PROGALIGN: PROGRESSIVE SEMANTIC ALIGNMENT WITH VISUAL CHAIN-OF-THOUGHT REASONING FOR OPEN-VOCABULARY HOI DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary human-object interaction (OV-HOI) detection is a fundamental task for human-centric scene understanding. Existing end-to-end HOI methods often suffer from insufficient fine-grained visual semantics mining and limited reasoning-process interpretability. To address these challenges, we propose ProgAlign, an end-to-end framework that integrates visual chain-of-thought reasoning with a progressive semantic alignment mechanism. Specifically, a visual chain-of-thought module is designed to construct a multi-step coarse-to-fine HOI evidence chain across spatial, contact, and interaction stages, enabling the model to progressively capture fine-grained interaction cues, which in turn helps improve the discrimination of visually similar interactions. Furthermore, a progressive semantic alignment mechanism is introduced to align Large Language Model-derived stage-specific semantic priors with intermediate visual representations, thereby helping alleviate black-box reasoning issues and enhancing the semantic consistency of the intermediate reasoning process. Extensive experiments on the SWIG-HOI and HICO-DET datasets demonstrate that ProgAlign achieves state-of-the-art performance among comparable open-vocabulary methods, validating its effectiveness in advancing OV-HOI understanding and distinguishing complex and fine-grained interactive relations.

1 INTRODUCTION

Human-Object Interaction (HOI) detection aims to localize interacting human-object pairs and identify their specific interaction categories, which is essential for tasks such as action recognition (Miao et al., 2025), human-computer interaction (Robinson et al., 2023), and anomaly detection (Yang et al., 2025). The key to HOI detection is to jointly model who interacts with what, how they are spatially related, and what action is being performed.

Existing HOI detection methods can be broadly categorized into two-stage and one-stage paradigms. Two-stage methods (Lei et al., 2023; 2024a; Wang et al., 2024; Cao et al., 2023) first perform human-object detection and then conduct interaction recognition, but suffer from high computational overhead due to exhaustive pair enumeration. In contrast, one-stage methods (Ning et al., 2023; Hu et al., 2025; Lei et al., 2024b; 2025; Zhong et al., 2021; Wang et al., 2022; 2020) unify detection and interaction reasoning into an end-to-end framework for more efficient HOI prediction as depicted in Fig. 1(a).

Existing one-stage approaches can be broadly divided into point-based and query-based methods. Point-based methods (Zhong et al., 2021; Wang et al., 2020) localize interaction centers and regress offsets, but can be sensitive to centroid ambiguity in complex scenes. Query-based methods (Ning et al., 2023; Hu et al., 2025; Lei et al., 2024b; 2025; Wang et al., 2022) instead exploit Transformer queries for global relational reasoning. With the success of CLIP (Radford et al., 2021), recent methods further introduce vision–language alignment for open-vocabulary generalization.

CLIP-based OV-HOI methods have achieved promising results through semantic alignment. THID (Wang et al., 2022) introduces natural-language supervision, while MP-HOI (Yang et al., 2024) learns transferable representations through multimodal contrastive alignment. CMD-SE (Lei et al., 2024b) and SGC-Net (Lin et al., 2025) further enhance fine-grained interaction modeling

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

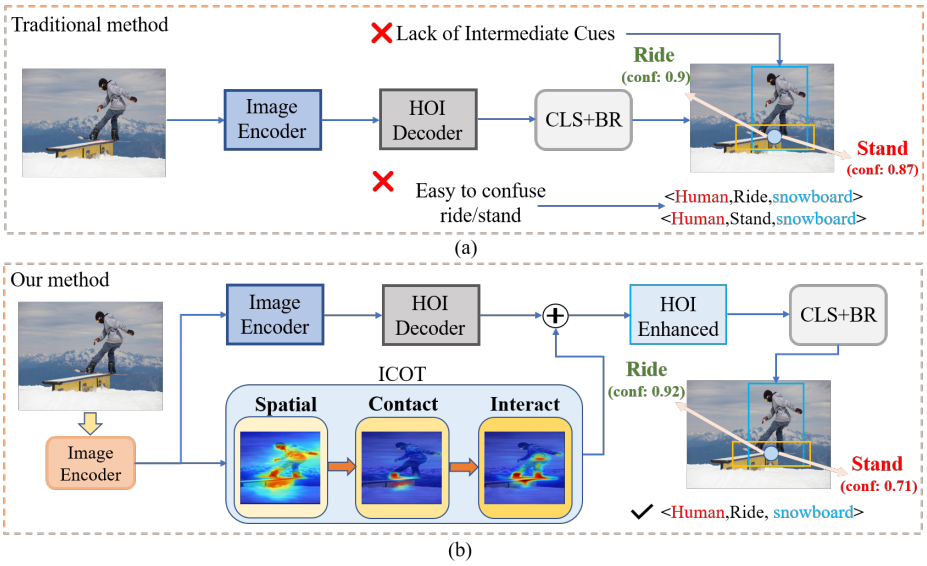


Figure 1: Comparison of (a) a conventional one-stage HOI detector and (b) ProgAlign. The conventional paradigm directly predicts interaction labels from visual features, whereas ProgAlign introduces spatial, contact, and interaction reasoning to provide explicit intermediate evidence for distinguishing similar interactions.

through multi-level decoding and granular semantic comparison. However, these methods mainly emphasize final or global semantic matching, leaving fine-grained local cues and intermediate reasoning representations insufficiently constrained.

To address these limitations, we propose ProgAlign, an end-to-end framework that combines visual chain-of-thought reasoning with progressive semantic alignment (PSA). An Image Chain-of-Thought (I-CoT) module decomposes HOI recognition into spatial, contact, and interaction stages, progressively mining fine-grained evidence from human–object layouts, contact patterns, and interaction-specific regions. In parallel, PSA aligns LLM-derived stage-specific semantic priors with the corresponding intermediate visual representations, providing explicit semantic guidance throughout the reasoning process.

The main contributions are summarized as follows:

- We propose an Image Chain-of-Thought (I-CoT) module that organizes HOI reasoning into spatial, contact, and interaction stages to progressively capture fine-grained interaction evidence.
- We introduce progressive semantic alignment (PSA), which aligns intermediate I-CoT representations with corresponding LLM-derived semantic priors to improve semantic consistency across reasoning stages.
- Extensive experiments on SWIG-HOI (Wang et al., 2021) and HICO-DET (Chao et al., 2018) demonstrate that ProgAlign achieves state-of-the-art performance among comparable open-vocabulary methods.

