CL-HOI: Cross-Level Human-Object Interaction Distillation from Vision Large Language Models

Anonymous ACL submission

Abstract

Human-object interaction (HOI) detection has seen advancements with Vision Language Models (VLMs), but these methods often depend on extensive manual annotations. Vision Large Language Models (VLLMs) can inherently recognize and reason about interactions at the image level but are computationally heavy and not designed for instance-level HOI detection. To overcome these limitations, we propose a Cross-Level HOI distillation (CL-HOI) framework, which distills instance-level HOIs from VLLMs' image-level understanding without the need for manual annotations. Our approach involves two stages: **context distillation**, where a Visual Linguistic Translator (VLT) converts visual information into linguistic form, and **interaction distillation**, where an Interaction Cognition Network (ICN) reasons about spatial, visual, and context relations. We design contrastive distillation losses to transfer image-level context and interaction knowledge from the teacher to the student model, enabling instance-level HOI detection. Evaluations on HICO-DET and V-COCO datasets demonstrate that our CL-HOI surpasses existing weakly supervised methods and VLLM supervised methods, showing its efficacy in detecting HOIs without manual labels.

1 Introduction

Human-object interaction (HOI) detection is a task to detect the human-object pairs and corresponding interactions in triplets of <human, interaction, object>. It is a fundamental task in computer vision and vital for surveillance (Alkilani and Shirkhodaie, 2013), human-robot interaction (Jahanmahin et al., 2022), and human-computer interaction (Chaves and Gerosa, 2021).

HOI detection has increasingly adopted vision language models (VLMs) (Radford et al., 2021), demonstrating significant potential in improving detection accuracy. Current HOI detectors primarily utilize these models to generate supplementary information or distill vision-text alignment, both of which enhance the learning process of HOI detectors. For generating supplementary information, some methods (Cao et al., 2024) employ VLMs to produce descriptions of interactions or integrate them as sub-modules to improve detection performance, thereby providing additional context and understanding. For distilling vision-text alignment abilities, there are generally two approaches: prompting (Wang et al., 2022) and mimicking (Liao et al., 2022). Prompting involves guiding vision encoders within VLMs to learn instance-level interaction features, effectively training the model to recognize and interpret specific interactions within an image. In contrast, mimicking aims to bridge the gap between the image-level features extracted by a customized vision encoder and those derived from the vision encoder in VLMs. Despite these advancements, these methods still rely on manual annotations at either the instance or image level.

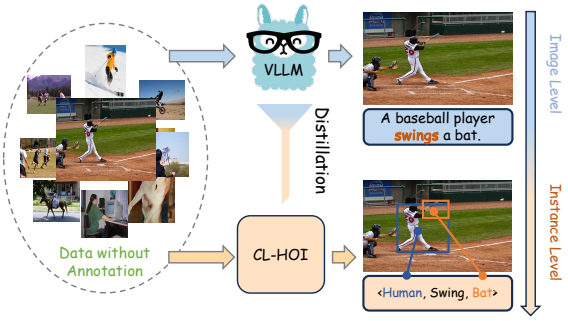

Figure 1: **Cross-Level HOI Distillation**: VLLMs exhibit strong capabilities in image-level interaction recognition and reasoning. However, they are not equipped to perform instance-level HOI detection, which requires associating detected humans and objects. Our CL-HOI is proposed to distill image-level abilities to instance-level HOI detections.

Recently, vision large language models (VLLMs) (Li et al., 2023; Dai et al., 2023) have shown significant promise across various vision tasks, including image captioning and

visual question answering. Trained on vast and diverse datasets, VLLMs exhibit remarkable zero-shot capabilities, enabling them to achieve image-level HOI understanding by recognizing human activities and reasoning about interactions, as illustrated in Fig. 1. However, these capabilities are computationally intensive and remain confined to the image level, lacking the ability to perform instance-level HOI detection. Instance-level reasoning adds complexity and nuance by identifying and interpreting specific interactions between detected humans and objects, a task that current VLLMs are not fully explored. Even though RLIPv2 (Yuan et al., 2023) has adopted BLIP to generate captions, it uses some language tools and additional networks to generate interaction supervision, which is inefficient.

Building on the motivation presented in Fig. 1, we introduce an annotation-free and Cross-Level HOI distillation (CL-HOI, student) framework to transfer the image-level HOI recognition and reasoning capabilities from the teacher VLLM (Dai et al., 2023) to instance-level HOI detection. The distillation is proposed in a two-stage manner: context distillation and interaction distillation. For context distillation, we first prompt the teacher VLLM to generate context captions for each image, focusing on human activities. We then introduce a multimodal Visual Linguistic Translator (VLT) in CL-HOI that integrates a vision-language model with learnable adapters, enabling the translation of image features into context features and facilitating the learning of image-level human activities from context captions. For interaction distillation, we employ the language model (Vicuna (Chiang et al., 2023)) within the VLLM to parse HOI triplets in the format <human, interaction, object>. Consequently, we introduce an Interaction Cognitive Network (ICN) in the student CL-HOI, which employs a novel chain-of-cognition strategy to effectively analyze HOIs through spatial, visual, and context cognition. This allows the model to learn instance-level interactions from parsed image-level HOI triplets, associating detected humans and objects. To enable cross-level distillation, we introduce contrastive distillation losses to capture contextual information in VLT and guide the ICN to detect interactions, thereby bridging the gap between image-level and instance-level detection.

Our contributions can be summarized as: (1) We introduce the pipeline that distills image-level knowledge from VLLMs to instance-level HOI detection without relying on manual annotation. (2) We propose a new solution called CL-HOI, which includes context distillation and interaction distillation stages with the Visual Linguistic Translator and Interaction Cognition Network to effectively generate explicit interaction representations. (3) We introduced contrastive distillation losses to facilitate the effective learning of context and interactions from VLLMs. We evaluated our methods on two benchmarking datasets. Our findings show that CL-HOI achieves state-of-the-art VLLMs supervised performance, with 17.5% mAP on HICO-DET and 36.63% Role AP on V-COCO.

