TOWARDS OPEN-WORLD HUMAN-OBJECT INTERACTION REASONING WITH MULTIMODAL LARGE LANGUAGE MODEL

Eastman Z Y, WU[1,2,3] *Yali Li*[1,2,3], *Shengjin Wang*[1,2,3†]

[1]Department of Electronic Engineering, Tsinghua University
[2]Beijing National Research Center for Information Science and Technology (BNRist), China
[3]National Engineering Research Center of Dangerous Articles and Explosives Detection Technologies, Beijing; 100084, China

ABSTRACT

Human-Object Interaction (HOI) detection extends conventional object detection by reasoning about higher-level semantic relationships between humans and objects. Most existing HOI detectors are built on DETR-style architectures and rely on external knowledge from large language models (LLMs) or vision-language models (VLMs). However, they still face three major challenges: (1) insufficient semantic understanding, which limits fine-grained interaction recognition; (2) restricted open-world knowledge due to reliance on predefined concepts; and (3) weak interpretability. To address these issues, we propose **HOI-MLLM**, a framework for HOI detection that leverages the reasoning capability of multimodal large language models (MLLMs). We construct balanced supervised fine-tuning (SFT) data with curated chain-of-thought (CoT) annotations to elicit intrinsic reasoning ability for HOI tasks. We further adopt a two-stage training strategy that combines SFT warm-up with GRPO-based post-training, guided by HOI-specific reward functions to enhance reasoning capability. Experiments on V-COCO and HICO-DET demonstrate that our HOI-MLLM achieves state-of-the-art performance. In addition, the generated reasoning chains improve interpretability, offering new insights into explainable HOI detection. Code will be available at HOI-MLLM.

***Index Terms*—** Human-object interaction, Multimodal large language model, Chain-of-thought, Open-world.

1. INTRODUCTION

Human-Object Interaction (HOI) detection aims to localize humans and objects and recognize their interactions. Beyond conventional object detection, HOI reasoning requires fine-grained, human-centric scene understanding that captures higher-level semantics. Such capability supports a wide range of applications, including collaborative robotics, intelligent surveillance, and activity recognition.

Recent advances in vision-language models (VLMs), such as CLIP [1], have driven progress in open-vocabulary HOI reasoning. Most existing HOI methods adopt VLM encoders within DETR[2]-style frameworks and augment them with external knowledge from large language models (LLMs). Despite these advances, as shown in Figure 1, current approaches remain limited: they rely on predefined concepts and categories, exhibit insufficient semantic reasoning, and provide weak interpretability, ultimately falling short in modeling fine-grained interactions. These limitations restrict their generalization to dynamic, open-world scenarios.

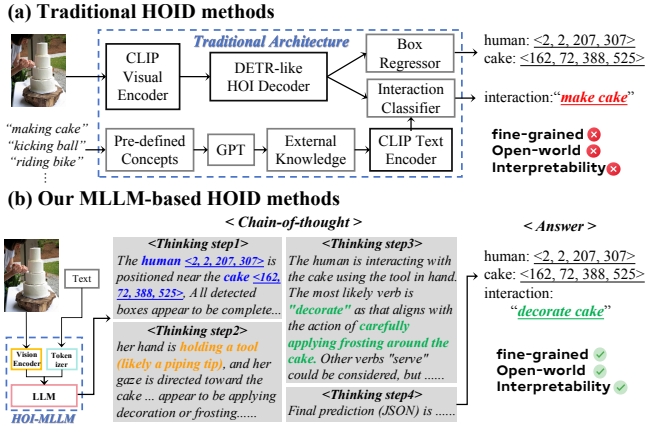

Fig. 1. **Motivation of our proposed HOI-MLLM**. (a) Traditional HOI detection methods are typically built on DETR-style architectures and rely on pre-defined concepts and categories. (b) Our method leverages chain-of-thought reasoning to enable fine-grained, interpretable HOI detection with open-world knowledge.