2 Related Work

2.1 OV-HOI Detection

OV-HOI detection aims to transfer relational knowledge to novel action–object compositions while distinguishing semantically similar interactions. Existing methods improve generalization through compositional learning, relational pre-training, and vision–language knowledge transfer. THID (Wang et al., 2022) uses decomposed interaction labels as textual supervision, RLIP (Yuan

et al., 2022) transfers external relational knowledge, and recent VLM-based methods such as CMD-SE (Lei et al., 2024b), SGC-Net (Lin et al., 2025), and INP-CC (Lei et al., 2025) enrich interaction representations with fine-grained semantics or prompts. However, semantic alignment is typically concentrated on final prediction representations, leaving intermediate interaction evidence weakly constrained. ProgAlign instead aligns stage-specific semantic priors with intermediate I-CoT representations.

2.2 Chain-of-Thought in Computer Vision

Chain-of-Thought (CoT) (Wei et al., 2022) decomposes complex reasoning into intermediate steps and has been extended to visual and multimodal tasks. ViperGPT (Surís et al., 2023) uses executable visual programs, while Multimodal-CoT (Zhang et al., 2023), IdealGPT (You et al., 2023), and DDCoT (Zheng et al., 2023) introduce language-based or staged multimodal reasoning. These approaches mainly express intermediate reasoning as language or programs and focus on general multimodal tasks. In contrast, I-CoT realizes a coarse-to-fine spatial–contact–interaction evidence chain directly within an HOI detector.

3 Methodology

As illustrated in Fig. 2, ProgAlign contains a visual reasoning branch, an LLM-enhanced text branch, and progressive semantic alignment. I-CoT organizes visual reasoning into spatial, contact, and interaction stages, while stage-specific LLM priors constrain their intermediate representations during training. Category-level textual features are finally matched with the enhanced visual representations for HOI prediction.

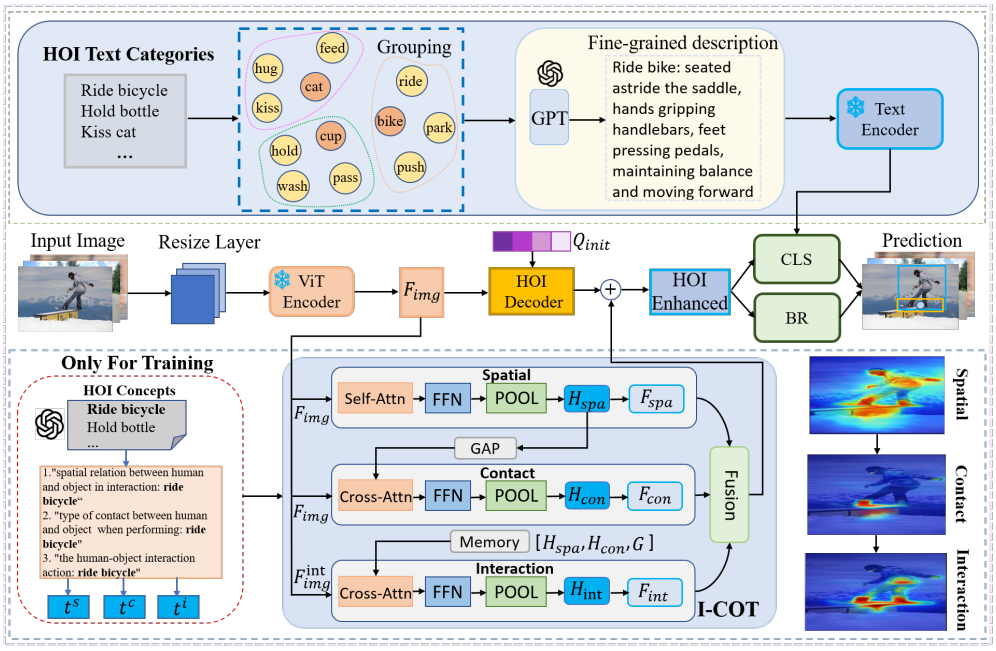


Figure 2: The overall architecture of the proposed ProgAlign framework. The pipeline integrates a three-stage I-CoT reasoning module and a standard HOI decoder for final triplet prediction. "Fimg", "GAP", "CLS", and "BR" denote visual features, global average pooling, the interaction classifier, and the bounding box regressor, respectively.

3.1 Visual Chain-of-Thought Reasoning

We instantiate visual chain-of-thought reasoning with an Image Chain-of-Thought (I-CoT) module that decomposes HOI understanding into three successive stages: spatial topology reasoning,

physical contact verification, and semantic interaction inference. Each stage conditions subsequent reasoning, forming a coarse-to-fine evidence chain from global spatial relations to local contact cues and interaction semantics.

Given the visual feature sequence $F_{\text{img}} \in \mathbb{R}^{B\times N\times C}$ produced by the image encoder, where B, N, and C denote the batch size, number of visual tokens, and feature dimension, respectively, I-CoT consists of three cascaded Transformer-based StageBlocks. The spatial stage adopts self-attention to model global dependencies among visual regions, whereas the contact and interaction stages employ cross-attention to incorporate preceding reasoning results. The three StageBlocks share the same architecture but use independent parameters. Each StageBlock outputs a pooled stage representation $H_k \in \mathbb{R}^{B\times C}$, where $k \in \{\text{spa}, \text{con}, \text{int}\}$. Implementation and token-modulation details are provided in Appendix A.3.

In the spatial topology reasoning stage, the model captures the relative positions and global configurations of humans and objects:

$$H_{\text{spa}} = \text{SB}_{\text{spa}}\left(F_{\text{img}}\right), \tag{1}$$

where $\text{SB}_{\text{spa}}(\cdot)$ denotes the spatial StageBlock, and H_{spa} encodes the global human–object spatial configuration for subsequent contact reasoning.

In the physical contact verification stage, H_{spa} serves as the spatial condition for learning contact-related representations:

$$H_{\text{con}} = \text{SB}_{\text{con}}\left(F_{\text{img}}, H_{\text{spa}}\right), \tag{2}$$

where H_{con} represents physical contact evidence conditioned on the global human–object layout. The spatial and contact attention responses are aggregated and normalized into patch-wise importance vectors $w_{\text{spa}}, w_{\text{con}} \in \mathbb{R}^{B\times N}$.

In the semantic interaction inference stage, the two importance vectors are combined as $w_{\text{joint}} = \alpha w_{\text{spa}} + (1-\alpha)w_{\text{con}}$, where α controls their relative contributions. The interaction-stage visual sequence is constructed as $F_{\text{img}}^{\text{int}} = F_{\text{img}} \odot w_{\text{joint}}$, where $\odot$ denotes token-wise feature reweighting. We further obtain $G = \text{GAP}(F_{\text{img}})$ and construct $Memory = [H_{\text{spa}}; H_{\text{con}}; G]$. The interaction representation is then computed as

$$H_{\text{int}} = \text{SB}_{\text{int}}\left(F_{\text{img}}^{\text{int}}, Memory\right), \tag{3}$$

where H_{int} integrates multi-level visual evidence for high-level interaction inference.

