig. 2. It includes N blocks, each with cascaded self- and cross-attention layers, designed to process four input, namely content query \mathbf{C}_q , positional query \mathbf{P}_q , content key \mathbf{C}_k , and positional key \mathbf{P}_k . The attention layers follow the standard process $\text{Atten}(*)$ as,

$$\text{Atten}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Norm} \left(\sigma \left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} \right) \mathbf{V} + \mathbf{V} \right), \quad (3)$$

where \mathbf{Q} , \mathbf{K} , \mathbf{V} are the query, key and value tokens. $\text{Norm}(\cdot)$ and $\sigma(\cdot)$ denotes the layer normalization and softmax function, respectively.

All tokens $\mathbf{g}_l \in \mathbb{R}^D$ are concatenated into a matrix $\mathbf{G}_0 \in \mathbb{R}^{m \times D}$, which is used as the content query \mathbf{C}_q . The corresponding positional matrix $\mathbf{X} \in \mathbb{R}^{m \times D}$ is used as the positional query \mathbf{P}_q . It is extracted from all bounding box pair $(\mathbf{b}_i, \mathbf{b}_j)$, by combining various attributes. D is the embedding dimension and m is the number of human-object pairs. Then, the output of self-attention process $\mathbf{A}_i^{b\text{-self}}$ in the i -th binary decoder block can be expressed as follows,

$$\mathbf{A}_i^{b\text{-self}} = \text{Atten}(\mathbf{G}_{i-1} + \mathbf{X}, \mathbf{G}_{i-1} + \mathbf{X}, \mathbf{G}_{i-1}), \quad (4)$$

where $i = 1, 2, \dots, N$, \mathbf{G}_{i-1} is the outputs of the previous decoder block.

To enhance interaction modeling, we further employ the visual encoder $\text{VisEnc}(\cdot)$ of VLMs to extract spatial features $\mathbf{V}_e \in \mathbb{R}^{H' \times W' \times D}$ from the whole image, accompanying by the position embedding $\mathbf{S} \in \mathbb{R}^{H' \times W' \times D}$. Here, \mathbf{V}_e serves as content keys and values, while \mathbf{S} acts as positional keys. Thus, the output of cross-attention layer $\mathbf{A}_i^{b\text{-cross}}$ in the i -th block can be formulated as,

$$\mathbf{A}_i^{b\text{-cross}} = \text{Atten}\left(\mathbf{A}_i^{b\text{-self}} + \mathbf{X}, \mathbf{V}_e + \mathbf{S}, \mathbf{V}_e\right). \quad (5)$$

The final output \mathbf{G}_i of i -th block is obtained as,

$$\mathbf{G}_i = \text{Norm}(\mathbf{A}_i^{b\text{-cross}} + \text{FFN}(\mathbf{A}_i^{b\text{-cross}})), \quad (6)$$

where $\text{FFN}(\cdot)$ denotes a feed-forward network. The output of the last block \mathbf{G}_N is then fed into a linear layer to predict the classification logits $\tilde{\mathbf{y}} \in \mathbb{R}^{m \times c}$ for human-object pairs' interactions, where c is the number of action categories.

Ternary Relationship. Beyond the binary connections, we further explore the ternary relationships within \mathcal{Z} by incorporating the concept of object affordances into the HOI task. To determine which objects can serve as functional tools in HOIs, we query a large language model (LLM) using a prompt of “*Can {X} be a tool for a human to interact with {Y}?*”, where X and Y represent different object categories (e.g., bat and ball). Although the prompt does not explicitly mention affordances, the LLM exhibits a strong implicit understanding of object functionality and affordance-related reasoning. Based on LLM's responses, we construct a set of object-tool pairs, which are stored in the knowledge bank \mathcal{B} as ordered pairs of the form $\langle Y, X \rangle$. To improve efficiency, these identified object-tool pairs are curated offline and used as external knowledge during training and inference.

Similar to the binary token set, a ternary token set $\mathcal{T} = \{\mathbf{t}_o\}_{o=1}^r$ of size r is constructed as follows,

$$\begin{aligned} \mathbf{t}_o &= \text{MLP}(\text{Concat}(\mathbf{u}'_i, \mathbf{u}'_j, \mathbf{u}'_k)), \\ \text{s.t. } &\begin{cases} i \neq j, j \neq k, i \neq k, \\ c_i = \text{"human"}, \\ \langle c_j, c_k \rangle \in \mathcal{B}. \end{cases} \end{aligned} \quad (7)$$

A ternary decoder, designed with a structure similar to the binary decoder but having different coefficients, is developed to leverage affordance-guided interaction from \mathcal{T} . Noting that each ternary token $\mathbf{t}_o \in \mathbb{R}^D$ is associated with a triplet

consisting of the human token \mathbf{u}'_i , object token \mathbf{u}'_j and tool token \mathbf{u}'_k . Based on their corresponding bounding box triplet $(\mathbf{b}_i, \mathbf{b}_j, \mathbf{b}_k)$, we compute the relationships between each pair of tokens, i.e., human-object, human-tool, and object-tool. These pairwise features are then concatenated and passed through an MLP to form the ternary positional matrix $\mathbf{W} \in \mathbb{R}^{r \times D}$, which serves as the positional query \mathbf{P}_q . Then, the ternary matrix $\mathbf{T}_0 \in \mathbb{R}^{r \times D}$, concatenated by all \mathbf{t}_o , is used as the content query \mathbf{C}_q . The output of self-attention $\mathbf{A}_i^{t\text{-self}}$ in i -th ternary decoder block is,

$$\mathbf{A}_i^{t\text{-self}} = \text{Atten}(\mathbf{T}_{i-1} + \mathbf{W}, \mathbf{T}_{i-1} + \mathbf{W}, \mathbf{T}_{i-1}), \quad (8)$$

where $i = 1, 2, \dots, N$, and \mathbf{T}_{i-1} represents the output from the previous decoder block.