2 Related Work

2.1 Human-Object Interaction Detection

Most existing HOI detection methods can be broadly categorized into groups, such as one-stage and two-stage methods. Two-stage methods (Gao et al., 2018; Ulutan et al., 2020; Zhang et al., 2022) typically apply an off-the-shelf object detector (Girshick, 2015; Carion et al., 2020) to detect the location of humans and objects, along with their semantic labels. Subsequently, detected boxes are fed to the interaction classification module for interaction modeling. In addition to relying on object appearances through object detectors, the two-stage models can leverage contextual features (Gao et al., 2018; Ulutan et al., 2020), spatial distribution (Zhang et al., 2021), and human pose (Fang et al., 2018; Wan et al., 2019) to enhance the analysis of Inspired by one-stage object detection (Carion et al., 2020), recent researchers attempted to develop end-to-end HOI detection (Zou et al., 2021; Liao et al., 2020; Kim et al., 2021; Lim et al., 2023) strategies for simultaneous detection of human/object localization and interaction classification. Some of the most recent methods (Yang et al., 2024; Li et al., 2024) aim to pre-train or train on extra relational data to improve performance.

Beyond one-stage and two-stage methods, which depend on expensive instance-level supervision, recent research has shifted towards weakly supervised methods. Kumaraswamy et al. (Kumaraswamy et al., 2021) proposed a momentum-independent learning and element-swapping strategy for both fully and weakly supervised HOI detection. More recently, Unal et al. (Unal and Kovashka, 2023) introduced a method to detect HOIs based on image-level interaction labels (e.g., <*jump*>, <*hold*>), utilizing these labels for interaction detection.

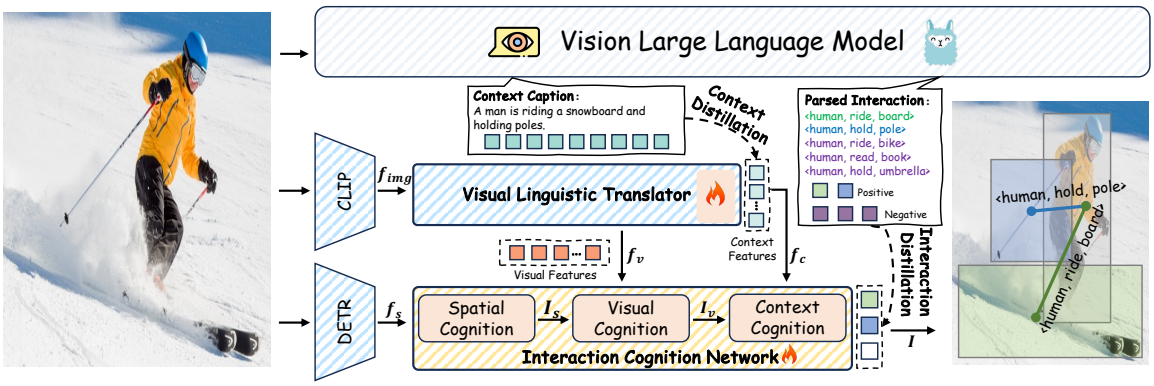

Figure 2: The proposed CL-HOI framework focuses on distilling instance-level interaction detection from the image-level recognition and reasoning abilities of the VLLM (teacher model). Guided by context captions and parsed interaction triplets from the teacher model, CL-HOI performs context and interaction distillation for the visual-linguistic translator (VLT) and Interaction Cognitive Network (ICN) in two stages of instance-level interaction learning without annotated supervision.

2.2 Vision-Language Models in HOI Detection

Many research works have investigated their application in incorporating various computer vision tasks with recent emerging developments in the vision language models. For instance, extensive research has been conducted on the popular vision-language models (e.g., CLIP (Radford et al., 2021), BLIP (Li et al., 2023)) in the context of HOI detection (Wang et al., 2022; Park et al., 2023; Liao et al., 2022). Gen-VLKT (Liao et al., 2022) introduces visual-linguistic knowledge transfer that relies on the semantic information extracted from a visual-linguistic pre-trained model to enhance interaction learning. Most existing multimodal-based HOI methods (Liao et al., 2022; Wang et al., 2022) desire to distill the knowledge from large models and achieve better performance in supervised learning. More recent foundation models like VLLM and LLM have also been explored in HOI detections. Usually, such foundation models (Gao et al., 2024; Cao et al., 2024) are prompted to provide additional cues to help detectors better learn HOIs.

2.3 Knowledge Distillation

Knowledge distillation (Gou et al., 2021) aims to enhance the learning capabilities of models by leveraging the strengths of large pre-trained models without relying on extensive labeled datasets. In knowledge distillation, a teacher model, usually a large and well-trained network, transfers its knowledge to a smaller student model, guiding the student to mimic the teacher's predictions. Knowledge distillation eliminates the need for labeled data by using the intrinsic information within the data itself to guide the learning process. The teacher model generates pseudo-labels or other forms of supervision signals from the data, which the student model then uses to learn. These signals can come from various forms of supervision, such as predicting parts of the data from other parts (Lee et al., 2021), identifying patterns within the data (Ericsson et al., 2022), or utilizing domain-specific transformations and augmentations (Bucci et al., 2021).

3 Methodology

3.1 Overall Architecture

The overall structure of our proposed CL-HOI framework is illustrated in Fig. 2. This framework is designed to distill human-object interactions (HOIs) from a vision large language model (VLLM) through a two-stage process: context distillation and interaction distillation. Within the student CL-HOI model, we initially use an off-the-shelf object detector (DETR) and a CLIP visual encoder to detect person-object pairs and extract spatial features $\underline{f_s}$ as well as image features $\underline{f_{img}}$. The spatial features $\underline{f_s}$ consist of human and object feature embeddings e_h and e_o, bounding boxes b_h and b_o, confidence scores s_h and s_o, and object categories c_o, forming the person-object pairs $\{e_h, e_o, b_h, b_o, s_h, s_o, c_o\}$. In context distillation, we first prompt the VLLMs to generate the image captions (see details in Appendix). And then, we propose a Visual Linguistic Translator (VLT) to transfer the image features $\underline{f_{img}}$ into context features $\underline{f_c}$ and visual features $\underline{f_v}$, thereby distilling context captions from the VLLM. In the interaction distillation stage, we first use the LLM in the VLLM to parse the interactions. We also introduce

a Chained Interaction Cognition Network (ICN) to reason about and prompt interactions based on spatial, visual, and contextual cognition, learning from the parsed interactions provided by the VLLM. Finally, we propose contrastive distillation losses to facilitate cross-level distillation and transfer the recognition and reasoning capabilities from the VLLM to the student model.