Finally, the pooled hidden states from the three reasoning stages are concatenated and projected into image-conditioned HOI queries:

$$Q_{\text{init}} = \text{MLP}\left(H_{\text{spa}} \oplus H_{\text{con}} \oplus H_{\text{int}}\right), \tag{4}$$

where $\oplus$ denotes feature concatenation. The projection produces $N_q C$ outputs and reshapes them into $Q_{\text{init}} \in \mathbb{R}^{B\times N_q\times C}$, where $N_q = 25$. These image-conditioned queries are added to the learnable HOI query embeddings before the decoder retrieves interaction evidence from F_{img}.

Overall, I-CoT progressively captures spatial structure, physical contact, and interaction semantics to improve the discrimination of visually similar interactions.

3.2 Progressive Semantic Alignment

Although I-CoT progressively models human–object interactions, detection objectives alone cannot explicitly guarantee that each intermediate stage learns semantics consistent with its designated reasoning role. We therefore introduce progressive semantic alignment to provide stage-specific constraints and improve semantic consistency across the reasoning trajectory.

For each interaction category $y \in \mathcal{Y}$, we use the gpt-oss-20b language model (OpenAI, 2025) to generate three textual descriptions corresponding to the relative spatial configuration, physical contact pattern, and interaction semantics. Each query focuses on the semantic aspect associated with its corresponding I-CoT stage. The complete prompt templates, model-serving settings, and representative descriptions are provided in Appendix B.

The generated responses are independently encoded by a frozen pre-trained CLIP text encoder to obtain the stage-wise semantic priors $\mathbf{t}_y^{\text{spa}}$, $\mathbf{t}_y^{\text{con}}$, and $\mathbf{t}_y^{\text{int}}$. The language model is used only for offline textual-prior construction, and all descriptions and encoded features are cached before detector training.

Since an image may contain multiple interaction categories, for the interaction-label set $\mathcal{Y}_i$ of the i-th sample, the image-level semantic target at stage k is constructed as

$$\bar{\mathbf{t}}_i^k = \mathrm{Norm}\left(\frac{1}{|\mathcal{Y}_i|}\sum_{y\in\mathcal{Y}_i}\mathbf{t}_y^k\right), \tag{5}$$

where $k \in \{\mathrm{spa}, \mathrm{con}, \mathrm{int}\}$, $|\mathcal{Y}_i|$ denotes the number of interaction categories in the i-th image, and $\mathrm{Norm}(\cdot)$ denotes L_2 normalization. The aggregated target summarizes the stage-level semantics of multiple interactions and serves as an image-level auxiliary constraint.

Meanwhile, each StageBlock produces a stage-level hidden state H_i^k, which is linearly projected and normalized as $\mathbf{f}_i^k = \mathrm{Norm}(W_k H_i^k)$ in the shared vision–language space, where W_k is a learnable stage-specific projection into the CLIP text embedding space. For a training batch of size B, the similarity between the i-th visual representation and the j-th textual target at stage k is computed as

$$\mathbf{S}_{ij}^k = \frac{\langle \mathbf{f}_i^k, \bar{\mathbf{t}}_j^k \rangle}{\tau}, \tag{6}$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity between normalized features and τ is a temperature parameter.

A row-wise Softmax is applied to obtain $p_{ij}^k = \exp(\mathbf{S}_{ij}^k)/\sum_{l=1}^{B}\exp(\mathbf{S}_{il}^k)$. The stage-level visual-to-text contrastive loss is defined as

$$\mathcal{L}_k = -\frac{1}{B}\sum_{i=1}^{B}\log p_{ii}^k, \tag{7}$$

where p_{ii}^k denotes the matching probability between the stage-level visual representation and the textual target from the same sample. This objective guides each reasoning stage to learn semantic information consistent with its designated role.

The alignment losses of the three reasoning stages are averaged to obtain the overall progressive semantic alignment objective:

$$\mathcal{L}_{\mathrm{icot}} = \frac{1}{3}\left(\mathcal{L}_{\mathrm{spa}} + \mathcal{L}_{\mathrm{con}} + \mathcal{L}_{\mathrm{int}}\right). \tag{8}$$

Progressive semantic alignment explicitly associates the spatial, contact, and interaction stages with human–object configurations, physical contact patterns, and high-level action semantics, respectively. Since the textual priors are pre-computed and the alignment objective is used only during training, it introduces no additional inference cost.

3.3 LLM-Enhanced Text Branch

The original HOI category names generally contain only coarse interaction semantics, which may be insufficient for distinguishing categories with similar actions or visual appearances. We therefore construct an LLM-enhanced text branch to generate more discriminative category-level textual representations.

Specifically, K-means clustering (MacQueen, 1967) organizes semantically related HOI categories into groups with similar actions, objects, or interaction meanings. For each target category, its name and the remaining categories in the same group are provided to the OpenAI gpt-oss-20b model (OpenAI, 2025), served through Ollama Cloud. The model generates a fine-grained description emphasizing observable interaction cues that distinguish the target category from its related counterparts. For example, the description of "ride bicycle" characterizes the human as sitting on the saddle, gripping the handlebars, pedaling, maintaining balance, and moving forward. Prompt templates and representative descriptions are provided in Appendix B.

The generated descriptions are encoded by a frozen pre-trained CLIP text encoder to obtain category-level textual features. All descriptions and their features are generated offline and cached before detector training. During HOI prediction, these textual features are matched with the enhanced visual representations produced by the HOI decoder and I-CoT module, and the resulting visual–text similarities are used for interaction classification.

The LLM-enhanced text branch and progressive semantic alignment serve different purposes: the former provides category-level textual features for final interaction classification, whereas the latter uses stage-specific semantic priors to align intermediate representations only during training.

3.4 Training and Inference

3.4.1 Training

Following previous HOI detection methods (Liao et al., 2022; Lei et al., 2025), we adopt the basic detection objective

$$\mathcal{L}_{\text{det}} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{b}}\mathcal{L}_{\text{b}} + \lambda_{\text{giou}}\mathcal{L}_{\text{giou}} + \lambda_{\text{conf}}\mathcal{L}_{\text{conf}}, \tag{9}$$

where $\mathcal{L}_{\text{cls}}$, $\mathcal{L}_{\text{b}}$, $\mathcal{L}_{\text{giou}}$, and $\mathcal{L}_{\text{conf}}$ denote classification, bounding box regression, GIoU, and confidence losses, respectively.