To address these issues, we propose **HOI-MLLM**, the first HOI detector built upon multimodal large language models (MLLMs) with chain-of-thought (CoT) reasoning. Our approach is built upon two key components. First, we construct balanced supervised data with carefully curated CoT annotations to elicit the intrinsic reasoning capability of MLLMs. Given the severe long-tail distribution of existing HOI benchmarks, we derive a smaller yet balanced subset from the original imbalanced data pool and transform each sample into a unified open-vocabulary HOI template that enables instance-level parsing. We further design a structured four-step CoT template, covering reflection on detection results, visual understanding, interaction reasoning, and final HOI output, which guides MLLMs to produce interpretable reasoning chains aligned with structured predictions. Second, we adopt a two-stage training strategy that combines supervised fine-tuning (SFT) with GRPO-based post-training. In this stage, we incorporate both image-level and instance-level HOI-specific reward functions to guide reinforcement learning and further strengthen the reasoning ability of the model. Through this design, HOI-MLLM is able to generate open-vocabulary predictions grounded in non-predefined, open-world knowledge.

Extensive experiments on two benchmarks demonstrate that HOI-MLLM achieves state-of-the-art performance. In addition,

This work is supported by the research fund under Grant No. 20242910035 from the Tsinghua University-Jiangsu CRRC Digital Technology Co.,Ltd. Joint Research Center for Data Driven Intelligence of Industry.

†corresponding author

the generated reasoning chains provide improved interpretability. Overall, HOI-MLLM establishes a strong baseline for open-world HOI reasoning and opens new opportunities for applying MLLMs to broader human-centric visual understanding. Our main contributions are summarized as follows:

- We present **HOI-MLLM**, the first HOI detection framework built upon MLLMs with structured CoT reasoning, establishing a new paradigm beyond conventional DETR-style architectures and accompanied by a unified evaluation protocol.
- We develop a two-stage training strategy that integrates SFT with GRPO-based post-training, where image-level and instance-level HOI-specific reward functions are introduced to further enhance ability through reinforcement learning.
- Extensive experiments on V-COCO and HICO-DET demonstrate that HOI-MLLM achieves state-of-the-art performance, while the generated reasoning chains improve interpretability and provide new insights into explainable HOI detection.

2. RELATED WORK

Traditional HOI detection methods. Since the introduction of DETR [2], query-based approaches have become the dominant paradigm in HOI detection. In general, existing methods can be categorized into one-stage [3, 4, 5, 6, 7, 8] and two-stage [9, 10, 11, 12, 13] frameworks. One-stage methods aim to jointly detect objects and infer interactions within a single model. For example, QPIC [14] extend the DETR [2] architecture by adding an interaction head to predict HOI relationships. In contrast, two-stage methods first rely on off-the-shelf object detectors to generate candidate human and object bounding boxes, and then perform interaction reasoning through pairing and classification. Many of these methods incorporate transformer-based architectures while augmenting them with additional cues such as human pose [15, 11, 13], linguistic cues [16, 17], or representations from VLMs [18, 19, 20, 21, 22].

Integrating MLLMs. With the rapid development of MLLMs, several recent works have explored their potential to enhance HOI detection. However, rather than directly constructing MLLM-based HOI detector, these methods typically make secondary use of the knowledge embedded in MLLMs. For instance, UniHOI [23] employs a retrieval-augmented generation (RAG) strategy to exploit the open-world knowledge in BLIP-2 [24], while ContextHOI [22] leverages GPT to generate textual features for pre-defined concepts to improve interaction classification.

3. METHOD

3.1. Data Preparation

Balanced Data Curation. Existing HOI datasets, such as HICO-DET, exhibit a severe long-tail distribution, where common classes contain more than 10k samples while many rare interactions have fewer than 10. For fine-tuning MLLMs, the quality and balance of training data are particularly critical. To address this issue, we adopt a greedy selection strategy that prioritizes samples involving rare interactions at the instance level. In this way, we obtain a more balanced HOI dataset for post training.

Structured Output Definition. Since MLLMs generate outputs in an autoregressive manner, the raw textual responses are inherently free-form and unstructured. To ensure reliable parsing, we design a structured output format that constrains generation into a consistent schema, which can then be systematically decoded into HOI predictions. The expected format is defined as:

```
HOI_result = {
  "human_boxes": [...], # n_gt × 4 List[array]
  "object_boxes": [...], # n_gt × 4 List[array]
  "object_classes": [...],# n_gt List[str]
  "interactions": [...], # n_gt List[str]
}
```

This format consists of four components: *human_boxes*, *object_boxes*, *object_classes*, and *interactions*, which together form n_{gt} triplets $\langle$human, interaction, object$\rangle$. For human-object pairs with multiple interactions, we duplicate the pair for each interaction rather.

