Supplementary Materials of
Re-mine, Learn and Reason: Exploring the Cross-modal Semantic Correlations for Language-guided HOI detection

The supplementary materials are organized as follows. In Appendix A, we elaborate on the motivations behind the development of the RmLR framework. In Appendix B, we present a more detailed description of our architecture. In Appendix C, we outline the datasets and evaluation metrics used in our experiments. In Appendix D, we provide an in-depth explanation of the training and inference procedures. In Appendix E, we discuss additional cases that demonstrate the interaction loss phenomenon. In Appendix F, we examine the effects of varying the number of layers in different modules. In Appendix G, we implement the Interaction Relation Encoder using a pre-trained human pose detection model and assess its performance. In Appendix H, we explore the connection between our RmLR method and other CLIP-based approaches. In Appendix I, we present additional detection results for further analysis.

A. Motivations for Our RmLR Framework

An effective HOI detector must concurrently handle both object detection and interaction relation recognition tasks. The latter imposes a more substantial requirement on the model's capability to comprehend visual features. Moreover, optimizing the model by solely mapping the <$person, action, object$> combinations in HOI datasets [16][6] to one-hot labels presents challenges due to the flexibility and diversity inherent in these annotations.

In recent years, several studies have investigated the integration of language prior knowledge from text to guide the learning of HOI models [66][21][35][57][60]. Incorporating linguistic modality information has led to modest improvements in the performance of existing HOI methods. However, a majority of these approaches employ a CLIP-like technique to condense the textual semantic features of multiple interaction actions into a fixed-length vector [47]. For set prediction problems such as HOI, this compression strategy imposes limitations on the transfer of cross-modal knowledge.

Consequently, we propose a novel cross-modal HOI detection framework that enhances visual feature extraction and cross-modal learning efficiency from two perspectives:

- Firstly, we perform a qualitative and quantitative analysis of the interaction information loss issue in two-stage visual HOI detectors. We provide supplementary examples in Appendix E to corroborate our observations. To tackle this problem, we introduce the Interactive Relation Encoder (IRE), designed to **re-mine** visual features specifically for HO interaction recognition.
- Secondly, considering that HOI prediction involves set prediction tasks, we introduce sentence- and word-level alignment strategies to facilitate effective cross-modal **learning** and ensure knowledge transfer from linguistic modalities.

By incorporating these richer multi-modal representations, we can ultimately achieve improved HOI recognition performance.

B. More Implementation Details

The Visual Feature Extractor and Entity Detection module of our RmLR are based on ResNet [17] and DETR[5], respectively. For the Interactive Relation Encoder and Interaction Reasoning Module, we use 2 and 1 Transformer encoder layers, respectively. We follow the two-stage HOI detector training paradigm[63], where we first pre-train DETR on a large-scale image dataset and then fine-tune it on HICO-DET and V-COCO datasets. The weights of DETR remain frozen during fine-tuning. To initialize the network for HICO-DET, we use DETR pre-trained on MS COCO [38]. However, for V-COCO, we exclude some of COCO's training images that are contained in the V-COCO test set when pre-training DETR. We use an FC layer to map the global context features to 512-dimensional vectors. Similarly, we use an FC layer to map the output of IRE's interactive feature to the same dimension (512). For the spatial features (entity tokens), we concatenate human and object tokens to construct a 1024-dimensional vector.

We employ the data augmentation and preprocessing techniques proposed in [63]. Specifically, we resize the input images such that the shorter side is within the range of 480 to 800 pixels and the longer side is limited to 1333 pixels. In