The overall training objective combines the detection loss with the progressive semantic alignment loss:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_{\text{icot}}\mathcal{L}_{\text{icot}}, \tag{10}$$

where λ_{icot} controls the contribution of the progressive semantic alignment objective.

3.4.2 Inference

Given the interaction classification score $s_{i,c}$ and bounding box confidence c_i, the final HOI prediction score is computed as

$$s'_{i,c} = s_{i,c} \cdot (c_i)^{\gamma}, \tag{11}$$

where γ is a scaling factor used to suppress low-quality predictions during inference (Zhang et al., 2022; 2021).

4 Experiments

4.1 Experimental Setting

Datasets. We evaluate the proposed ProgAlign framework on two widely used OV-HOI benchmarks: HICO-DET (Chao et al., 2018) and SWIG-HOI (Wang et al., 2021), on which prior OV-HOI works (Wang et al., 2022; Lei et al., 2024b; 2025; Lin et al., 2025) are exclusively evaluated. HICO-DET contains 47,776 images with 117,871 annotated HOI instances covering 600 interaction categories. Following the zero-shot setting adopted by previous OV-HOI work (Lin et al., 2025; Lei et al., 2025), we hold out 120 rare interactions as unseen categories and use the remaining 480 interactions as seen categories. SWIG-HOI covers more than 400 human actions and 1,000 object categories, and its test set contains approximately 14,000 images and 5,500 interactions.

Evaluation Metric. Following previous works (Liao et al., 2022; Lei et al., 2024b; Mao et al., 2023; Lin et al., 2025; Lei et al., 2025), we adopt mean Average Precision (mAP) as the evaluation metric and report performance on seen categories, unseen categories, and the full interaction set. A predicted HOI triplet is considered a true positive if the following two conditions are simultaneously satisfied: (1) the Intersection-over-Union (IoU) between both the predicted human/object bounding boxes and the corresponding ground-truth boxes is greater than 0.5, (2) the predicted interaction category is correct.

Implementation Details. We implement ProgAlign using PyTorch on two NVIDIA A800 GPUs. Images are resized to 224×224, and the visual encoder is initialized with ViT-B/16 pre-trained weights. The model is trained for 120 epochs using AdamW (Loshchilov & Hutter, 2017), with an initial learning rate of 1×10^{-4}, weight decay of 5×10^{-5}, 10 warm-up epochs, and cosine decay to 1×10^{-7}. The effective batch size is 96. We use 25 HOI queries, 6 attention heads, and a hidden dimension of 768. We set $\tau = 0.07$, $\lambda_{\text{icot}} = 0.1$, and $\alpha = 0.5$. Further implementation details are provided in Appendix A.

Table 1: Performance comparison on the HICO-DET dataset (mAP %).

Method	Pre-trained Detector	Unseen	Seen	Full
Zero-shot Methods				
SCL (Iyer et al., 2022)	✓	19.07	30.39	28.08
GEN-VLKT (Liao et al., 2022)	✓	21.36	32.91	30.56
HOICLIP (Ning et al., 2023)	✓	23.48	34.47	32.26
CLIP4HOI (Mao et al., 2023)	✓	<u>28.47</u>	**35.48**	**34.08**
HOIGen (Guo et al., 2024)	✓	**31.01**	<u>34.57</u>	<u>33.86</u>
Open-vocabulary Methods				
THID (Wang et al., 2022)	×	15.53	24.32	22.38
CMD-SE (Lei et al., 2024b)	×	16.70	23.95	22.35
INP-CC (Lei et al., 2025)	×	17.38	24.74	23.12
SGC-Net (Lin et al., 2025)	×	<u>23.27</u>	<u>28.34</u>	<u>27.22</u>
ProgAlign (Ours)	×	**23.58**	**28.71**	**27.63**

Table 2: Performance comparison on the SWIG-HOI dataset (mAP %).

Method	Non-rare	Rare	Unseen	Full
CHOID (Wang et al., 2021)	10.93	6.63	2.64	6.64
QPIC (Tamura et al., 2021)	16.95	10.84	6.21	11.12
GEN-VLKT (Liao et al., 2022)	20.91	10.41	–	10.87
MP-HOI (Yang et al., 2024)	20.28	14.78	–	12.61
THID (Wang et al., 2022)	17.67	12.82	10.04	13.26
CMD-SE (Lei et al., 2024b)	21.46	14.64	10.70	15.26
INP-CC (Lei et al., 2025)	22.84	<u>16.74</u>	11.02	16.74
SGC-Net (Lin et al., 2025)	<u>23.67</u>	16.55	<u>12.46</u>	<u>17.20</u>
ProgAlign (Ours)	**24.02**	**17.91**	**12.58**	**18.11**

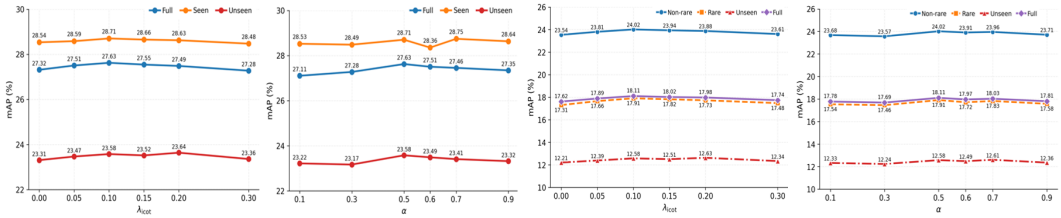
4.2 Comparison with Other Methods

Results and analysis on HICO-DET. We compare our method with the state-of-the-art approaches under the zero-shot setting on HICO-DET (Table 1). Although recent zero-shot methods (Liao et al., 2022; Guo et al., 2024; Mao et al., 2023; Iyer et al., 2022; Ning et al., 2023) utilize CLIP text embeddings for interaction classification, they rely on DETR models pre-trained on COCO (Gupta & Malik, 2015), limiting scalability in open-world scenarios. Moreover, this comparison is inherently biased since HICO-DET and COCO share identical object categories. Open-vocabulary methods alleviate this limitation by avoiding pre-training on detection datasets. Our ProgAlign surpasses SGC-Net (Lin et al., 2025) by 0.41 mAP points in Full mAP, achieving the best Full mAP among the compared open-vocabulary methods. Together with the ablation results, this gain supports the complementary contributions of progressive I-CoT reasoning and progressive semantic alignment. Aligning textual semantics with intermediate visual features reduces ambiguity in the learned representations and improves fine-grained interaction discrimination.