We continue to use the VLM's spatial features \mathbf{V}_e and position embeddings \mathbf{S} as content keys and positional keys. Thus, ternary cross-attention $\mathbf{A}_i^{t\text{-cross}}$ in i -th block is,

$$\mathbf{A}_i^{t\text{-cross}} = \text{Atten}\left(\mathbf{A}_i^{t\text{-self}} + \mathbf{W}, \mathbf{V}_e + \mathbf{S}, \mathbf{V}_e\right). \quad (9)$$

The final output \mathbf{T}_i of i -th ternary decoder block is given as,

$$\mathbf{T}_i = \text{Norm}(\mathbf{A}_i^{t\text{-cross}} + \text{FFN}(\mathbf{A}_i^{t\text{-cross}})). \quad (10)$$

Similarly, \mathbf{T}_N from the final block of ternary decoder is then passed through a classifier to generate the affordance-guided prediction logits $\mathbf{y}' \in \mathbb{R}^{r \times c}$.

Typically, the number of human-object pairs, denoted as m , does not match the number of human-tool-object triplets, represented by r . As a result, the logits $\tilde{\mathbf{y}}$ and \mathbf{y}' might have different dimensions. It is necessary to effectively fuse them together, ensuring that the disparate information from binary and ternary sets can be integrated in a meaningful way. For the l -th human-object pair, whenever the triplet for \mathbf{t}_o includes this specific pair, the ternary logits \mathbf{y}'_o shall contribute to classifying that interaction. Considering that different tools might be involved in such interactions, we define the subset of ternary tokens associated with the same l -th human-object pair as \mathcal{K}_l . Thus, the refined interaction logits $\hat{\mathbf{y}}$ are defined as follows,

$$\hat{\mathbf{y}}_l = \tilde{\mathbf{y}}_l + \alpha \cdot \sum_{\mathbf{t}_o \in \mathcal{K}_l} \mathbf{y}'_o, \quad (11)$$

where α is the weighting parameter.

C. Contextualized Prompt Learning

To fulfill the missing piece of corresponding visual information in conventional prompt learning, we propose Contextualized Prompt Learning (CPL) to incorporate specific regional features $\mathbf{D} \in \mathbb{R}^{m \times C'}$. As shown in the lower part of Fig. 2, these contextual representations are integrated with the global visual context \mathbf{V}_g obtained by projecting \mathbf{V}_e through a projection layer, in order to enhance the effect of learnable text tokens for better interaction prediction.

The contextual features $\mathbf{D} \in \mathbb{R}^{m \times C'}$ is the concatenation of all $\mathbf{d}_l \in \mathbb{R}^{C'}$, using pure visual features of candidate human-object pairs as follows,

$$\mathbf{d}_l = \text{MLP}(\text{Concat}(\mathbf{u}_i, \mathbf{u}_j)). \quad (12)$$

The corresponding object labels (i.e., c_j) are used to construct a prompt in the form of “[PREFIX] person [ACT] {object}”, where [PREFIX] and [ACT] can be a sequence of learnable tokens, i.e., $[v]_1[v]_2\dots[v]_A$. Each $[v]_a$ ($a \in \{1, \dots, A\}$) is a vector with the same dimension as word embeddings (i.e., 512 or 768 for CLIP), and A is a hyperparameter specifying the number of learnable tokens. Then, they are passed through the VLM’s text encoder and a projection layer to produce the textual features $M_0 \in \mathbb{R}^{m \times C'}$.

A new contextual decoder, whose block shares the same structure as the binary and ternary ones, is proposed to fuse regional and global contextual visual information (i.e., D and V_g) with adaptive textual features M_0 for interaction reasoning. Specifically, two decoder blocks (global and regional), i.e., $N = 2$, are designed to sequentially inject V_g and D at the cross-attention layer. Meanwhile, the positional queries or keys are set to 0, since the location of instances is less relevant to the prompt. Given M_0 as the input, the global visual context V_g is utilized to form the context key and value at the first cross-attention layer. It is repeated m times to obtain $V'_g \in \mathbb{R}^{m \times C'}$, aligned with the input’s dimensions. Thus, the output M_1 of the first block can be formulated as,

$$A_1^{u-cross} = \text{Atten}(\text{Atten}(M_0, M_0, M_0), V'_g, V'_g), \quad (13)$$

$$M_1 = \text{Norm}(A_1^{u-cross} + \text{FFN}(A_1^{u-cross})). \quad (14)$$

The regional features D are applied to further refine the contextualized representation at the second block as,

$$A_2^{u-cross} = \text{Atten}(\text{Atten}(M_1, M_1, M_1), D, D). \quad (15)$$

The final output M_2 of the contextual decoder is given as,

$$M_2 = \text{Norm}(A_2^{u-cross} + \text{FFN}(A_2^{u-cross})). \quad (16)$$

By integrating the overall scene context and fine-grained instance details, feature representation becomes more comprehensive than methods that rely solely on text-based prompts, which helps to understand complex interactions.