3.2. Reasoning with Chain-of-Thought

Chain-of-Thought (CoT) prompting has been shown to significantly improve performance on complex reasoning tasks such as code generation and mathematical problem solving. HOI detection is also a composite task, which involves object localization, human-object pairing, and interaction reasoning. Inspired by this, we introduce CoT supervision to elicit the reasoning capability of MLLMs in HOI detection. Specifically, we construct a small set of CoT-augmented samples in the early stage of training. For each sample, we use strong MLLMs such as GPT-4o-mini to generate detailed step-by-step reasoning according to a carefully designed template.

Step1. Reflection on detection results: assess accuracy of human/object boxes; identify missing or redundant ones.

Step2. Detailed Visual Cues Mining: describe human actions, posture, and contextual cues.

Step3. Interaction reasoning: match humans to objects, infer the most likely action, and justify the choice.

Step4. Final HOI output: provide the HOI prediction strictly in JSON format without additional explanations.

By incorporating these CoT-augmented samples during supervised fine-tuning, we activate the intrinsic reasoning ability of MLLMs for HOI tasks. As a result, the model not only learns to produce structured outputs but also acquires intermediate reasoning patterns that improve interpretability and generalization.

3.3. Mixed Training Objective

As discussed earlier, HOI detectors are typically divided into one-stage and two-stage methods. We unify both paradigms in **HOI-MLLM** by formulating HOI as structured text generation from image–text inputs. Given an image $\mathcal{I}$, a plain prompt $\mathcal{P}$, an object-detection prompt $\mathcal{P}_{\text{ob}}(\mathcal{B})$, the model Φ_θ auto-regressively produces tokens parsed into HOI predictions. During training, we jointly optimize one-stage and two-stage objectives to improve robustness.

We further introduce a control variable $\tau \in \{0, 1\}$ to indicate whether CoT reasoning is required. In practice, this is implemented by whether the prompt contains a CoT prefix. Let T denote the structured HOI tokens and R the reasoning tokens. Our SFT objective combines three supervision streams:

$$\begin{aligned}\theta^* \;=\; \arg\min_{\theta}\; \mathbb{E}_{(\mathcal{I},\mathcal{P},T,R)\sim\mathcal{D}}\big[&\mathcal{L}_{\text{LM}}\big(\Phi_\theta(\mathcal{I},\mathcal{P},\tau{=}0),\,T\big)\\ &+\mathcal{L}_{\text{LM}}\big(\Phi_\theta(\mathcal{I},\mathcal{P}_{\text{ob}}(\mathcal{B}),\tau{=}0),\,T\big)\\ &+\mathcal{L}_{\text{LM}}\big(\Phi_\theta(\mathcal{I},\mathcal{P}_{\text{ob}}(\mathcal{B}),\tau{=}1),\,[R;T]\big)\big],\end{aligned} \tag{1}$$

where $\mathcal{L}_{\text{LM}}$ is token-level cross-entropy and $[R;T]$ concatenates reasoning and answer tokens for CoT samples.

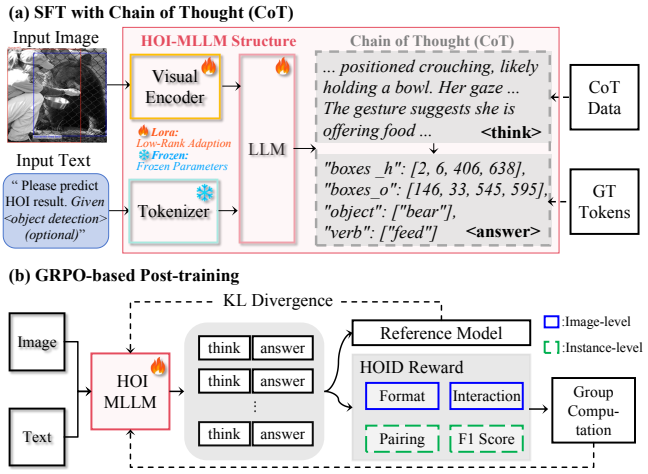

Fig. 2. Overview of the HOI-MLLM training pipeline. (a) **SFT with CoT:** the model generates step-by-step reasoning and structured predictions, supervised by curated CoT data and ground-truth tokens. (b) **GRPO Post-training:** the model is further optimized with GRPO, guided by HOI-specific rewards.