Results and analysis on SWIG-HOI. We further evaluate the proposed method on the more challenging SWIG-HOI dataset. As shown in Table 2, ProgAlign achieves **18.11%** Full mAP, outperforming the previous state-of-the-art method SGC-Net (Lin et al., 2025). Meanwhile, the results also indicate that methods relying on pre-trained detectors, such as MP-HOI (Yang et al., 2024), suffer from limited scalability as the vocabulary space expands, resulting in inferior performance in open-vocabulary scenarios. These results further demonstrate the effectiveness of the proposed I-CoT mechanism. Unlike conventional frameworks that rely on coarse global semantic alignment, I-CoT decomposes complex interactions into a structured progressive reasoning process, enabling more effective mining of fine-grained interaction semantics from localized visual representations.

Consequently, the proposed method captures subtle contact cues more accurately and improves generalization ability in OV-HOI scenarios.

4.3 Parameter Sensitivity Analysis



(a) (b) (c) (d)

Figure 3: Parameter sensitivity on HICO-DET and SWIG-HOI. (a,c) Progressive semantic alignment weight λ_{icot}; (b,d) spatial–contact fusion coefficient α. (a,b) HICO-DET; (c,d) SWIG-HOI.

We investigate the sensitivity of ProgAlign to the progressive semantic alignment weight λ_{icot} and the spatial–contact fusion coefficient α on HICO-DET and SWIG-HOI. Only the investigated parameter is varied while the remaining settings are fixed.

Progressive semantic alignment weight. As shown in Fig. 3, increasing λ_{icot} from 0 to 0.1 improves overall performance, with Full mAP on HICO-DET increasing from 27.32% to 27.63%. Further increasing the weight gradually reduces Full and Seen performance, suggesting that overly strong alignment may interfere with the primary detection objective. Although $\lambda_{\text{icot}} = 0.2$ slightly improves Unseen mAP, $\lambda_{\text{icot}} = 0.1$ provides the best overall balance.

Spatial–contact fusion coefficient. The coefficient α balances spatial and contact evidence. The best Full mAP is obtained at $\alpha = 0.5$ on both datasets, indicating that balanced global spatial structure and local contact cues are most effective. Performance remains relatively stable around this setting, and we therefore use $\alpha = 0.5$ in all experiments.

Overall, the results show that ProgAlign is robust to moderate hyperparameter variations, with $\lambda_{\text{icot}} = 0.1$ and $\alpha = 0.5$ providing the strongest overall performance.

4.4 Ablation Studies

We conduct controlled ablations on SWIG-HOI using SGC-Net (Lin et al., 2025) with its GSA and HGC modules removed as the baseline. All variants retain the same LLM-enhanced category descriptions and frozen CLIP text encoder.

Table 3: Ablation studies on SWIG-HOI (mAP %). (a) Effect of the main components. (b) Effect of progressive I-CoT reasoning stages.

(a) Main components

Configuration	Non-rare	Rare	Unseen	Full
Baseline	15.69	11.53	7.32	11.45
+ I-CoT	23.43	16.69	11.64	17.15
+ PSA	22.41	16.02	10.97	16.30
+ I-CoT + PSA	**24.02**	**17.91**	**12.58**	**18.11**

(b) Progressive I-CoT stages (without PSA)

Configuration	Non-rare	Rare	Unseen	Full
Baseline	15.69	11.53	7.32	11.45
+ Spatial	21.96	15.84	10.98	15.91
+ Contact	23.05	**17.12**	11.52	16.76
+ Interaction	**23.43**	16.69	**11.64**	**17.15**

Main components. As shown in Table 3, introducing I-CoT substantially improves Full mAP from 11.45% to 17.15%, with consistent gains across all category splits, demonstrating the benefit of progressive spatial–contact–interaction reasoning. PSA alone also achieves a clear improvement to 16.30% Full mAP, showing that stage-specific semantic guidance is effective even without the complete reasoning module. Combining I-CoT and PSA further increases Full mAP to 18.11% and achieves the best results on all splits, indicating that progressive visual reasoning and semantic alignment provide complementary benefits.

Progressive reasoning stages. With PSA disabled, adding the spatial stage raises Full mAP from 11.45% to 15.91%, confirming the importance of explicitly modeling global human–object configurations. Incorporating contact reasoning further improves Full mAP to 16.76% and yields the highest Rare mAP, suggesting that local physical-contact cues are particularly useful for low-frequency interactions. The complete three-stage I-CoT achieves the best Non-rare, Unseen, and Full mAP of 23.43%, 11.64%, and 17.15%, respectively. Although Rare mAP decreases slightly after adding interaction reasoning, the overall improvement supports the complete coarse-to-fine reasoning design.

4.5 Qualitative Analysis

To qualitatively examine the progressive reasoning process of I-CoT and the effect of progressive semantic alignment (PSA), Fig. 4 compares attention heatmaps of the full model and its variant without PSA across snowboarding, riding a bicycle, and kayaking. With PSA, the spatial stage captures the overall human–object configuration, the contact stage focuses more clearly on key physical interfaces, and the interaction stage highlights interaction-discriminative regions, revealing a progressive transition from global relations to local contact evidence and interaction semantics. Without PSA, the attention responses become more diffuse or shifted toward less relevant regions, particularly around contact interfaces, suggesting that stage-specific semantic guidance helps clarify the role of each reasoning stage and improves localization of fine-grained interaction evidence.

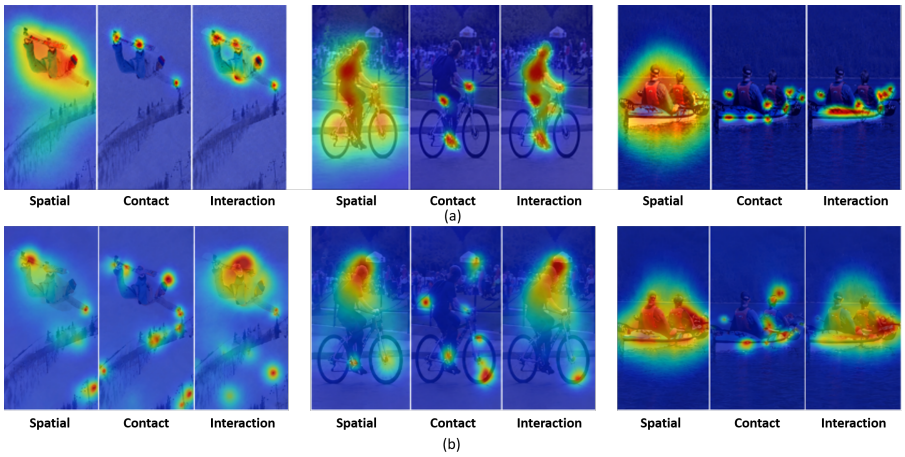


Figure 4: Three-stage attention heatmaps with (a) and without (b) PSA.