Finally, we train a classifier for M_2 to output the semantic interaction logits $\hat{y} \in \mathbb{R}^{m \times c}$. It is then integrated with the previously refined interaction logits \hat{y} as,

$$\hat{y}'_l = \hat{y}_l + \beta \cdot \hat{y}_l, \quad (17)$$

where β is the weighting parameter.

D. Training and Inference

Training. During training, the Focal Loss (FL) is used on the predicted action logits as follows,

$$\mathcal{L} = \frac{1}{\sum_{i=1}^m \sum_{j=1}^c y_{i,j}} \sum_{i=1}^m \sum_{j=1}^c \text{FL}(\hat{y}'_{i,j}, y_{i,j}), \quad (18)$$

where c is the number of action classes, $y_{i,j} \in \{0, 1\}$ indicates whether the ground truth of the i -th human-object pair contains the j -th action class.

Inference. To make full use of the pre-trained object detector, we incorporate the object confidence scores into the final scores of each human-object pair as,

$$s = (s_h s_o)^\lambda \cdot \delta(\hat{y}'), \quad (19)$$

where hyperparameter $\lambda > 1$, $\delta(*)$ is the sigmoid function.

IV. EXPERIMENTS

A. Experiment Setup

Dataset. We evaluate our model on two widely used benchmarks, HICO-DET [45] and V-COCO [46]. HICO-DET consists of 37,633 training and 9,546 test images, covering 600 HOI categories derived from 80 object and 117 action classes, split into 138 rare and 462 non-rare categories. V-COCO is derived from COCO [47] and contains 10,326 images (5,400 for training, 4,964 for testing) with annotations for 80 object and 24 action categories.

Evaluation Metrics. The mean average precision (mAP) is used to evaluate performance. A predicted HOI triplet is considered a true positive if it satisfies two conditions: 1) the Intersection over Union (IoU) between the predicted and ground-truth bounding boxes for both the human and the object exceeds 0.5; 2) the predicted action and object categories match the ground-truth labels.

Zero-shot Setting. Following prior works [15], [34], we conduct our zero-shot experiments under four distinct configurations: Rare First Unseen Combination (RF-UC), Non-rare First Unseen Combination (NF-UC), Unseen Verb (UV), and Unseen Object (UO). In the RF-UC setting, we select tail HOI categories as unseen categories, while in the NF-UC setting, we use head HOI categories as unseen categories. Under the UV and UO settings, some verb or object categories are not included in the training set, respectively.

Implementation Details. The pre-trained DETR model with a ResNet50 [48] backbone is selected as our object detector. AdamW [49] is used as the optimizer, with both the learning rate and weight decay being 10^{-4} . We set λ to 1 during training and 2.8 during inference. Unless otherwise specified, all models are trained for 15 epochs, with a learning rate drop by a factor of 5 at the 10th epoch. The visual encoder is based on ViT-B/16 and ViT-L/14 CLIP, and during training, the parameters of CLIP remain frozen. In the MRM module, we use two blocks for both the binary and ternary decoders and set the weighting coefficient α to 1. [PREFIX] token is manually designed, with [ACT] token length set to 4. All experiments are conducted on 8 NVIDIA 4090 GPUs and the batch size is 16. The computational environment runs Ubuntu 22.04, with Python version 3.7, PyTorch version 1.10.0, torchvision version 0.11.0, and CUDA version 11.3.

B. Comparison to the State-of-The-Art

The experimental results on the HICO-Det and V-COCO datasets are presented in Table I. We compared with both one-stage methods, such as CATN [39], ERNet [40], SGHOI+ [41], etc., and two-stage methods, like CMMP [14], EZ-HOI [15], LAIN [43], HOLa [44]. For the HICO-DET dataset, our proposed model demonstrates remarkable performance, outperforming the most recent work HOLa by a margin of **0.94** mAP for full categories. Notably, the performance improvement is most significant in the rare categories, where our model outperforms CMMP and EZ-HOI by margins of **2.92** mAP and **2.97** mAP, respectively. For the V-COCO dataset, our model achieves 60.9 and 66.9 role AP in S1 and S2,

TABLE I

COMPARISON WITH STATE-OF-THE-ART METHODS ON HICO-DET AND V-COCO. **BOLD** AND UNDERLINE ITEMS INDICATE THE BEST AND SECOND-BEST RESULTS, RESPECTIVELY. ONE-STAGE AND TWO-STAGE METHODS ARE HIGHLIGHTED SEPARATELY.