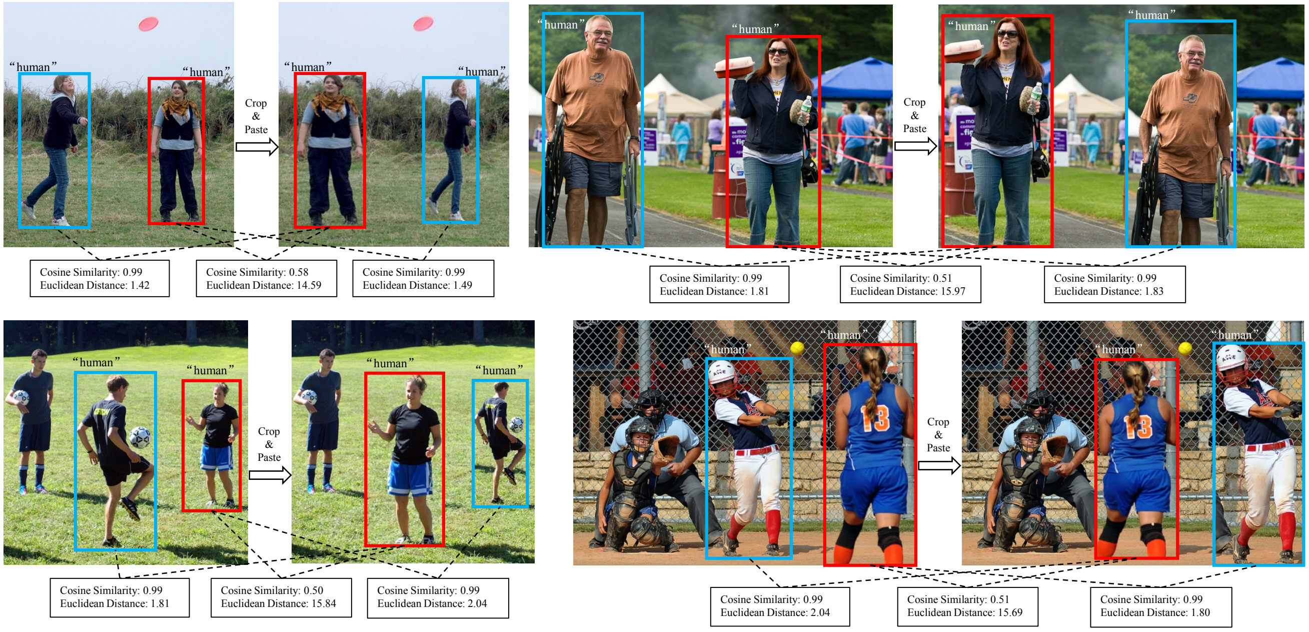


Figure 6. We provide further examples to elucidate the phenomenon of interaction information loss in two-stage Transformer-based HOI detectors. Figure 6 showcases instances from the HICO-DET and V-COCO datasets, where we evaluate the output tokens of DETR [5] using both cosine similarity and Euclidean distance metrics. Our results corroborate earlier observations that the output tokens of the detection model predominantly pertain to spatial positioning and object categories, rather than the interaction information. This is exemplified by the fact that individuals situated in the same position exhibit similar features, regardless of the actions they perform.

Table 8. Effect of the #Layers of Different Modules on the V-COCO test set. "CML-SA" indicates self-attention layers in cross-modal learning.

#Layer		V-COCO	
IRM	CML-SA	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
1	1	63.71	69.76
1	2	63.78	69.81
2	1	63.59	69.62
2	2	63.75	69.77

our cross-modal learning approach, we use two self-attention layers and one cross-attention layer, with a hidden state dimension of 1024. We set $\gamma = 0.2$ and $\beta = 0.5$ for the Focal loss, following [63]. To determine new hyper-parameters, we perform cross-validation. We use the Adam optimizer with an initial learning rate of 10^{-4} and cosine learning rate decay strategy. Our model is trained with a batch size of 8 for 20 epochs on four 3080 GPUs.

C. Details of Datasets and Evaluation Metrics

V-COCO [16]. V-COCO is a popular dataset for benchmarking HOI detection, which is built upon the MS-COCO dataset. The mean average precision (mAP) is used for evaluation. For object occlusion cases, two evaluation scenarios are considered. Scenario 1 ($AP^{\#1}_{role}$) considers a strict evaluation criterion that requires the prediction of a null bounding box with coordinates [0, 0, 0, 0], Scenario 2 ($AP^{\#2}_{role}$) relaxes this condition for such cases by ignoring the predicted bounding box for evaluation.

HICO-DET [6]. We follow the previous methods [34] to evaluate on the HICO-DET. The mAP metric is computed in $Default\ settings$ and $Known\,Objects\,Setting$ for three categories: **Full** (all 600 HOI classes), **Rare** (138 classes that have less than 10 training samples), **Non-rare** (462 classes that have more than 10 training samples). Here the $Default\ setting$ represents that the mAP is calculated over all testing images, while $Known\,Object\,Setting$ measures the AP of each object solely over the images containing that object class.