3.2 Context Distillation Stage

In the context distillation stage, we design a Visual Linguistic Translator (VLT) to generate context features f_c and visual features f_v by translating the image features f_{img} from visual space to linguistic context space, supervised by context captions prompted from the VLLM. Specifically, we insert two tuning adapters before and after the Q-Former (Li et al., 2023). Given the image features f_{img} from the CLIP encoder (Radford et al., 2021), the first adapter is used to map and project the image features to the Q-Former input space. Then, the projected feature with a set of initialized context queries C_0 are passed through the Q-Former to update the projected image features and context queries, obtaining updated image features f'_{img} with context features f_c. Thus, the final context features can learn rich context information that is closely aligned with the image features. Besides the context feature generation, we introduce an additional tuning adapter to generate visual features f_v. This can be achieved by aligning the image features f'_{img} with the context features f_c in the same dimension using an additional learnable projection layer.

Moreover, we propose a context distillation loss in Section 3.4 to enable the bidirectional exchange between the visual and linguistic modalities within the proposed translator. Thus, it ensures that context and visual features mutually enhance each other, leading to a more accurate interpretation of human-object interaction reasoning in the subsequent phase.

3.3 Interaction Distillation Stage

In this stage, we introduce a chained Interaction Cognition Network (ICN) that sequentially implements spatial, visual, and context cognitions to interpret and reason interactions from parsed interactions. The process begins with spatial cognition, where geometric priors from detected person-object pairs are leveraged to generate initial interaction features with spatial information, denoted as I_s. This is followed by visual cognition, which refines these features I_s by incorporating detailed interaction features from the visual features f_v and generates I_v. Finally, the context cognition stage integrates context features f_c into I_v to obtain the interaction features enriched with contextual information, denoted as I_c. We utilize I_c as the final interaction features I for interaction learning. This chain of cognitive interaction reasoning creates a multi-layered interpretation of interactions, seamlessly blending geometric positioning, visual cues, and contextual narratives.

Spatial Cognition. To capture spatial relationships, we first generate a set of interaction features with spatial information I_s from the detected spatial features f_s represented as person-object pairs $\{e_h, e_o, b_h, b_o, s_h, s_o, c_o\}$. We derive spatial features based on relative positions (c_h, c_o), dimensions (w_h, h_h, w_o, h_o), areas (a_h, a_o), aspect ratios (a'_h, a'_o), intersection over union (IoU), and relative distances and directions (d, d') from person-object pairs to capture spatial relationships. These spatial features are projected into spatial priors S using a linear layer. Simultaneously, content priors are refined through a self-attention mechanism within a Transformer encoder, modulating the embeddings e_h and e_o with their corresponding bounding boxes b_h and b_o by

$$E = \mathrm{SelfAttn}(e + \mathrm{PE}(b), e + \mathrm{PE}(b), e), \quad (1)$$

where $PE(\cdot)$ is the positional encoding function. The spatial S and content features E are then concatenated to form a set of interaction features with spatial information $I_s = \mathrm{Concat}(E_h, E_o)$.

Visual Cognition. The visual cognition in the chain aims to refine I_s to interaction features with visual information I_v by integrating visual interaction features f_v and guided by human-object union-region bounding boxes b_u. This process initiates with generating I_{v0} by augmenting I_s with positional encodings $PE(c_u^x, c_u^y)$ (Zhang et al., 2023) specific to the center position (c_u^x, c_u^y) of b_u. A multilayer perceptron (MLP) is proposed for dimensionality matching and feature enhancement. Concurrently, the visual features f_v receive their positional encodings based on their spatial coordinates (c_u^x, c_u^y), preparing them to serve as keys and values in the upcoming cross-attention operation. The cross-attention mechanism takes the prepared queries, keys, and values to produce updated in-

teraction features I_v imbued with rich visual interaction features. The process concludes with a pass through a feed-forward network (FFN), further refining these interaction features with visual information. The procedure to generate P_v can be formulated as:

$$I_{v0} = \mathrm{MLP}(\mathrm{Concat}(I_s, \mathrm{PE}(c_u^x, c_u^y))), \tag{2}$$

$$I_v = \mathrm{FFN}(\mathrm{CrossAttn}(I_{v0}, \mathrm{Concat}(f_v, \mathrm{PE}(c_u^x, c_u^y)), f_v)). \tag{3}$$

Context Cognition. The context cognition phase within the cognition chain is crucial for obtaining interaction features within context information f_c by weaving the context features f_c to I_c. This process involves a cross-attention mechanism where the interaction features I_v from visual cognition are used as queries, and the context features f_c with rich context features serve as both keys and values. This attention-based interaction allows for the contextual and linguistic details to be aligned with and imposed on the interaction features, resulting in a new set of updated interaction features I_c. These are then fine-tuned through a feed-forward network (FFN), which refines the features and injects them with depth for a comprehensive understanding of interactions. The process to generate I_c can be formulated as:

$$I_c = \mathrm{FFN}(\mathrm{CrossAttn}(I_v, f_c, f_c)). \tag{4}$$

For simplicity in subsequent discussions, we use I to denote the final interaction features I_c, which are now thoroughly informed by contextual information, facilitating a more sophisticated interpretation of human-object interactions. Based on the final interaction features, we proposed interaction distillation losses in Section 3.4 to achieve the cross-level distillation between the image-level supervision and instance-level interaction features.

3.4 Contrastive Distillation Losses

In the CL-HOI framework, contrastive distillation losses are essential for enabling information flow from the image level to the instance level. These losses are designed to leverage global similarities between features without requiring explicit matching between individual instances, drawing inspiration from DetCLIPv2 (Yao et al., 2023). This approach enables the model to learn from the overall relationships between features, facilitating the transfer of knowledge from broader image-level representations to more specific instance-level features, even when the exact correspondence between instances is unknown. Given any two features $f_1 \in \mathbb{R}^{M\times D}$ and $f_2 \in \mathbb{R}^{N\times D}$, the similarities can be calculated as:

$$s(f_1, f_2) = \frac{1}{M}\sum_{m\in M}\sum_{n\in N}\left(\frac{\exp(\mathrm{sim}_{m,n})}{\sum_{k\in N}\exp(\mathrm{sim}_{m,k})}\right)\mathrm{sim}_{m,n}, \tag{5}$$

$$\mathrm{sim}_{m,n} = \langle f_1[m,\cdot], f_2[n,\cdot]\rangle, \quad \forall m \in M,\, n \in N. \tag{6}$$

Based on the similarity, we design context and interaction distillation losses to guide the two-stage distillation.