Method	Backbone	Default (mAP % ↑)			Known Object (mAP % ↑)			V-COCO (%)	
		Full	Rare	Non-Rare	Full	Rare	Non-Rare	$AP_{role}^{S_1}$	$AP_{role}^{S_2}$
One-stage Methods:									
CATN [39] (ICCV'21)	R50	31.86	25.15	33.84	34.44	27.69	36.45	60.1	-
GEN-VLKT [32] (CVPR'2022)	R50+ViT-B/32	33.75	29.25	35.10	36.78	32.75	37.99	62.4	64.5
ERNet [40] (TIP'2023)	EfficientNetV2-L	34.25	28.70	36.33	-	-	-	61.6	-
SG2HOI+ [41] (TIP'2023)	R50+ViT-B/32	33.14	29.27	35.72	35.73	32.01	36.43	<u>63.6</u>	65.2
HODN [20] (TMM'2023)	R50	33.14	28.54	34.52	35.86	31.18	37.26	67.0	69.1
Multi-Step [42] (ACM MM'2023)	R101	34.42	30.03	35.73	37.71	33.74	<u>38.89</u>	61.3	<u>67.0</u>
HOICLIP [34] (CVPR'2023)	R50+ViT-B/32	34.69	31.12	35.74	37.61	34.47	38.54	63.5	64.8
CEFA [23] (ACM MM'2024)	R50+ViT-B/32	<u>35.00</u>	<u>32.30</u>	<u>35.81</u>	<u>38.23</u>	<u>35.62</u>	39.02	63.5	-
DP-ADN [24] (AAAI'2024)	R50+ViT-B/32	35.91	35.82	35.44	38.99	39.61	38.80	62.6	64.8
Two-stage Methods:									
PViC [†] [31] (ICCV'2023)	R50	34.69	32.14	35.45	38.14	35.38	38.97	59.7	65.4
ADA-CM [35] (ICCV'2023)	R50+ViT-L/14	38.40	37.52	38.66	-	-	-	58.6	64.0
CLIP4HOI [36] (NEURIPS'2023)	R50+ViT-B/16	35.33	33.95	35.74	-	-	-	-	66.3
CMMP [14] (ECCV'2024)	R50+ViT-L/14	38.14	37.75	38.25	-	-	-	-	64.0
Pose-Aware [7] (CVPR'2024)	R50	35.86	32.48	36.86	39.48	36.10	40.49	61.1	<u>66.6</u>
EZ-HOI [15] (NEURIPS'2024)	R50+ViT-L/14	38.61	37.70	38.90	-	-	-	60.5	66.2
LAIN [43] (CVPR'2025)	R50+ViT-B/16	36.02	35.70	36.11	-	-	-	-	65.1
HOLa [44] (ICCV'2025)	R50+ViT-L/14	<u>39.05</u>	<u>38.66</u>	<u>39.17</u>	-	-	-	60.3	66.0
CRL-B (Ours)	R50+ViT-B/16	<u>36.70</u>	<u>35.16</u>	<u>37.16</u>	<u>40.17</u>	<u>39.02</u>	<u>40.51</u>	60.2	65.9
CRL-L (Ours)	R50+ViT-L/14	39.99	40.67	39.78	43.35	44.43	43.02	60.9	66.9

[†] The released code of PViC for V-COCO is no longer available, thus the results reproduced in [7] are reported instead.

TABLE II

ABLATION STUDY OF OUR MODEL. UA: UNARY ASSOCIATION, TR: TERNARY RELATIONSHIP, GB: GLOBAL BLOCK, RB: REGIONAL BLOCK. ✓ INDICATES THAT THE MODULE IS USED.

MRM		CPL		HICO-Det (Default)		
UA	TR	GB	RB	Full	Rare	Non-Rare
-	-	-	-	35.45	33.66	35.99
✓	-	-	-	35.76	<u>34.10</u>	36.25
-	✓	-	-	35.94	34.00	36.52
✓	✓	-	-	36.00	34.07	36.58
-	-	✓	-	36.30	34.03	36.98
-	-	-	✓	35.85	33.25	36.63
-	-	✓	✓	<u>36.39</u>	<u>33.59</u>	37.22
✓	✓	✓	✓	36.70	35.16	<u>37.16</u>

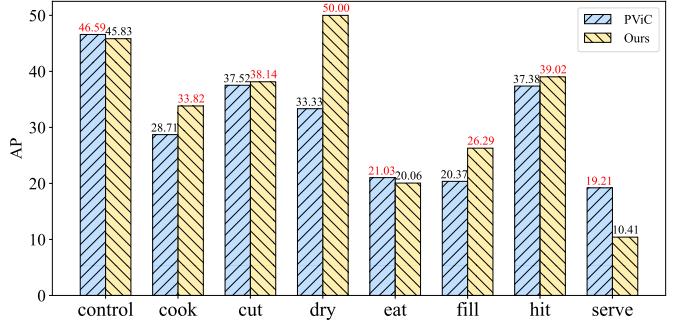


Fig. 3. Comparison of per-category accuracy between CRL-B and PViC [31] on HICO-Det-HTO.

C. Ablations Study

To demonstrate the effectiveness of our framework, we conducted several ablation studies on the HICO-Det dataset. Noting that our baseline is built on the inferior Variant E3 of PViC [31], which uses a ResNet-50 C5 backbone without an additional feature head. By replacing the decoder's original input feature with CLIP visual features V_e , this enhanced baseline achieves higher mAP scores compared to PViC, i.e., 35.45, 33.66 and 35.99 in full, rare and non-rare settings respectively.

Network Architecture Design. As shown in Table II, the introduction of Multivariate Relationship Modeling (MRM) consistently delivers performance gains. Concretely, it leads to **0.55** (full), **0.41** (rare) and **0.59** (non-rare) mAP improvements. In contrast, Contextualized Prompt Learning (CPL) achieves

surpassing recent work HOLa by margins of **0.7** mAP and **0.9** mAP, respectively.

To further assess the impact of tool affordance-guided interaction modeling, we construct a benchmark subset from HICO-Det, named HICO-Det-HTO. This subset exclusively includes interactions where explicit tool usage is involved. The final subset includes 783 test images. A category-wise accuracy analysis for each action is conducted on this new dataset, allowing a comparison with the baseline method PViC [31]. As shown in Fig. 3, it can be observed that our model outperforms PViC on several actions, including “cook”, “cut”, “dry”, “fill”, and “hit”.

TABLE III
ABLATIVE EXPERIMENTS FOR HYPER-PARAMETERS IN CPL. MANUAL: HAND-CRAFTED PROMPT.