Fig. 4 shows a progressive transition from global human–object layouts to contact interfaces and interaction-discriminative regions. Without PSA, attention is more diffuse or shifted, qualitatively supporting the role of stage-specific semantic guidance.

5 Conclusion

We propose ProgAlign for OV-HOI detection, coupling three-stage visual chain-of-thought reasoning with progressive semantic alignment. I-CoT organizes spatial, contact, and interaction evidence, while PSA aligns intermediate representations with stage-specific semantic priors. Experiments on HICO-DET and SWIG-HOI demonstrate consistent gains over comparable open-vocabulary methods.

AI USE STATEMENT

We used the OpenAI gpt-oss-20b model, served through Ollama Cloud, to generate the category-level descriptions used by the final text branch and the stage-specific textual descriptions used as semantic priors. The generation procedure, prompts, output processing, and offline caching are documented in Appendix B. Generative AI was also used to assist with manuscript editing and formatting. All AI-assisted content was reviewed and verified by the authors. The authors take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The main paper describes the model, objectives, datasets, training setup, evaluation protocol, ablation studies, parameter-sensitivity analysis, and qualitative results. Appendix A reports execution, training, and implementation details; Appendix B documents textual-knowledge generation; and Appendix C gives the exact parameter-sensitivity values.

REFERENCES

Yichao Cao, Qingfei Tang, Feng Yang, Xiu Su, Shan You, Xiaobo Lu, and Chang Xu. Re-mine, learn and reason: Exploring the cross-modal semantic correlations for language-guided hoi detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 23492–23503, 2023.

Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pp. 381–389. IEEE, 2018.

Yixin Guo, Yu Liu, Jianghao Li, Weimin Wang, and Qi 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*, pp. 1711–1720, 2024.

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

Yupeng Hu, Changxing Ding, Chang Sun, Shaoli Huang, and Xiangmin Xu. Bilateral collaboration with large vision-language models for open vocabulary human-object interaction detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 20126–20136, 2025.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022.

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

Ting Lei, Fabian Caba, Qingchao Chen, Hailin Jin, Yuxin Peng, and Yang Liu. Efficient adaptive human-object interaction detection with concept-guided memory. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6480–6490, 2023.

Ting Lei, Shaofeng Yin, and Yang Liu. Exploring the potential of large foundation models for open-vocabulary hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16657–16667, 2024b.

Ting Lei, Shaofeng Yin, Qingchao Chen, Yuxin Peng, and Yang Liu. Open-vocabulary hoi detection with interaction-aware prompt and concept calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23945–23957, 2025.

Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si 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*, pp. 20123–20132, 2022.

Xin Lin, Chong Shi, Zuopeng Yang, Haojin Tang, and Zhili Zhou. Sgc-net: stratified granular comparison network for open-vocabulary hoi detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4539–4549, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

J MacQueen. Multivariate observations. In *Proceedings ofthe 5th Berkeley symposium on mathematical statisticsand probability*, volume 1, pp. 281–297. University of California press Oakland, CA, USA, 1967.

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

Qiguang Miao, Wentian Xin, Ruyi Liu, Yi Liu, Mengyao Wu, Cheng Shi, and Chi-Man Pun. Adaptive pitfall: Exploring the effectiveness of adaptation in skeleton-based action recognition. *IEEE Transactions on Multimedia*, 27:56–71, 2025. doi: 10.1109/TMM.2024.3521774.

Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming 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*, pp. 23507–23517, 2023.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025.

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

Nicole Robinson, Brendan Tidd, Dylan Campbell, Dana Kulić, and Peter Corke. Robotic vision for human-robot interaction and collaboration: A survey and systematic review. *ACM Transactions on Human-Robot Interaction*, 12(1):1–66, 2023.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11888–11898, 2023.

Masato Tamura, Hiroki Ohashi, and Tomoaki 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*, pp. 10410–10419, 2021.

Guangzhi Wang, Yangyang Guo, Ziwei Xu, and Mohan Kankanhalli. Bilateral adaptation for human-object interaction detection with occlusion-robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 27970–27980, 2024.

Suchen Wang, Kim-Hui Yap, Henghui Ding, Jiyan Wu, Junsong Yuan, and Yap-Peng Tan. Discovering human interactions with large-vocabulary objects via query and multi-scale detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13475–13484, 2021.

Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 939–948, 2022.

Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4116–4125, 2020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Dejie Yang, Zijing Zhao, and Yang Liu. Planllm: Video procedure planning with refinable large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9166–9174, 2025.

Jie Yang, Bingliang Li, Ailing Zeng, Lei Zhang, and Ruimao Zhang. Open-world human-object interaction detection via multi-modal prompts. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 16954–16964, 2024.

Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. Idealgpt: Iteratively decomposing vision and language reasoning via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11289–11303, 2023.

Hangjie Yuan, Jianwen Jiang, Samuel Albanie, Tao Feng, Ziyuan Huang, Dong Ni, and Mingqian Tang. Rlip: Relational language-image pre-training for human-object interaction detection. *Advances in Neural Information Processing Systems*, 35:37416–37431, 2022.

Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13319–13327, 2021.

Frederic Z Zhang, Dylan Campbell, and Stephen 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*, pp. 20104–20112, 2022.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023.

Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13234–13243, 2021.

ProgAlign: Progressive Semantic Alignment with Visual Chain-of-Thought Reasoning for Open-Vocabulary HOI Detection

CONTENTS

A Evaluation and Reproducibility Details 14

A.1 Experimental Runs and Execution Environment 14

A.2 Training and Model Configuration 14

A.3 StageBlock Implementation Details 14

B Textual Knowledge Construction 15

B.1 Language Model and Offline Generation 15

B.2 HOI Category Grouping 16

B.3 Category-Level Description Prompt 16

B.4 Stage-Specific Semantic Prompts 17

B.5 Output Processing and Offline Caching 17

B.6 Representative Generated Descriptions ..... **??**

C Exact Parameter Sensitivity Values 18

A Evaluation and Reproducibility Details

A.1 Experimental Runs and Execution Environment

Unless otherwise stated, results produced by our implementation are averaged over four complete training and evaluation runs. All controlled comparisons use the same dataset split and training protocol, and only the explicitly named component or hyperparameter is changed. Results quoted from previous publications follow their original reporting protocols. The base random seed is set to 22. During distributed training, the process with rank r uses $22 + r$, and the corresponding seed is applied to the Python, NumPy, and PyTorch random number generators.