3.4. GRPO-based Post-training

Group Relative Policy Optimization (GRPO) [25] is a recent reinforcement learning algorithm that has shown effectiveness on complex reasoning tasks such as mathematics and code generation. After warming up HOI-MLLM with SFT and CoT, we further optimize it using GRPO. Both the policy model and the reference model are initialized from the warm-up checkpoint, as shown in Figure 2.

To align the autoregressive outputs with HOI-specific objectives, we design task-oriented rewards at two granularities: *image-level* and *instance-level*. Image-level rewards include: (1) **Format**, which enforces structural validity (e.g., presence of `<think>` and `<answer>` sections and consistent output length), and (2) **Interaction**, which evaluates whether predicted interactions match the ground truth at the image level, analogous to a captioning loss. Instance-level rewards include: (1) **Pairing**, which checks whether predicted humans and objects are correctly matched for each interaction (a match is correct if both bounding boxes achieve IoU > 0.5 with ground truth), and (2) **F1 score**, computed over predicted HOI triplets $\langle$human, interaction, object$\rangle$ within a batch.

By jointly optimizing image-level and instance-level objectives, GRPO enables HOI-MLLM to generate structurally valid outputs while achieving fine-grained reasoning and accurate HOI detection.

4. EXPERIMENTS

4.1. Experiment Settings

Datasets. We evaluate our method on two widely used HOI benchmarks: HICO-DET [9], V-COCO [26]. **HICO-DET** contains 47,100 images with 600 HOI categories, spanning 80 object classes and 117 verb classes. **V-COCO**, built upon MS-COCO [27], defines 24 interaction categories involving 80 objects and 24 interactions.

Implementation Details. We adopt Qwen-VL-2.5 3B as the backbone of our HOI-MLLM and apply LoRA-based fine-tuning to both the visual encoder and the large language model. We set the LoRA rank to 64, the scaling factor α to 128, and the dropout rate to 0.05. Training is performed with a global batch size of 64, an initial learning rate of 2×10^{-4}, weight decay of 0.1, and a warmup ratio of 0.03. We employ a cosine learning rate scheduler. All experiments are conducted using 8 NVIDIA RTX 4090 GPUs.

V-COCO	mF1	mPrec	mRec
End-to-End OVD & HOID			
HOI-MLLM (ours)	51.69	61.57	45.81
R50-DETR for Object Detection			
PViC [28]	67.12	68.13	67.32
CMMP [19]	66.55	60.73	74.76
CMMP† [19]	67.65	60.98	77.47
ADA-CM [20]	64.67	58.67	73.16
ADA-CM† [20]	67.80	64.21	73.17
EZ-HOI [29]	68.01	61.59	77.04
EZ-HOI† [29]	68.42	62.56	**77.68**
HOI-MLLM (ours)	**69.28**	**74.44**	65.56
Oracle Setting			
PViC† [28]	79.57	83.74	84.03
CMMP [19]	79.41	73.07	87.93
CMMP† [19]	81.06	75.19	**89.22**
ADA-CM [20]	78.18	73.84	83.95
ADA-CM† [20]	79.02	74.64	85.22
EZ-HOI [29]	80.60	77.86	84.41
EZ-HOI† [29]	81.02	78.91	84.84
HOI-MLLM (ours)	**81.94**	**85.32**	79.35

Table 1. Performance comparison of our HOI-MLLM with state-of-the-arts HOID methods on V-COCO dataset.

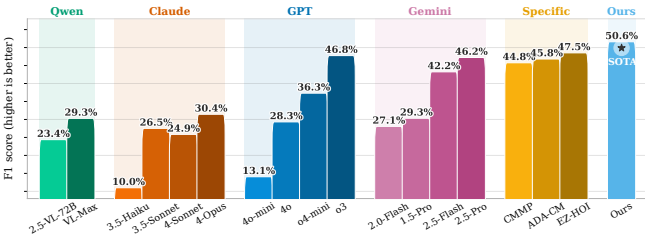

Fig. 3. Comparison with Cutting-Edge General-Purpose MLLMs.