(a) Prompt length		(b) Prompt setting				(c) CPL weight		
[ACT]	Full	Rare	Non-Rare	[PREFIX]	[ACT]	Full	Rare	Non-Rare
2	36.56	33.92	37.36	Manual	4	36.70	35.16	37.16
4	36.70	35.16	37.16		4	36.30	33.96	37.00
6	36.05	33.88	36.69		4	36.28	34.64	36.76
8	36.46	34.92	36.92		8	36.57	34.07	37.32
10	36.57	34.62	37.15		8	36.54	34.41	37.17

promising mAP increases in the full (**0.94**) and non-rare (**1.23**) settings. However, there is a slight decrease of **0.07** in mAP for the rare category. Fortunately, it can be effectively compensated by integrating the MRM module. When combined, they form a comprehensive model that delivers outstanding overall performance, notably achieving an impressive mAP boost of **1.50** in rare cases.

Multivariate Relationship Modeling. The modeled binary connection is inherently included in the baseline, which cannot be removed. Therefore, we focus on unary association (UA) and ternary relationship (TR) modeling. As shown in the “MRM” column in Table II, both are effective in improving interaction recognition. For the Full, Rare and Non-rare metrics, associating the textual information in unary modeling can improve the mAP by **0.31**, **0.44** and **0.26**, respectively. Meanwhile, the sole use of ternary relationship modeling boost the mAP by **0.49**, **0.34**, and **0.53**. The best performance is achieved when both of them are combined.

Contextualized Prompt Learning. The proposed Contextual Decoder is crucial in CPL to refine textual features derived from learnable prompts. It comprises two blocks, namely the global block (GB) and the regional block (RB). As illustrated in the “CPL” column in Table II, both layers individually enhance HOI performance, and their combined use results in the most significant improvements.

Moreover, those predefined parameters in CPL also need to be carefully analyzed. In Table III (a), we vary the length of learnable tokens [ACT] in our prompt template “A photo of a person [ACT] {object}”. The best performance is achieved with 4 tokens, while using more than this could potentially increase the complexity of learning. Instead of using the hand-crafted [PREFIX] tokens like “a photo of a”, we can use learnable ones, resulting in the format “[PREFIX] person [ACT] {object}”. As shown in Table III (b), switching to learnable [PREFIX] tokens appears to be less effective. This suggests that having too much flexibility in prompts might be suboptimal, aligning with the findings presented in Table III (a). Finally, as given in Table III (c), the performance is not particularly sensitive to changes in the CPL weight (i.e., β).

D. Qualitative Results

As shown in Fig. 5, we present qualitative comparisons between our method and PViC [31] on the HICO-Det dataset. In the first row of the figure, we present three examples involving tool-related interactions. For actions related to “hit”, our model demonstrates more effective recognition of the “hit” interaction. Specifically, in the leftmost image of the first row, PViC tends to predict the interaction as “hold” or

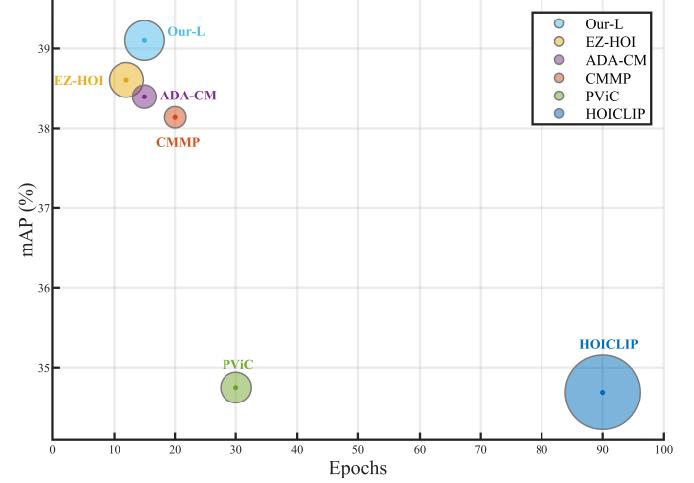


Fig. 4. Comparison of model performance with respect to learnable parameters and training epochs.

“carry”, whereas guided by tool-related semantics, our model correctly identifies the interaction as “fill”. It can be observed that our method performs robust in tool-related interaction cases and can consistently deliver more satisfying results. In the second row, we select examples of tool-irrelevant interactions for qualitative comparison. However, as illustrated by the rightmost example involving the action “paint”, painting typically requires a tool such as a brush. Since the predefined object detector cannot recognize brushes (as “paintbrush” is not among the 80 COCO categories), we did not include this sample in the HICO-Det-HTO dataset. In this example, our model assigns a higher score proportion to the “paint” interaction, while PViC incorrectly classifies the interaction mostly as “inspect”. In the two rightmost examples, PViC exhibits a tendency to classify the interactions as “hold”, while our approach successfully recognizes the more accurate interactions, namely “eat” and “hug”.