Context Distillation Loss. This loss is designed to optimize the context features f_c to align with the context captions. It contrasts the similarity between the encoded context features f_c and the caption features derived from the VLLM. In detail, generated captions in a mini-batch with B captions are first encoded into caption features f_{cap} using the text encoder of CLIP. Then, the context and caption features are then paired as $\{f_c^i, f_{cap}^i\}_{i=1}^B$. Then, we calculate the contrastive caption loss as:

$$L_c = -\frac{1}{B}\sum_{i\in B}\log\frac{\exp\left(s(f_c^i, f_{cap}^i)\right)}{\sum_{j\in B}\exp\left(s(f_c^i, f_{cap}^j)\right)}. \tag{7}$$

Interaction Distillation Loss. In annotation-free settings, the absence of one-to-one instance-level matching necessitates carefully designing interaction distillation losses to improve HOI detection. We introduce three types of losses: 1) Image-to-Text (I2T) loss, 2) Text-to-Image (T2I) loss, and 3) Soft-Relation (SR) loss. For interaction features I in a mini-batch, we first encode verbs generated from the VLLM for each image into f_{verb} using the text encoder in CLIP. These verb features f_{verb} paired with interaction features are treated as positives f_{pos}. For negatives f_{neg}, we randomly sample negative verbs that are different from positive verbs from a verb dictionary (non-overlaps with ground-truth verb labels in original datasets) and encode them using the text encoder in CLIP. Further details on this process and the construction of the verb dictionary are provided in the Supplementary Materials. Thus, the interaction features and verb features in a mini-batch can be organized into $\{I^i \in \mathbb{R}^{K\times D}, f_{pos}^i \in \mathbb{R}^{P\times D}, f_{neg}^i \in \mathbb{R}^{Q\times D}\}_{i=1}^B$.

I2T loss aims to ensure the learned interactions are tightly associated with positive verbs

from image-level aspects. Specifically, the aggregated image-level interaction features are encouraged to align with positive verbs while contrasting against negative verbs. For each $I^i \in \mathbb{R}^{K\times D}$ in a mini-batch, we first average instance-level interaction features into image-level interaction features $I^i_{img} \in \mathbb{R}^{1\times D}$. We then calculate the image-to-text loss as:

$$L_{I2T} = -\frac{1}{B}\sum_{i\in B}\log\frac{\exp(s(I^i_{img}, f^i_{pos}))}{\exp(s(I^i_{img}, f^i_{pos})) + \exp(s(I^i_{img}, f^i_{neg}))}. \quad (8)$$

T2I loss is designed to ensure that positive verbs are represented in the learned interactions while negative verbs are excluded. The text-to-image loss is calculated as:

$$L_{T2I} = -\frac{1}{B}\sum_{i\in B}\log\frac{\exp\left(s(f^i_{pos}, I^i)\right)}{\sum_{j\in B}\exp\left(s(f^i_{pos}, I^j)\right)}. \quad (9)$$

SR loss is inherited from weakly-supervised HOI detection, which uses CLIP (Radford et al., 2021) to generate pseudo-HOI labels from the generated interaction labels and cropped human-object regions. Details of the soft-relation loss L_{SR} are provided in the supplementary materials. The final loss is calculated as follows:

$$L = L_c + L_{I2T} + L_{T2I} + L_{SR}. \quad (10)$$

4 Experiments

4.1 Setup

Datasets. We evaluate the proposed method with well-established HOI detection datasets, HICO-DET (Chao et al., 2018) and V-COCO (Gupta and Malik, 2015). HICO-DET consists of 37,633 and 9,546 images for training and testing, where the images are annotated with bounding boxes for human-object pairs and their interaction labels. This dataset includes 80 object categories (consistent with the MS COCO dataset (Gupta and Malik, 2015)) and 117 interaction categories, resulting in a total of 600 unique <interaction, object> pairs. The V-COCO dataset consists of 5,400 and 4,946 images for training and testing, respectively. It is a subset of the MS COCO dataset, which is annotated with 80 object classes and 26 interaction categories. Each image is associated with five captions.

Evaluation Metrics. We adopt commonly used evaluation metrics to evaluate the effectiveness of our proposed model, which uses Full mAP for HICO-DET and Role AP for V-COCO. It is worth noting that the evaluation matric of Full mAP and Role AP are analogous. Both metrics require that the predicted human and object bounding boxes exhibit an IoU of at least 0.5 with their respective HOI targets, and the predicted interaction category must align with the target interaction label.

Methods	Sup.	Ann.	Backbone	mAP
iCAN (Gao et al., 2018)	Full	✓	RN50	14.84
VSGNet (Ulutan et al., 2020)	Full	✓	RN152	19.80
SCG (Zhang et al., 2021)	Full	✓	RN50 FPN	21.85
IDN (Li et al., 2020)	Full	✓	RN50	23.36
HOTR (Kim et al., 2021)	Full	✓	RN50+Tr.	25.10
MSTR (Kim et al., 2022)	Full	✓	RN50+Tr.	31.17
Gen-VLKT (Liao et al., 2022)	Full	✓	RN50+Tr.	33.75
RLIPv2 (Yuan et al., 2023)	Full	✓	RN50+Tr.	33.57
PViC (Zhang et al., 2023)	Full	✓	RN50+Tr.	33.80
CMMP (Lei et al., 2024)	Full	✓	RN50+Tr	33.24
UPT (Baseline) (Zhang et al., 2022)	Full	✓	RN50+Tr.	31.66
CL-HOI (Ours)	Full	✓	RN50+Tr.	34.10
MX-HOI (Kumaraswamy et al., 2021)	Weak	✓	RN101	16.14
SCG (Zhang et al., 2021)	Weak	✓	RN50 FPN	7.05
IO-HOI (Unal and Kovashka, 2023)	Weak	✓	RN50 FPN	8.38
RLIPv2 (Yuan et al., 2023)	VLLM	✗	RN50+Tr.	15.05
CL-HOI (Ours)	VLLM	✗	RN50+Tr.	15.62
CL-HOI (Ours)	VLLM	✗	Swin-L+Tr.	**17.55**

Table 1: Benchmark comparisons on the HICO-DET full mAP with both fully and weakly supervised settings. Sup: Supervision; Ann: Annotation; Tr: Transformer.