All experiments are conducted on an HPC cluster. Each experimental run uses two NVIDIA A800 80 GB GPUs allocated from a CentOS Linux 7 node equipped with an Intel Xeon Gold 6348 processor, 56 CPU cores, 1,024 GB of system memory, and eight NVIDIA A800 GPUs. The principal hardware and software settings are summarized in Table 4.

Table 4: Execution environment, random-seed configuration, and number of experimental runs.

Setting	Configuration
Platform	HPC cluster
Operating system	CentOS Linux 7
CPU	Intel Xeon Gold 6348
CPU cores per node	56
System memory	1,024 GB per node
GPU node configuration	Eight NVIDIA A800 GPUs
GPUs used per run	2
GPU memory	80 GB per GPU
System CUDA	12.1
Python	3.8.20
PyTorch	1.8.0+`cu111`
torchvision	0.9.0+`cu111`
PyTorch CUDA build	11.1
cuDNN	8.0.5
Compiler	GCC 4.8.5 (Red Hat 4.8.5-44)
Base random seed	22
Runs for implemented settings	4

The complete Python dependency list, with package versions verified under Python 3.8.20, will be released together with the source code.

A.2 Training and Model Configuration

Table 5 summarizes the principal training and model configurations. The visual encoder is initialized with ViT-B/16 pre-trained weights, while the CLIP text encoder remains frozen. Optimization is performed with AdamW for 120 epochs using two distributed training processes. The effective global batch size is 96.

The learning-rate schedule consists of 10 warm-up epochs followed by cosine decay from an initial learning rate of 1×10^{-4} to a minimum learning rate of 1×10^{-7}. Repeat-factor sampling and focal classification loss are enabled.

A.3 StageBlock Implementation Details

The three I-CoT StageBlocks share the same architecture but use independent parameters. Each StageBlock first normalizes its input before multi-head attention, adds the attention output to the normalized query sequence, and then applies a second layer normalization followed by a residual feed-forward update. The hidden dimension is 768, and the feed-forward network uses dimensions

Table 5: Training and model configurations used in the experiments.

Setting	Configuration
Visual encoder	ViT-B/16
Text encoder	Frozen CLIP text encoder
Optimizer	AdamW
Distributed processes	2
Effective global batch size	96
Total training epochs	120
Initial learning rate	1×10^{-4}
Minimum learning rate	1×10^{-7}
Learning-rate schedule	Warm-up cosine decay
Warm-up epochs	10
Weight decay	5×10^{-5}
Data-loader workers	0
Repeat-factor sampling	Enabled
Focal loss	Enabled
Focal-loss α_{f}	0.25
Focal-loss γ_{f}	2.5
HOI matching cost	5
HOI query tokens N_q	25
Attention heads	6
Hidden dimension	768
Contrastive temperature τ	0.07
Alignment weight λ_{icot}	0.1
Fusion coefficient α	0.5

$768 \to 1536 \to 768$ with a GELU activation. Multi-head attention uses 6 heads and a dropout rate of 0.1.

The spatial StageBlock applies self-attention to F_{img}. The contact StageBlock uses the visual tokens as queries and the pooled spatial hidden state H_{spa} as a single conditional key–value token. The interaction StageBlock uses the visual tokens as queries and $Memory = [H_{\mathrm{spa}}; H_{\mathrm{con}}; G]$ as three conditional key–value tokens, where $G = \mathrm{GAP}(F_{\mathrm{img}})$ denotes the global visual representation.

The token-level output of each StageBlock is average-pooled into a 768-dimensional stage hidden state and projected into a normalized 512-dimensional feature for progressive semantic alignment. Attention-derived spatial and contact importance weights modulate the visual tokens between successive reasoning stages. Finally, the three pooled hidden states are concatenated and projected into 25 image-conditioned HOI queries, which are added to the learnable HOI query embeddings before the HOI decoder.

B Textual Knowledge Construction

ProgAlign uses two complementary forms of textual knowledge. Category-level descriptions provide the textual features used for final interaction classification, whereas stage-specific descriptions provide spatial, contact, and interaction priors for the auxiliary training objective.

B.1 Language Model and Offline Generation

All descriptions are generated offline with the OpenAI gpt-oss-20b open-weight reasoning model (OpenAI, 2025), served through Ollama Cloud under the runtime tag `gpt-oss:20b-cloud`. One English response is retained for each prompt, and the decoding parameters follow the platform defaults. The language model is not invoked during detector training or inference.

Table 6: Language-model serving and generation configuration.

Setting	Configuration
Model developer	OpenAI
Model name	`gpt-oss-20b`
Serving platform	Ollama Cloud
Runtime tag	`gpt-oss:20b-cloud`
Outputs per prompt	1
Output language	English
Decoding parameters	Platform defaults

Accepted responses are generated once before training and reused across all model variants and experimental runs.

B.2 HOI Category Grouping

Category grouping is performed independently for HICO-DET and SWIG-HOI. Each interaction label is converted into a canonical verb–object phrase by replacing underscores with spaces while preserving its semantics.

The normalized category names are encoded by the same frozen CLIP text encoder used in the final text branch. The resulting features are L_2-normalized and grouped by K-means (MacQueen, 1967). Categories within the same group typically share related actions, objects, or interaction meanings, providing comparative context for fine-grained description generation. The grouping procedure uses only category names and does not access images, bounding boxes, or instance-level annotations from unseen categories. Group assignments are computed once and fixed for all experiments.

B.3 Category-Level Description Prompt

For each target category, the model receives the target name and the remaining category names in the same group. These related categories provide comparative context for identifying observable cues that distinguish visually similar interactions.

System instruction.

```
You are an expert in visual Human-Object interaction recognition.
Given a target HOI category and several semantically related
categories, describe the observable visual evidence of the target
interaction and emphasize the cues that distinguish it from the
related interactions.  Do not include information that cannot
reasonably be inferred from an image.
```

User prompt.

```
Target HOI category:  {TARGET CATEGORY}
Related HOI categories in the same group:  {RELATED CATEGORY LIST}
Generate one concise fine-grained visual description of the target
interaction.  The description must characterize the human pose and
action execution, relative Human-Object configuration, relevant
physical contact, and object state; emphasize differences from
related categories; contain only observable visual information;
and use one English sentence.
Output format:  {TARGET CATEGORY}:  {FINE-GRAINED DESCRIPTION}
```

The target category is excluded from the related-category list, and one accepted category-level description is retained for each HOI category.

B.4 Stage-Specific Semantic Prompts

Three independent prompts construct semantic priors corresponding to the spatial, contact, and interaction stages of I-CoT.