Moreover, as shown in Fig.6, we visualize attention maps corresponding to some of the images in Fig.5. Specifically, the first three examples illustrate the attention maps of the tool-related images in the first row of Fig.5, where the attention is taken from the last layer of the ternary decoder. We observe that, compared to PViC, our model places greater focus on the tools involved in the interactions. The last image in Fig.6 corresponds to the leftmost image in the second row of Fig.5. As previously mentioned, the brush in this image cannot be detected by the object detector, and thus cannot be recognized as a tool. Consequently, the ternary token cannot

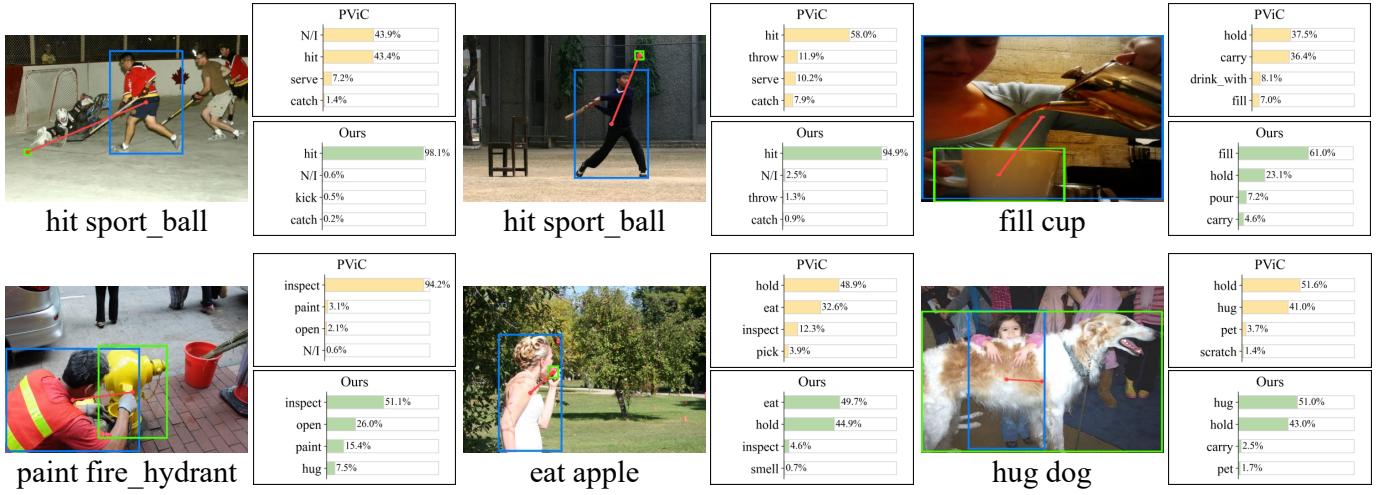


Fig. 5. Qualitative results on HICO-Det test set with fine-tuned DETR-R50 as the object detector. Bounding boxes of humans and objects are drawn with blue and green boxes. The textual annotation below the figure represents the ground truth. N/I denotes no interaction.

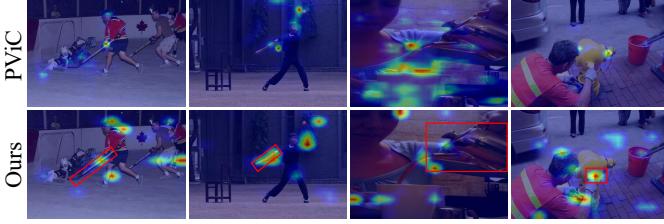


Fig. 6. The visualization of the cross-attention maps on a subset of images from Fig. 5. Bounding boxes of tools are drawn with red boxes. Best viewed in color.

be constructed, and no attention map is available from the ternary decoder. Instead, we visualize the attention map from the last layer of the pairwise decoder, and it shows that our model still attends to the brush held in the person’s hand.

During our experiments, we observed that existing datasets such as HICO-Det are not well-suited for annotating certain tool-related categories. For example, in some images depicting “a person hitting a baseball”, but the annotations are often limited to actions such as “person swinging a bat” or “person holding a bat”, which fail to capture the full semantics of the interaction. Moreover, the number of images involving measurable tool-related interactions is relatively small, making it difficult to effectively leverage tool-related information. In future work, we plan to expand the dataset by collecting more images, either from real-world sources or by generating synthetic images using generative models such as diffusion models.

E. Discussion

Model Complexity. Considering that the introduction of the ternary decoder and contextual decoder introduces additional trainable parameters, we conduct a comprehensive comparison with several recent one-stage and two-stage HOI models. Specifically, we compare our method with EZ-HOI [15], CMMP [14], and ADA-CM [35], all of which adopt ViT-

TABLE IV
ZERO-SHOT COMPARISONS WITH SOTA METHODS ON HICO-DET.

Method	Type	Unseen	Seen	Full
CMMP [14]	RF-UC	29.45	32.87	32.18
EZ-HOI [34]	RF-UC	29.02	34.15	33.13
LAIN [43]	RF-UC	31.83	35.06	34.41
HOLa [44]	RF-UC	30.61	35.08	34.19
Ours	RF-UC	<u>31.73</u>	36.61	35.63
CMMP [14]	NF-UC	32.09	29.71	30.18
EZ-HOI [34]	NF-UC	33.66	30.55	31.17
LAIN [43]	NF-UC	36.41	<u>32.44</u>	33.32
HOLa [44]	NF-UC	<u>35.25</u>	31.64	<u>32.36</u>
Ours	NF-UC	28.95	32.59	31.87
CMMP [14]	UO	33.76	31.15	31.59
EZ-HOI [34]	UO	33.28	32.06	32.27
LAIN [43]	UO	37.88	<u>33.55</u>	<u>34.27</u>
HOLa [44]	UO	36.45	33.02	33.59
Ours	UO	<u>36.74</u>	34.53	34.90
CMMP [14]	UV	26.23	32.75	31.84
EZ-HOI [34]	UV	25.10	33.49	32.32
LAIN [43]	UV	28.96	33.80	<u>33.12</u>
HOLa [44]	UV	<u>27.91</u>	<u>35.09</u>	34.09
Ours	UV	13.05	36.05	32.83

L/14 as the backbone. We also include comparisons with our baseline PViC [31] and the one-stage method HOICLIP [34].