Methods	Sup.	Ann.	Backbone	Role AP
iCAN (Gao et al., 2018)	Full	✓	RN50	52.04
VSGNet (Ulutan et al., 2020)	Full	✓	RN152	57.00
SCG (Zhang et al., 2021)	Full	✓	RN50 FPN	58.01
IDN (Li et al., 2020)	Full	✓	RN50	60.30
HOTR (Kim et al., 2021)	Full	✓	RN50+Tr.	64.40
MSTR (Kim et al., 2022)	Full	✓	RN50+Tr.	65.20
Gen-VLKT (Liao et al., 2022)	Full	✓	RN50+Tr.	64.80
PViC (Zhang et al., 2023)	Full	✓	RN50+Tr.	67.80
RLIPv2 (Yuan et al., 2023)	Full	✓	RN50+Tr.	64.50
CMMP (Lei et al., 2024)	Full	✓	RN50+Tr.	61.20
UPT (Baseline) (Zhang et al., 2022)	Full	✓	RN50+Tr.	64.50
CL-HOI (Ours)	Full	✓	RN50+Tr.	65.44
SCG (Zhang et al., 2021)	Weak	✓	RN50 FPN	20.05
IO-HOI (Unal and Kovashka, 2023)	Weak	✓	RN50 FPN	29.59
CL-HOI (Ours)	VLLM	✗	RN50+Tr.	26.74
CL-HOI (Ours)	VLLM	✗	Swin-L+Tr.	**36.63**

Table 2: Benchmark comparisons on the V-COCO datasets with respect to Role AP with both fully and weakly supervised methods. Sup: Supervision; Ann: Annotation; Tr: Transformer.

Baseline and Training Procedure. Our method is built upon a two-stage HOI detection method (Zhang et al., 2022). We utilize DETR (Carion et al., 2020) that are pre-trained on MS COCO for object detection. We initialed the learning rate as $1e^{-4}$ and subsequently reduced it to $1e^{-5}$ after the

Method	Annotation	Backbone	Role AP
SCG (Zhang et al., 2021)	✓	RN50 FPN	20.05
IO-HOI (CC) (Unal and Kovashka, 2023)	✓	RN50 FPN	17.71
CL-HOI (Ours)	✗	RN50+Tr.	26.74
CL-HOI (Ours)	✗	RN50+Tr.	**30.07**

Table 3: The comparison of our CL-HOI with weakly-supervised methods. Different from Table. 2, we also evaluate our CL-HOI with well-annotated captions from V-COCO.

3rd epoch. The batch size is set to 16, and the model is trained for the five epochs.

4.2 Comparison to State-of-the-Art Methods

The effectiveness of our proposed method is benchmarked through comparative analysis with current state-of-the-art (SOTA) methods in fully and weakly supervised Human-Object Interaction (HOI) detection. As shown in Table 1 and Table 2, our proposed CL-HOI outperforms existing weakly-supervised methods on the HICO-DET and V-COCO datasets without the need for manual annotations. On the HICO-DET dataset, our CL-HOI, when using Swin-L as the backbone, achieves the best performance among weakly-supervised methods, reaching 17.55% mAP. On the V-COCO dataset, CL-HOI with Swin-L outperforms all weakly-supervised methods, achieving 36.63% Role AP. Even when adopting ResNet-50 as the backbone, CL-HOI demonstrates competitive performance compared to existing weakly-supervised methods on both HICO-DET and V-COCO datasets. For fully supervised comparisons, our CL-HOI shows performance comparable to existing methods. Notably, CL-HOI outperforms the baseline methods on both HICO-DET and V-COCO datasets. Overall, CL-HOI excels in HOI detection with minimal supervision, surpassing even some fully supervised counterparts.

4.3 Weakly-Supervised CL-HOI

To robustly demonstrate the effectiveness of our CL-HOI framework, we conduct a comprehensive evaluation on the V-COCO dataset under weakly-supervised settings, which is particularly notable for utilizing well-annotated captions. These captions serve as a rich source of supervision for our framework. The performance of CL-HOI under these settings is compared with other existing weakly-supervised methods and CL-HOI variants. This comparative analysis is detailed in Table 3. As the comparisons indicate, our CL-HOI outperforms state-of-the-art methods under weakly-supervised settings. For a fair comparison, we include IO-HOI, which is trained using the well-annotated Conceptual Captions dataset (Sharma et al., 2018). Moreover, even when trained on annotation-free captions, our CL-HOI continues to outperform existing methods.

Prompting			HICO-DET	V-COCO
Spatial	Visual	Context	mAP	Role AP
✗	✗	✗	9.42	24.21
✓	✗	✗	11.24	25.99
✓	✓	✗	14.67	26.45
✓	✓	✓	**15.62**	**26.74**

Table 4: Ablation study on the impact of chain-of-cognition cognition in interaction cognition network on HICO-DET and V-COCO datasets based on our baseline UPT.

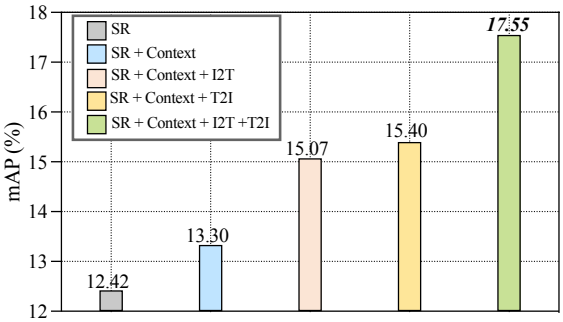

Figure 3: Ablation study on the impact of contrastive distillation losses on HICO-DET dataset.

4.4 Ablation Study

Impact of Different Teacher VLLM Models. To evaluate the impact of different teacher models on HOI detection, we performed an ablation study using two VLLMs: BLIP-2 and InstructBLIP. As shown in Table 5, BLIP-2 with the RN50+Transformer backbone achieved an mAP of 9.63%, which improved to 12.98 % when using the more powerful Swin-L+Transformer backbone. InstructBLIP consistently outperformed BLIP-2, with an mAP of 15.62% using RN50+Tr. and 17.55% with Swin-L+Transformer. These results highlight the effectiveness of switching the teacher model from BLIP-2 to InstructBLIP, especially with stronger backbones.

Impact of the Chain-of-Cognition Strategy. First, we assess the impact of spatial cognition on HOI detection performance. Before evaluation, we use the baseline to train with the interaction distillation losses. Table 4 shows that the baseline yields a mAP of 9.42% and a Role AP of 24.21%. We

VLLM	Bacbone	mAP
BLIP-2 (Li et al., 2023)	RN50+Tr.	9.63
	Swin-L+Tr.	12.98
InstructBLIP (Dai et al., 2023)	RN50+Tr.	15.62
	Swin-L+Tr.	17.55

Table 5: Ablation study of different teacher models on the HICO-DET dataset. We evaluate two teacher models based on our CL-HOI.