D. Details of Training and Inference

To guarantee the effectiveness and efficiency of our approach, we systematically design three stages that ensure robust visual feature extraction and successful cross-modal knowledge transfer: (i) **Re-mining Visual Interaction-Relevant Features**: This stage employs a visual feature extractor and the IRE module to capture low-level features and model interactive relations; (ii) **Cross-Modal Alignment for Visual and Textual Representations**: This stage devises sentence- and word-level alignment strategies to establish correlations between the semantic information of different modalities; (iii) **Reasoning Using Linguistic Knowledge**:

Algorithm 1: The training and inference process of RmLR framework.

```
Input: Pre-trained object detector, pre-trained text
       encoder [52], maximum training epochs $N$.
Init $\tau = 0$;
Initialize and freeze the weights of $\mathcal{F}_{ED}$ with
  pre-trained object detector weights;
while $\tau \leq N$ do
    1. Learning Visual Features
    (1) Extract the low-level features $\mathcal{X}^v$ for input $\mathcal{I}$;
    (2) Flatten and project the $\mathcal{X}^v$ into $\mathcal{Z}^v$;
    (3) Entity Detection:
      $(\mathcal{S}^v, \mathcal{B}^v, \mathcal{C}^v) = \mathcal{F}_{ED}\left(\mathcal{Z}^v, \mathcal{Q}_o\right)$;
    (4) Exhaustively generate HO pairs and filter
      away invalid combinations;
    (5) Obtaining pair-wise token features $\tilde{S}^v$;
    (6) Interactive relation modeling via Transformer
      encoder layer: $\mathcal{X}_e^v = \mathcal{F}_{enc}\left(\mathcal{X}^v\right)$;
    (7) Masked RoI operation is adopted to generate
      union region features $m^v$;
    (8) Calculate the global context feature $g^v$;
    (9) Concatenate the $[g^v, \tilde{s}^v, m^v]$ to obtain overall
      visual features for HO pairs;
    2. Learning Cross-modal Content
    (1) Serialize annotation labels as sentence $\mathcal{T}$;
    (2) Tokenize the $\mathcal{T}$ into $\mathcal{Z}^l$ and then map to $\mathcal{X}^l$;
    (3) Calculate the $[CLS]$ tokens and word
      embeddings: $(\mathcal{E}_{cls}, \mathcal{E}^w) = \mathcal{F}_{TE}\left(\mathcal{X}^l\right)$;
    (4) Self-attention for the $\mathcal{M}^v$ and $\mathcal{O}^v$;
    (5) Associate the HO candidates with annotations
      to obtain the $\mathcal{M}^{va}$ and $\mathcal{O}^{va}$;
    (6) Cross-alignment for the two modality
      representations to obtain $\widehat{\mathcal{M}}^{va}$ and $\widehat{\mathcal{O}}^{va}$;
    (7) Calculate the L1 loss $\mathcal{L}^m$ and $L^a$ for IRM and
      IRE, respectively;
    3. Reasoning Using Knowledge
    (1) Reasoning using linguistic knowledge
      enhanced visual features and logits;
    (2) Calculate the overall loss;
    (3) Optimize the learnable weights of RmLR;
end
Output: The optimized weights of RmLR.
```

This stage utilizes an interaction reasoning module to integrate visual and linguistically-enhanced representations.

In this section, we present a comprehensive pseudo-code that outlines the training and inference procedures of RmLR in Algorithm 1. The three stages within this pseudo-code correspond to the three phases previously discussed. For the sake of simplicity, we exclude the training process of the object detection model in the first stage.

E. More Cases about Interaction Information Loss Phenomenon

In the main paper, we have proposed that two-stage Transformer-based HOI detectors tend to lose interactive information. In this section, we present additional evidence to support this claim. Figure 6 shows more examples, where we measure the output tokens of DETR [5] not only with cosine similarity but also with Euclidean distance. The results obtained using Euclidean distance also support the conclusion drawn in Figure 1, that the output tokens of the detection model are only related to position information. These results further reinforce the claim that the two-stage HOI detectors suffer from a loss of interactive information.