Evaluation Metrics. Unlike conventional HOI detectors that produce logits over predefined categories, MLLMs generate free-form natural language responses, which makes the direct application of the mAP metric non-trivial. Accordingly, we evaluate model performance using mean F1-score, mean Precision, and mean Recall.

4.2. Comparison results

Settings. In our comparative experiments, we evaluate under two settings. In the `R50-DETR` setting, object detections are generated by an off-the-shelf DETR and provided as textual input, ensuring both HOI-MLLM and baselines reason over the same detections. In the `Oracle` setting, ground-truth object boxes are given to all methods, enabling a fair comparison of interaction reasoning capability.

Benchmark Results. On the V-COCO dataset, our HOI-MLLM achieves the best overall performance under both the R50-DETR and Oracle settings. In particular, HOI-MLLM attains an mF1 of 69.28 and mPrec of 74.44 in the R50-DETR setting, surpassing all existing baselines. Compared with scale-up variants (marked with †) that integrating ViT-L/14@336px CLIP, our method still outperforms CMMP† and EZ-HOI† by +1.63% and +0.86% mF1, respectively. More importantly, HOI-MLLM exhibits significantly higher precision, highlighting its stronger capability for fine-grained scene understanding and reducing spurious predictions. In the Oracle setting, HOI-MLLM achieve 81.94 mF1 and 85.32 mean precision.

On the HICO-DET dataset, our method shows competitive but slightly suboptimal results in the R50-DETR setting. Nevertheless, under the Oracle setting, HOI-MLLM achieves SOTA performance

HICO-DET	mF1			mPre	mRec
	Full	Rare	N.Rarwane	Full	Full
End-to-End OVD & HOID					
HOI-MLLM	24.28	25.75	23.27	26.72	26.26
R50-DETR for Object Detection					
PViC [28]	28.68	23.61	30.20	25.29	44.10
CMMP [19]	26.08	23.16	26.95	24.40	36.64
CMMP† [19]	33.24	30.24	34.14	33.24	43.52
ADA-CM [20]	28.20	24.64	29.26	27.37	35.62
ADA-CM† [20]	**34.14**	**31.33**	**34.98**	32.25	**44.69**
EZ-HOI [29]	27.82	23.25	29.18	26.00	36.58
EZ-HOI† [29]	31.44	29.43	32.04	30.57	40.20
HOI-MLLM (ours)	30.32	27.12	31.68	**35.71**	36.21
Oracle Setting					
PViC† [28]	44.41	46.49	43.79	39.73	61.63
CMMP [19]	36.55	37.27	36.34	30.74	59.22
CMMP† [19]	44.76	46.48	44.25	40.77	60.32
ADA-CM [20]	39.78	36.24	40.84	36.68	53.68
ADA-CM† [20]	45.81	42.57	46.78	41.97	61.39
EZ-HOI [29]	42.72	33.42	45.50	37.82	59.48
EZ-HOI† [29]	47.52	43.81	48.62	43.32	**63.24**
HOI-MLLM (ours)	**50.61**	**51.16**	**49.90**	**51.37**	56.47

Table 2. Performance comparison of our HOI-MLLM with mainstream MLLMs and traditional HOI detection methods, under a unified evaluation protocol. HICO-DET datasets.

#	SFT	Mixed	CoT	GRPO	mF1	mPre	mRec
A1	×	×	×	×	31.31	41.92	27.46
A2	✓	×	×	×	64.07	71.00	59.46
A3	✓	✓	×	×	66.97	68.66	66.07
A4	✓	✓	✓	×	68.10	73.85	63.71
A5	✓	✓	✓	✓	69.28	74.44	65.56

Table 3. Ablation studies on the V-COCO, R50-DETR setting.

with 50.61 mF1, outperforming all prior methods, such as EZ-HOI† (+3.59% mF1). Furthermore, HOI-MLLM is able to operate in a fully end-to-end manner, jointly performing open-vocabulary object detection and interaction reasoning without reliance on a separate object detector. This end-to-end capability highlights the potential of MLLMs as unified solutions for open-world HOI reasoning.