Datasets	Training	Backbone	mAP	Role AP
HICO-DET	w/o training	RN50+Tr.	5.3	-
	w/ training	RN50+Tr.	15.62	-
V-COCO	w/o training	RN50+Tr.	-	15.40
	w/ training	RN50+Tr.	-	26.74

Table 6: Performance comparisons for HICO-DET and V-COCO datasets with training and without training (random initialization).

then add the spatial cognition to the baseline. The results show that spatial cognition improves the mAP by 1.82% and the Role AP by 1.78%, validating its effectiveness in providing crucial geometric information for HOI detection. We then investigate the impact of visual cognition on the HOI detection performance. We add visual cognition to spatial cognition as the second chain. The results in Table 4 show that visual cognition improves the mAP by 3.43% and the Role AP by 0.46%. The improvements demonstrate that the visual features are essential for HOI detection. Lastly, we evaluate the impact of context cognition on CL-HOI. We introduce context cognition after visual cognition as the final chain. Table 4 shows that context cognition improves the mAP to 15.62% and the Role AP to 26.74%. This confirms that contextual cues are also instrumental in accurately reasoning HOIs.

Impact of the Contrastive Distillation Losses. We also perform ablation studies on our proposed contrastive distillation losses in Fig. 3 using Swin-L as the backbone. First, we applied only the soft-relation loss to the model, resulting in a mAP of 12.42%. Next, we added the context loss, which improved the mAP to 13.30%. Following that, we incorporated the image-to-text and text-to-image losses, contributing additional improvements of 1.77% and 2.10%, respectively. Finally, when all the losses were combined, the model achieved a mAP of 17.55%. These incremental improvements demonstrate that our proposed contrastive distillation losses significantly enhance the detection performance, making them a valuable component of our HOI detection methodology.

Random Initialization. We compare the performance of the trained model with a randomly initialized one on two benchmark datasets, as shown in Table 6. On HICO-DET, random initialization yields 5.30%, improving to 15.62% after training. Similarly, Role AP improves from 15.40 % to 26.74% with our framework. These results demonstrate the effectiveness of cross-level distillation in transferring image-level interaction reasoning to instance-level HOI detection.

4.5 Qualitative Analysis

The visualization in Fig. 4 highlights CL-HOI's improved ability to detect human-object interactions compared to the baseline. In the first comparison, CL-HOI correctly identifies "throw", while the baseline mislabels it as "catch". Similarly, CL-HOI detects "hit" in the second group, which the baseline misses. In the third set, CL-HOI accurately identifies no interaction with a bicycle, unlike the baseline, which mistakenly detects "sit".

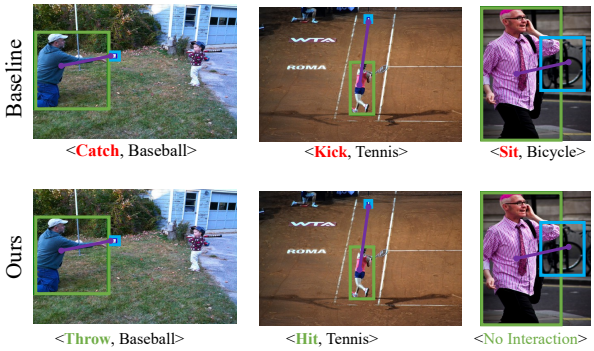

Figure 4: Qualitative comparisons of HOI detection performance between our baseline and CL-HOI.

5 Conclusion

We introduce CL-HOI, a human-object interaction detector that reduces reliance on manual annotations by distilling instance-level detection from VLLMs. Using a two-stage context and interaction distillation framework, we leverage a VLLM as a teacher model for cross-level knowledge transfer through a Visual-Linguistic Translator (VLT) and Interaction Cognition Network (ICN). Contrastive distillation losses further enhance this transfer. CL-HOI achieves competitive results on V-COCO and HICO-DET, showcasing the effectiveness of our distillation approach.

6 Limitations

Our proposed CL-HOI aims to distill instance-level HOI detection from VLLMs, which are able to reason and recognize image-level human-object interactions. However, there are two bottlenecks for such cross-level distillation: the recognition accuracy of VLLMs and the relation ambiguities. The VLLMs fail to accurately recognize the image-level HOIs since they are not fine-tuned on HOI-related datasets. Thus, the image-level supervision are not as accurate as supervision in fully and weakly supervised methods. On the other hand, the image-level generated supervision is blind to the matching relations with the human-object pairs in the image. These two bottlenecks lead to an insufficient performance when evaluating on HICO-DET and V-COCO datasets.

References

Amjad Alkilani and Amir Shirkhodaie. 2013. Acoustic signature recognition technique for human-object interactions (hoi) in persistent surveillance systems. In *Signal Processing, Sensor Fusion, and Target Recognition XXII*, volume 8745, pages 459–469. SPIE.

Silvia Bucci, Antonio D'Innocente, Yujun Liao, Fabio M Carlucci, Barbara Caputo, and Tatiana Tommasi. 2021. Self-supervised learning across domains. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5516–5528.

Yichao Cao, Qingfei Tang, Xiu Su, Song Chen, Shan You, Xiaobo Lu, and Chang Xu. 2024. Detecting any human-object interaction relationship: Universal hoi detector with spatial prompt learning on foundation models. *Advances in Neural Information Processing Systems*, 36.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. 2018. Learning to detect human-object interactions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.

Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, 37(8):729–758.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales. 2022. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62.

Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. 2018. Pairwise body-part attention for recognizing human-object interactions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 51–67.

Chen Gao, Yuliang Zou, and Jia-Bin Huang. 2018. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*.

Jianjun Gao, Kim-Hui Yap, Kejun Wu, Duc Tri Phan, Kratika Garg, and Boon Siew Han. 2024. Contextual human object interaction understanding from pre-trained large language model. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13436–13440. IEEE.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Saurabh Gupta and Jitendra Malik. 2015. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*.

Roohollah Jahanmahin, Sara Masoud, Jeremy Rickli, and Ana Djuric. 2022. Human-robot interactions in manufacturing: A survey of human behavior modeling. *Robotics and Computer-Integrated Manufacturing*, 78:102404.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 74–83.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19556–19565.

Suresh Kirthi Kumaraswamy, Miaojing Shi, and Ewa Kijak. 2021. Detecting human-object interaction with mixed supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1228–1237.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. 2021. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34:309–323.

Ting Lei, Shaofeng Yin, Yuxin Peng, and Yang Liu. 2024. Exploring conditional multi-modal prompts for zero-shot hoi detection. *arXiv preprint arXiv:2408.02484*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. 2020. Hoi analysis: Integrating and decomposing human-object interaction. *Advances in Neural Information Processing Systems*, 33:5011–5022.