To provide a fair and holistic evaluation, we take into account not only model accuracy but also the number of learnable parameters and training epochs. As shown in Fig. 4, all experiments are conducted on the HICO-DET test set. When using the CLIP ViT-L/14 variant, the visual encoder outputs 1024-dimensional features, which significantly increases the parameter size of the downstream attention modules. To mitigate this overhead, we adopt lightweight down-projection layers that reduce the feature dimensionality from 1024 to 512, effectively lowering the overall model complexity. As a result, our model contains 18.3M learnable parameters. While this is slightly higher than EZ-HOI (14.1M) and CMMP (5.4M), our method achieves better performance on the full set, reaching 39.10 mAP.

Zero-shot Setting. Table IV compares the zero-shot HOI detection performance of our method with four SOTA approaches (CMMP, EZ-HOI, LAIN, HOLA) on HICO-Det across four scenarios (RF-UC, NF-UC, UO, UV), evaluating Unseen, Seen, and Full metrics. Since the experiment is conducted under a zero-shot setting, we freeze the last three classifiers and adopt the Verb Adapter from HOICLIP [34] as the weights for these three classifiers. Our method exhibits notable strengths: it ranks first in the Seen metric across all scenarios (e.g., 36.61% in RF-UC, 36.05% in UV) and secures top positions in the Full metric for RF-UC (35.63%) and UO (34.90%), with competitive Unseen performance in these two scenarios (31.73% for RF-UC, 36.74% for UO). However, it has room for improvement: in NF-UC, its Unseen score (28.95%) drags the Full metric to 31.87. In the unseen verb (UV) task, the low Unseen score (13.05%) restricts the Full metric to only 32.83%. The poor performance of unseen verbs stems from two key issues: tool-related ones like “stab” and “swing” are often misclassified as familiar actions such as “hit” or “serve” due to imprecise feature links, while abstract verbs like “inspect” and “install” lack consistent connections to known actions. Their vague boundaries and context-dependent use further hinder knowledge transfer. To fix this, future work should refine tool-action representations to capture subtle differences and integrate goal information for abstract verbs.

V. CONCLUSION

In this paper, we have proposed CRL, a two-stage HOI detection framework that addresses two key limitations of existing methods: the lack of tool-mediated modeling and the inability to integrate visual features into prompt learning, limiting their capacity to capture context-dependent relationships. Our approach has shown notable effectiveness in challenging HOI scenarios with complex tool-assisted interactions and ambiguous contextual cues, highlighting the benefits of multivariate relationship modeling and contextualized prompt learning. In future work, we plan to explore fine-grained affordance-guided modeling and broaden generalization to open-world HOI scenarios.

REFERENCES

- [1] S. Liu and X. Ma, “Attention-driven appearance-motion fusion network for action recognition,” *IEEE Transactions on Multimedia*, vol. 25, pp. 2573–2584, 2022.
- [2] L. Xu, Q. Wu, L. Pan, F. Meng, H. Li, C. He, H. Wang, S. Cheng, and Y. Dai, “Towards continual egocentric activity recognition: A multi-modal egocentric activity dataset for continual learning,” *IEEE Transactions on Multimedia*, vol. 26, pp. 2430–2443, 2023.
- [3] M. Wang, Z. Huang, X. Kong, G. Shen, G. Dai, J. Wang, and Y. Liu, “Action detail matters: Refining video recognition with local action queries,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 19 132–19 142.
- [4] C. Tao, C. Wang, S. Lin, S. Cai, D. Li, and J. Qian, “Feature reconstruction with disruption for unsupervised video anomaly detection,” *IEEE Transactions on Multimedia*, 2024.
- [5] D. Wang, X. Lu, Q. Wang, Y. Tian, B. Wan, and L. He, “Gist, content, target-oriented: A 3-level human-like framework for video moment retrieval,” *IEEE Transactions on Multimedia*, 2024.
- [6] Y. Shi, X. Wu, H. Lin, and J. Luo, “Commonsense knowledge prompting for few-shot action recognition in videos,” *IEEE Transactions on Multimedia*, vol. 26, pp. 8395–8405, 2024.
- [7] E. Z. Wu, Y. Li, Y. Wang, and S. Wang, “Exploring pose-aware human-object interaction via hybrid learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 815–17 825.
- [8] F. Z. Zhang, D. Campbell, and S. Gould, “Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 104–20 112.
- [9] Z. Hou, B. Yu, Y. Qiao, X. Peng, and D. Tao, “Affordance transfer learning for human-object interaction detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 495–504.
- [10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.
- [11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [12] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to prompt for vision-language models,” *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.
- [13] S. Wang, Y. Duan, H. Ding, Y.-P. Tan, K.-H. Yap, and J. Yuan, “Learning transferable human-object interaction detector with natural language supervision,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 939–948.
- [14] T. Lei, S. Yin, Y. Peng, and Y. Liu, “Exploring conditional multimodal prompts for zero-shot hoi detection,” in *European Conference on Computer Vision*. Springer, 2024, pp. 1–19.
- [15] Q. Lei, B. Wang, and R. Tan, “Ez-hoi: Vlm adaptation via guided prompt learning for zero-shot hoi detection,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 55 831–55 857, 2024.
- [16] C. Zou, B. Wang, Y. Hu, J. Liu, Q. Wu, Y. Zhao, B. Li, C. Zhang, C. Zhang, Y. Wei *et al.*, “End-to-end human object interaction detection with hoi transformer,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 825–11 834.
- [17] B. Kim, J. Lee, J. Kang, E.-S. Kim, and H. J. Kim, “Hotr: End-to-end human-object interaction detection with transformers,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 74–83.
- [18] M. Tamura, H. Ohashi, and T. Yoshinaga, “Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 410–10 419.
- [19] C. Xie, F. Zeng, Y. Hu, S. Liang, and Y. Wei, “Category query learning for human-object interaction classification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 275–15 284.
- [20] S. Fang, Z. Lin, K. Yan, J. Li, X. Lin, and R. Ji, “Hodn: Disentangling human-object feature for hoi detection,” *IEEE Transactions on Multimedia*, 2023.
- [21] A. Zhang, Y. Liao, S. Liu, M. Lu, Y. Wang, C. Gao, and X. Li, “Mining the benefits of two-stage and one-stage hoi detection,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 209–17 220, 2021.
- [22] S. Kim, D. Jung, and M. Cho, “Relational context learning for human-object interaction detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2925–2934.
- [23] L. Zhang, W. Suo, P. Wang, and Y. Zhang, “A plug-and-play method for rare human-object interactions detection by bridging domain gap,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 8613–8622.
- [24] J. Gao, K. Liang, T. Wei, W. Chen, Z. Ma, and J. Guo, “Dual-prior augmented decoding network for long tail distribution in hoi detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, 2024, pp. 1806–1814.
- [25] T. He, L. Gao, J. Song, and Y.-F. Li, “Exploiting scene graphs for human-object interaction detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 984–15 993.
- [26] X. Liu, Y.-L. Li, X. Wu, Y.-W. Tai, C. Lu, and C.-K. Tang, “Interactiveness field in human-object interactions,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 113–20 122.
- [27] X. Wu, Y.-L. Li, X. Liu, J. Zhang, Y. Wu, and C. Lu, “Mining cross-person cues for body-part interactiveness learning in hoi detection,” in