Common system instruction.

```
You are an expert in visual Human-Object interaction analysis.
Given an HOI category, provide one concise description focusing
only on the requested reasoning aspect.  Use observable and
category-relevant visual evidence without adding unverifiable
scene details.
```

Spatial-stage prompt.

```
HOI category:  {HOI CATEGORY}
Describe the relative spatial configuration between the human and
the object in this interaction.  Focus on their relative position,
orientation, distance, and overall Human-Object layout.  Use one
concise English sentence.
```

Contact-stage prompt.

```
HOI category:  {HOI CATEGORY}
Describe the physical contact between the human and the object in
this interaction.  Focus on the body parts involved, the contacted
object regions, and the type of physical support or manipulation.
Use one concise English sentence.
```

Interaction-stage prompt.

```
HOI category:  {HOI CATEGORY}
Describe the interaction semantics expressed by this category.
Focus on the action goal, the role of the object, and the
observable execution of the interaction.  Use one concise English
sentence.
```

Each prompt is queried independently so that the resulting text corresponds to one designated reasoning aspect.

B.5 Output Processing and Offline Caching

Leading and trailing spaces, redundant quotation marks, and unnecessary line breaks are removed from the generated responses. A response is regenerated when it is empty, violates the requested format, merely repeats the category name, or does not describe the requested semantic aspect. Accepted responses are not manually rewritten.

Category-level descriptions are encoded by the frozen CLIP text encoder to obtain the textual features used for final interaction classification. Spatial, contact, and interaction descriptions are encoded independently to obtain the three stage-specific semantic priors. Each HOI category is therefore associated with one category-level description and three stage-specific descriptions. All descriptions and encoded features are cached before detector training, so gpt-oss-20b adds no training-time calls or inference cost.

B.6 Representative Generated Descriptions. For "ride bicycle," the related categories include "sit on bicycle," "repair bicycle," and "stand beside bicycle." The generated category-level description is: *The person is seated on the saddle, grips the handlebars, places the feet on the pedals, maintains balance, and propels the bicycle forward rather than merely holding or standing beside it.*

The corresponding spatial description is: *The person is positioned above and astride the bicycle, with the body aligned along the bicycle frame.* The contact description is: *The hands contact the handlebars, the feet contact the pedals, and the body is supported by the saddle.* The interaction description is: *The person operates the bicycle by balancing, steering, and pedaling it to move forward.*

C Exact Parameter Sensitivity Values

Tables 7 and 8 report the exact ten-point sensitivity results for the alignment weight λ_{icot} and fusion coefficient α on HICO-DET and SWIG-HOI. Only the evaluated hyperparameter is varied in each sweep, and the best result within each sweep is shown in bold.

Table 7: Exact parameter-sensitivity results on HICO-DET (mAP %). The best result within each hyperparameter sweep is shown in bold.

Setting	Value	Unseen	Seen	Full
λ_{icot}	0	23.31	28.54	27.32
λ_{icot}	0.025	23.39	28.57	27.42
λ_{icot}	0.05	23.47	28.59	27.51
λ_{icot}	0.075	23.54	28.65	27.58
λ_{icot}	0.10	23.58	**28.71**	**27.63**
λ_{icot}	0.125	23.56	28.69	27.60
λ_{icot}	0.15	23.52	28.66	27.55
λ_{icot}	0.20	**23.64**	28.63	27.49
λ_{icot}	0.25	23.49	28.56	27.39
λ_{icot}	0.30	23.36	28.48	27.28
α	0.1	23.22	28.53	27.11
α	0.2	23.20	28.51	27.20
α	0.3	23.17	28.49	27.28
α	0.4	23.43	28.62	27.49
α	0.5	**23.58**	28.71	**27.63**
α	0.6	23.49	28.36	27.51
α	0.7	23.41	**28.75**	27.46
α	0.8	23.37	28.69	27.40
α	0.9	23.32	28.64	27.35
α	1.0	23.26	28.57	27.24

Table 8: Exact parameter-sensitivity results on SWIG-HOI (mAP %). The best result within each hyperparameter sweep is shown in bold.

Setting	Value	Non-rare	Rare	Unseen	Full
λ_{icot}	0	23.54	17.31	12.21	17.62
λ_{icot}	0.025	23.68	17.49	12.30	17.76
λ_{icot}	0.05	23.81	17.66	12.39	17.89
λ_{icot}	0.075	23.93	17.80	12.50	18.02
λ_{icot}	0.10	**24.02**	**17.91**	12.58	**18.11**
λ_{icot}	0.125	23.99	17.86	12.56	18.09
λ_{icot}	0.15	23.94	17.82	12.51	18.02
λ_{icot}	0.20	23.88	17.73	**12.63**	17.98
λ_{icot}	0.25	23.75	17.61	12.47	17.86
λ_{icot}	0.30	23.61	17.48	12.34	17.74
α	0.1	23.68	17.54	12.33	17.78
α	0.2	23.62	17.50	12.29	17.73
α	0.3	23.57	17.46	12.24	17.69
α	0.4	23.86	17.69	12.45	17.92
α	0.5	**24.02**	**17.91**	12.58	**18.11**
α	0.6	23.91	17.72	12.49	17.97
α	0.7	23.96	17.83	**12.61**	18.03
α	0.8	23.84	17.70	12.56	17.91
α	0.9	23.71	17.58	12.36	17.81
α	1.0	23.59	17.43	12.27	17.66

HICO-DET. For λ_{icot}, the best Full and Seen mAP values are obtained at $\lambda_{\text{icot}} = 0.10$, while the highest Unseen mAP is obtained at $\lambda_{\text{icot}} = 0.20$. Considering the overall performance across the

three evaluation splits, we adopt $\lambda_{\text{icot}} = 0.10$ as the default semantic alignment weight. Within the α sweep, $\alpha = 0.5$ achieves the best Unseen and Full mAP, whereas $\alpha = 0.7$ produces a slightly higher Seen mAP. We therefore select $\alpha = 0.5$, which provides the strongest overall performance on HICO-DET.

SWIG-HOI. For λ_{icot}, the setting $\lambda_{\text{icot}} = 0.10$ achieves the best Non-rare, Rare, and Full mAP, while $\lambda_{\text{icot}} = 0.20$ produces the highest Unseen mAP. The selected value $\lambda_{\text{icot}} = 0.10$ therefore provides the best overall balance across the four evaluation splits. Within the α sweep, $\alpha = 0.5$ obtains the best Non-rare, Rare, and Full mAP, whereas $\alpha = 0.7$ achieves a slightly higher Unseen mAP. Based on the overall results, we use $\alpha = 0.5$ in all main experiments.