Zhuolong Li, Xingao Li, Changxing Ding, and Xiangmin Xu. 2024. Disentangled pre-training for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28191–28201.

Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. 2020. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490.

Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. 2022. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20123–20132.

JunYi Lim, Vishnu Monn Baskaran, Joanne Mun-Yee Lim, KokSheik Wong, John See, and Massimo Tistarelli. 2023. Ernet: An efficient and reliable human-object interaction detection network. *IEEE Transactions on Image Processing*, 32:964–979.

Jeeseung Park, Jin-Woo Park, and Jong-Seok Lee. 2023. Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17152–17162.

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*, pages 8748–8763. PMLR.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. 2020. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13617–13626.

Mesut Erhan Unal and Adriana Kovashka. 2023. Weakly-supervised hoi detection from interaction labels only and language/vision-language priors. *arXiv preprint arXiv:2303.05546*.

Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. 2019. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9469–9478.

Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. 2022. Learning transferable human-object interaction detector with natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 939–948.

Jie Yang, Bingliang Li, Ailing Zeng, Lei Zhang, and Ruimao Zhang. 2024. Open-world human-object interaction detection via multi-modal prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16954–16964.

Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. 2023. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23497–23506.

Hangjie Yuan, Shiwei Zhang, Xiang Wang, Samuel Albanie, Yining Pan, Tao Feng, Jianwen Jiang, Dong Ni, Yingya Zhang, and Deli Zhao. 2023. Rlipv2: Fast scaling of relational language-image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21649–21661.

Frederic Z Zhang, Dylan Campbell, and Stephen Gould. 2021. Spatially conditioned graphs for detecting human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13319–13327.

Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. 2022. 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 (CVPR)*, pages 20104–20112.

Frederic Z Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. 2023. Exploring predicate visual context in detecting of human-object interactions. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision, pages 10411–10421.

Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. 2021. End-to-end human object interaction detection with hoi transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11825–11834.

A Appendix

A.1 Vision Large Language Model

In this paper, we introduce the InstructBlip (Dai et al., 2023) as the teacher model. In this section, we elaborate on the prompting implementation of the teacher model. We first prompt the teacher VLLM to generate the image-level context captions describing human activities. Given the input image, we use the IntructBlip "What are the persons doing in the image?" as $prompt_1$ to generate *context captions*. Based on the context captions, we further prompt the LLM (Vicuna (Chiang et al., 2023)) in the InstructBlip to parse all possible human-object interactions. Specifically, we prompt "Briefly list human-object interactions in a triplet of <interaction, object> from the sentence." as $prompts_2$ to parse interactions from *context captions*. Then, we use the *context captions* and *interaction triplets* to supervise the learning of our proposed visual-linguistic translator and interaction cognition network. This process enables the context and interaction distillation and transfers the interaction recognition and reasoning capabilities from the VLLM to our proposed CL-HOI. We give an example to show how we generate the captions and parse the verbs in Fig. 5. The final outputs of the VLLM are given in "*Swing, Baseball Bat.", "*Catch, Baseball.". We need some simple postprocessing to remove unnecessary punch marks.

A.2 Negative Verb Generation

Providing the verbs parsed from the context captions, we identify these verbs as positive verbs. In cases where a caption lacks an action verb, we omit both the caption and its corresponding image from further consideration. This selective approach ensures that our analysis focuses only on images with captions that effectively describe actions. The negative verb generation process begins with constructing a verb dictionary, the contents of which are illustrated in Fig. 6. Notably, there is no overlap between the dictionary and ground truth verbs in both V-COCO and HICO-DET datasets. For each image, we randomly select two verbs from this dictionary that are not present in the positive verbs list, labeling these as negative verbs. Therefore, the data for training comprise context captions and positive and negative verbs for each input image.

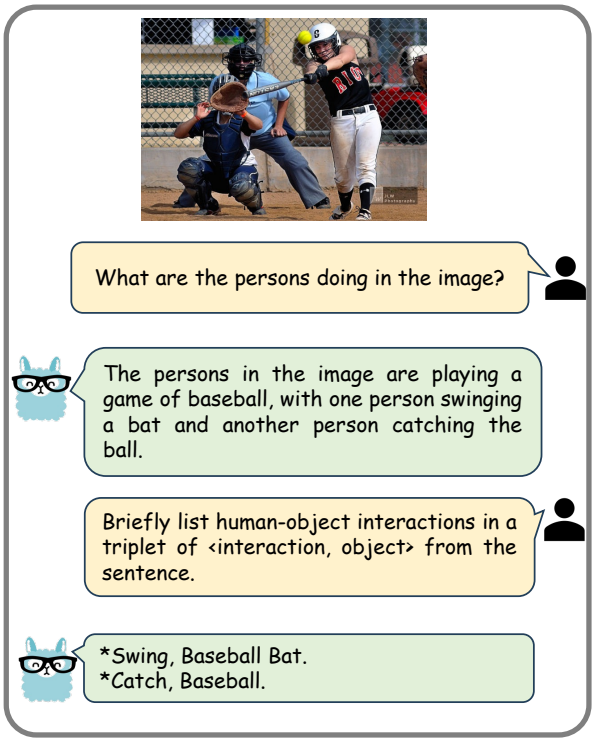

Figure 5: The example of the prompting implementation of the teacher VLLM.

have, do, say, get, go, know, take, see, come, think, want, give, use, find, tell, ask, like, call, try, need, feel, become, leave, put, mean, keep, let, begin, seem, help, start, show, hear, play, live, believe, bring, happen, write, provide, meet, include, continue, learn, change, lead, understand, follow, create, speak, allow, add, spend, grow, win, offer, remember, love, consider, appear, wait, send, expect, build, stay, fall, reach, remain, suggest, raise, pass, sell, require, report, decide

Figure 6: The verbs in the dictionary that are used to generate negative verbs in interaction distillation loss. For each valid image, we randomly choose two verbs from the dictionary as negatives.

A.3 Soft-Relation Loss

In this paper, we also adopt the soft-relation loss in our interaction distillation loss. The soft-relation loss uses pseudo labels to supervise the interaction learning. Specifically, we first organize the detected human and object into union bounding boxes and the crop corresponding regions from images. After that, we use the CLIP vision encoder to encode these union regions as $f_u \in \mathbb{R}^{K \times D}$. On the other hand, we use the text encoder to encode posi-

Figure 7: Examples of context captions and interactions generated from the teacher VLLM.