F. Selection of #Layers of Different Modules

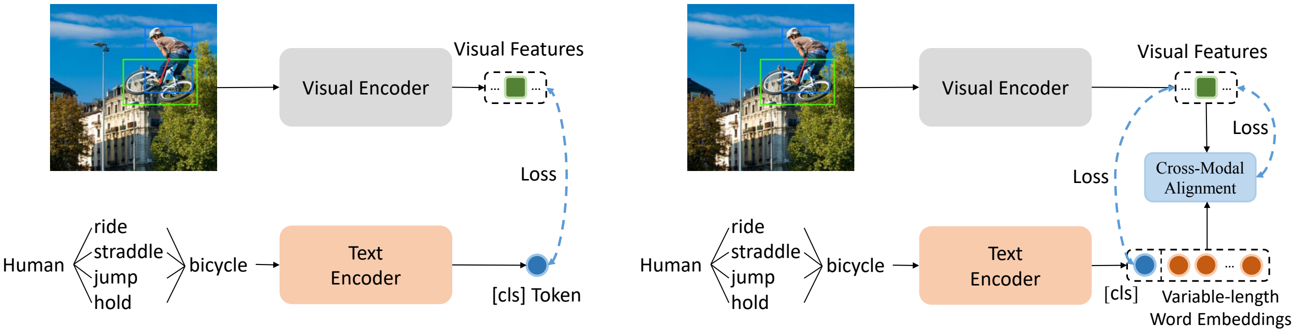
In this section, we present a comprehensive comparison of the number of layers among various models. To fully unleash the potential of our method, we also conducted experiments to compare the performance of different numbers of layers in the IRM module, as presented in Table 8. Moreover, we also investigated the effect of varying the number of self-attention layers in cross-modal learning. Our results demonstrate that the performance improvement of the model is constrained by only increasing the number of layers in IRM and CML-SA.

G. Modified IRE module using Human Pose Information

In Section 3.2 and Table 2, we presented the need for re-mining interaction-relevant information in two-stage HOI detectors. In our proposed RmLR framework, the IRE module is a learnable component for interactive relationship modeling, gradually acquiring the ability to capture HO interaction cues under HOI annotation and textual semantic information supervision. Furthermore, we replace the IRE module with an explicit human posture recognition model to learn the union interaction feature of HO candidates. This model is pre-trained on the CrowPose [30] dataset, and we freeze its weight for model training and reasoning as an explicit interaction learning module. We conducted comparative experiments on ResNet-50 and ResNet-101-based RmLR on two datasets, and the results in Table 9 show that the IRE module trained with additional datasets can further enhance the RmLR framework. This finding also confirms the necessity of re-mining interaction features from a different perspective.

H. Relationship between RmLR with other HOI-VLM Methods

As discussed in Section 2.3, current state-of-the-art HOI-VLM methods can be categorized into two groups: VLP-based and knowledge distillation-based approaches. VLP



(a) Previous [cls] token-based methods

(b) Our cross-modal alignment and transfer method

Figure 7. The HOI task involves predicting multiple interaction categories for one human-object pair, making it a set prediction problem. Our RmLR approach employs a more refined knowledge transfer operation compared to the previous HOI-VLM method, which ensures the effectiveness and efficiency of cross-modal learning of HOI detector.