- European Conference on Computer Vision.* Springer, 2022, pp. 121–136.
- [28] Y. Zhang, Y. Pan, T. Yao, R. Huang, T. Mei, and C.-W. Chen, “Exploring structure-aware transformer over interaction proposals for human-object interaction detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 548–19 557.
- [29] S. Zheng, B. Xu, and Q. Jin, “Open-category human-object interaction pre-training via language modeling framework,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 392–19 402.
- [30] Y. Guo, Y. Liu, J. Li, W. Wang, and Q. Jia, “Unseen no more: Unlocking the potential of clip for generative zero-shot hoi detection,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 1711–1720.
- [31] F. Z. Zhang, Y. Yuan, D. Campbell, Z. Zhong, and S. Gould, “Exploring predicate visual context in detecting of human-object interactions,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 411–10 421.
- [32] Y. Liao, A. Zhang, M. Lu, Y. Wang, X. Li, and S. Liu, “Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 123–20 132.
- [33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [34] S. Ning, L. Qiu, Y. Liu, and X. He, “Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 507–23 517.
- [35] T. Lei, F. Caba, Q. Chen, H. Jin, Y. Peng, and Y. Liu, “Efficient adaptive human-object interaction detection with concept-guided memory,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6480–6490.
- [36] Y. Mao, J. Deng, W. Zhou, L. Li, Y. Fang, and H. Li, “Clip4hoi: towards adapting clip for practical zero-shot hoi detection,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 45 895–45 906, 2023.
- [37] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Conditional prompt learning for vision-language models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 816–16 825.
- [38] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European conference on computer vision*. Springer, 2020, pp. 213–229.
- [39] L. Dong, Z. Li, K. Xu, Z. Zhang, L. Yan, S. Zhong, and X. Zou, “Category-aware transformer network for better human-object interaction detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 538–19 547.
- [40] J. Lim, V. M. Baskaran, J. M.-Y. Lim, K. Wong, J. See, and M. Tistarelli, “Ernet: An efficient and reliable human-object interaction detection network,” *IEEE Transactions on Image Processing*, vol. 32, pp. 964–979, 2023.
- [41] T. He, L. Gao, J. Song, and Y.-F. Li, “Toward a unified transformer-based framework for scene graph generation and human-object interaction detection,” *IEEE Transactions on Image Processing*, vol. 32, pp. 6274–6288, 2023.
- [42] Y. Zhou, G. Tan, M. Li, and C. Gou, “Learning from easy to hard pairs: Multi-step reasoning network for human-object interaction detection,” in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 4368–4377.
- [43] S. Kim, D. Jung, and M. Cho, “Locality-aware zero-shot human-object interaction detection,” *arXiv preprint arXiv:2505.19503*, 2025.
- [44] Q. Lei, B. Wang, and R. T. Tan, “Hola: Zero-shot hoi detection with low-rank decomposed vlm feature adaptation,” *arXiv preprint arXiv:2507.15542*, 2025.
- [45] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, “Learning to detect human-object interactions,” in *2018 ieee winter conference on applications of computer vision (wacv)*. IEEE, 2018, pp. 381–389.
- [46] S. Gupta and J. Malik, “Visual semantic role labeling,” *arXiv preprint arXiv:1505.04474*, 2015.
- [47] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.
- [48] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [49] I. Loshchilov, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.