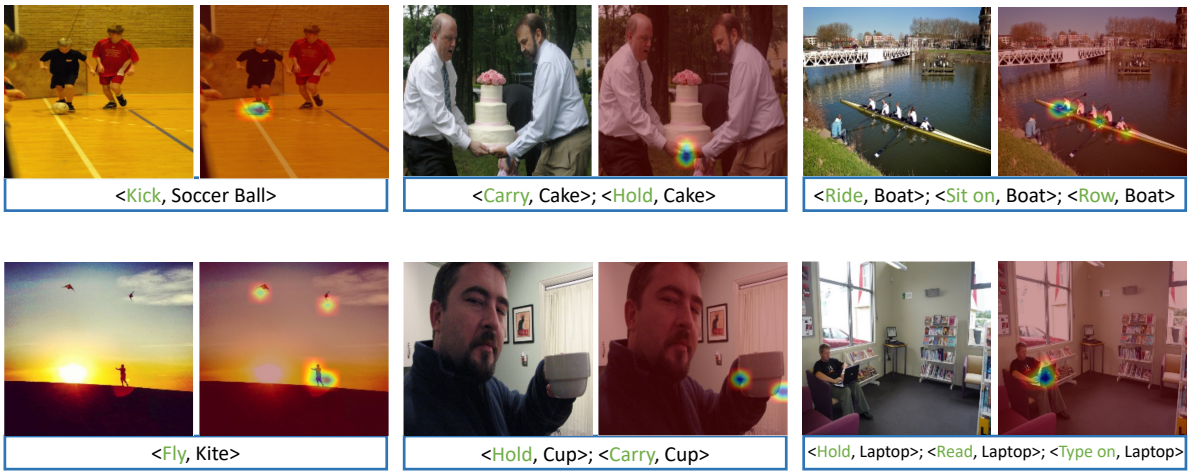

Figure 8: Visualized attention maps from CL-HOI, demonstrating the correctness and effectiveness of the context and interaction distillation. The six examples illustrate that CL-HOI accurately attends to the regions where interactions occur.

tive verbs and objects from the parsed interaction triplets as $f_{pos} \in \mathbb{R}^{P\times D}$ and $f_{obj} \in \mathbb{R}^{P\times D}$. Then, we calculate the similarities between f_u and f_{pos} as well as f_{obj} respectively as:

$$sim_1 = \frac{f_u \cdot f_{pos}^\top}{\|f_u\|\|f_{pos}\|^\top}, sim_2 = \frac{f_u \cdot f_{obj}^\top}{\|f_u\|\|f_{obj}\|^\top}, \quad (11)$$

where $\|\cdot\|$ denotes to row-wise l_2 norm, and $sim_1, sim_2 \in \mathbb{R}^{K\times P}$. The final pseudo-labels Y_s are generated from these two similarities:

$$\tilde{sim}_1[k,p] = \begin{cases} 1 & \text{if } sim_1[k,p] \geq 0.5, \\ 0 & \text{if } sim_1[k,p] < 0.5, \end{cases} \quad (12)$$

$$\tilde{sim}_2[k,p] = \begin{cases} 1 & \text{if } sim_2[k,p] \geq 0.5, \\ 0 & \text{if } sim_2[k,p] < 0.5, \end{cases} \quad (13)$$

$$Y_s = \tilde{sim}_1 \odot \tilde{sim}_2, \quad (14)$$

where $k \in K$ and $p \in P$, and $Y_s \in \mathbb{R}^{K\times P}$.

Given the pseudo labels, we then calculate the soft-relation loss. Firstly, we calculate the similarities between interaction features I and encoded positive verbs f_{pos} as the final prediction as $\hat{Y}$ as

$$\hat{Y} = \frac{I \cdot f_{pos}^{\top}}{\|I\| \|f_{pos}\|^{\top}}. \tag{15}$$

The final soft-relation loss is calculated as follows:

$$L_{sr} = -\sum_{k=1}^{K} \sum_{p=1}^{P} \left[Y_{k,p} \log(\hat{Y}_{k,p}) + (1 - Y_{k,p}) \log(1 - \hat{Y}_{k,p}) \right]. \tag{16}$$

A.4 Inference

In the testing and inference phases of our approach, we employ a method that centers around the use of interaction features, denoted as I, to predict the interaction class, represented as y. This is achieved through a process of calculating the similarity between I and each of the encoded interaction labels within Y. The calculation can be presented as:

$$\hat{Y} = \frac{I \cdot Y^{\top}}{\|I\| \|Y\|^{\top}}, \tag{17}$$

$$s_i = \text{sigmoid}(\hat{y}_i')^{(1-\lambda)} \cdot (s_o \cdot s_h)^{\lambda}, \tag{18}$$

where λ is a balancing factor that helps in appropriately weighing the contributions.

A.5 Label Generated from Teacher Model

Since the VLLM only provides image-level interaction labels, the grounding between weak labels and detected humans or objects is missing, unlike fully supervised settings. Additionally, the teacher model's supervision signals are incomplete, making the task harder than typical weakly supervised methods. We compare the labels from the original HICO-DET dataset with our VLLM-generated labels for the top 30 interaction categories in Fig. 9. The results show a significant gap, particularly for categories with more instances, where original labels are nearly double our generated ones.

A.6 Qualitative Analysis

Generated Context Captions and Interactions. In Fig. 7, we show more examples of the context captions and interactions generated by teacher VLLM on the HICO-DET dataset. From the examples, it is evident that the VLLM is capable of recognizing image-level human activities and reasoning human-object interactions. For instance, the VLLM recognizes "The persons in the first example are flying kites in a grassy field." and parses the "fly, kite" from the generated caption.

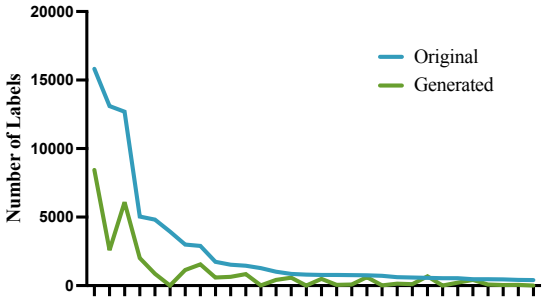

Figure 9: The quantitative comparisons of the top thirty labels with respect to numbers from the original HICO-DET dataset and labels generated from the VLLM.

Attention Visualization. We include more visualization results of the attention map of the last layer of VLT in Fig. 8. The results demonstrate that CL-HOI is able to capture context and interaction information to aid in attending to interactive areas, even when objects are not close to humans.