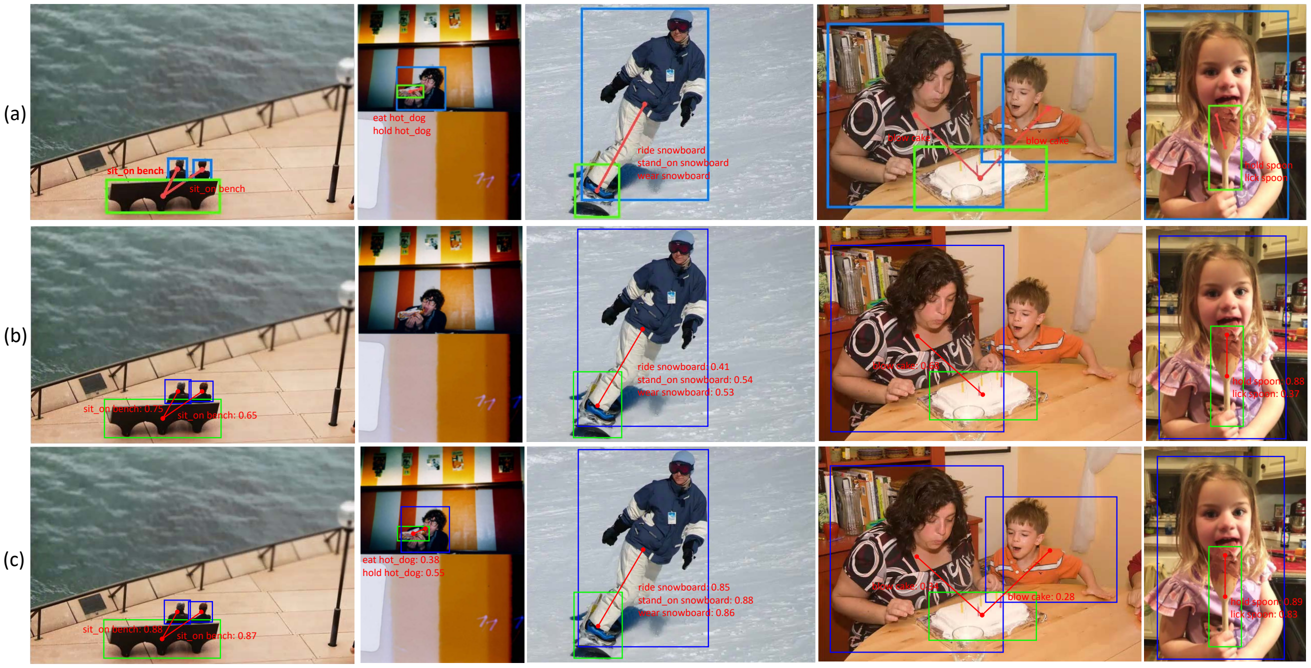


Figure 8. Visualization of HOI annotations and detection results from UPT [63] and proposed RmLR method. From top to bottom, the images depict the ground truth annotations, UPT results, and our results, respectively. These comparisons reveal that UPT suffers from false negative and low confidence results. In contrast, our RmLR method achieves more accurate and confident HOI detection results.

methods typically rely on large-scale Vision-and-Language datasets for cross-modal pre-training and fusion of text and image features. In contrast, our proposed method falls under the knowledge distillation category, which enhances the optimization of visual models by transferring knowledge from pre-trained language models.

Our approach is innovative in two key aspects compared to existing HOI-VLM methods:

Firstly, most current knowledge distillation-based methods are somewhat simplistic, such as some CLIP-based HOI detection methods [57][37]. These methods directly map annotation text to a fixed-length feature vector and use it to guide the visual model in learning semantic information. While they have achieved some success in exploring HOI-VLM, they still suffer from several drawbacks that need to be addressed. The HOI task is essentially a set prediction problem, where an image may contain multiple HO pairs with various interactions within each pair. Our experimental results in Section 4.4 and Table 2 demonstrate that simply compressing the semantic information of these interactions into a fixed-length sentence representation (*i.e.*, [cls] tokens) limits the effectiveness of HOI recognition. This approach

Table 9. Performance comparison on the V-COCO test set. The “Extra Dataset" represent the dataset other than HOI datasets.

		HICO-DET (Default Setting)			V-COCO	
Backbone	Extra Dataset	Full	Rare	Non-rare	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
ResNet-50	-	36.93	29.02	39.29	63.78	69.81
ResNet-50	CrowdPose	37.15	30.18	40.23	63.93	69.97
ResNet-101	-	37.41	28.81	39.97	64.17	70.23
ResNet-101	CrowdPose	**38.29**	**31.05**	**40.37**	**64.38**	**70.45**

constrains the full utilization and effective transfer of linguistic information. Therefore, implementing cross-modal alignment and association from text to visual modality is essential to ensure the successful transfer of linguistic prior knowledge to the visual model.

Secondly, our method differs from the general VLP approach because a large vision-and-language dataset is not required in the training process. The training of the HOI detector can be completed solely through efficient fine-tuning and knowledge transfer on the HOI dataset. Furthermore, our RmLR method exhibits exceptionally high training efficiency on HOI datasets. Based on a four 3080 GPU server, the training process of the ResNet-50-based RmLR model takes only about 1.5 hours on the V-COCO dataset and about 12 hours on the HICO-DET dataset.

I. Visualization for the HOI Detection

We present visualizations of HOI annotations and detection results on the HICO-DET [6] test set in Figure 8. The annotations in (a) demonstrate that an image may contain multiple HO pairs, and various interactions may occur within a single HO pair. Therefore, HOI detectors must predict an HO pair and interaction category set. The detection results of the UPT and our RmLR methods are shown in (b) and (c), respectively. These results reveal that the UPT method [63] is susceptible to false negatives and low confidence results. Even for some obvious interactions, the UPT method produces highly fluctuating prediction confidence. In contrast, our method achieves more accurate results for both HO pair and interaction category set prediction. These visualizations reinforce the quantification results presented in Table 1, suggesting that our RmLR framework possesses a significantly stronger interaction understanding capability.