Compared with Close-source MLLMs. We further compare HOI-MLLM with several state-of-the-art MLLMs, including the Qwen, Claude, GPT, and Gemini families, as shown in Figure 3. Although these general-purpose MLLMs contain significantly larger parameter scales, they exhibit limited performance on HOI detection (typically below 40% F1). In contrast, our method achieves 50.6% F1, establishing a new state-of-the-art. Notably, HOI-MLLM surpasses the strongest baselines such as GPT-o3 (46.8%) and Gemini-2.5-Pro (46.2%) by more than +4% absolute F1. This clear margin underscores the importance of domain-specific data construction, CoT supervision, and HOI-oriented reinforcement learning post-training.

Ablation Study on the V-COCO dataset, as shown in Table 3. Without fine-tuning (A1), the general-purpose MLLM performs poorly on HOI detection, yielding only 31.31 mF1. Introducing SFT alone (A2) brings a substantial improvement (64.07 mF1), indicating its necessity for basic task alignment. Incorporating mixed training objectives (A3) and CoT supervision (A4) further enhances both precision and recall, demonstrating the benefits explicit reasoning guidance. Finally, GRPO post-training (A5) achieves the best result of 69.28 mF1, confirming that HOI-specific reinforcement learning provides additional gains and strengthens reasoning robustness.

Qualitative Results. In Figure 4, our HOI-MLLM demonstrates clear advantages over traditional HOI detectors. 1) Open-vocabulary Detection. It can recognize instances beyond predefined categories, such as identifying "spoon with visible chocolate residue" instead of

Fig. 4. Our HOI-MLLM showcases advantages beyond traditional methods: blue indicates open-vocabulary detections, purple reflects open-world knowledge, red demonstrates reflective reasoning, and green emphasizes fine-grained visual cues such as posture, gesture.

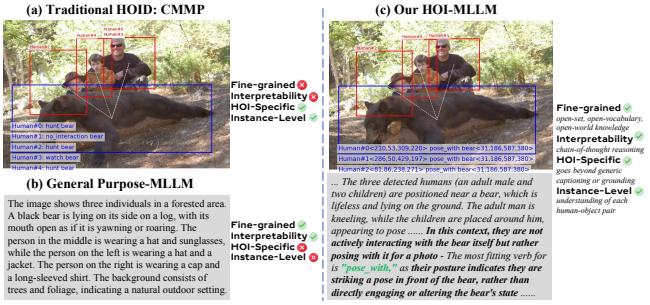

Fig. 5. Our HOI-MLLM *v.s* Traditional HOID & General MLLM.

merely "spoon." 2) Open-world Knowledge. It identifies "tagging" in a baseball scene, whereas conventional methods are limited to "holding" due to their lack of understanding of rules. 3) Reflective Reasoning. It corrects errors through self-reflection, such as revising a mistaken classification of a dog as a human. 4) Fine-grained perception. It actively captures subtle visual cues, including posture, gestures and object details, to refine interaction reasoning. Together, these capabilities highlight the robustness and interpretability in open-world settings. In Figure 5, traditional HOI detectors are constrained by closed-set predictions, often producing inaccurate results such as "*human hunt bear.*" General-purpose MLLMs lack a dedicated focus on interactions, generating scene descriptions that are unstructured and overly generic. Our method bridges this gap by producing HOI-specific reasoning chains and structured outputs at the human-object instance level, enabling fine-grained and accurate understanding, for example "*human pose_with bear.*"

5. CONCLUSION

We proposed **HOI-MLLM**, the first HOI detector built upon MLLMs with chain-of-thought reasoning. By constructing balanced HOI data with explicit chain-of-thought supervision and adopting a two-stage training strategy that combines SFT with GRPO-based post-training, our approach effectively elicits the intrinsic reasoning capability of MLLMs for HOI tasks. Experiments on HICO-DET and V-COCO demonstrate that HOI-MLLM achieves state-of-the-art performance while producing interpretable and open-world predictions. Overall, this work shifts HOI research from pattern recognition toward reasoning-based understanding, establishing a solid baseline for future advances in human-centric visual understanding.

6. REFERENCES

[1] 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*. PMLR, 2021, pp. 8748–8763.

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

[3] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng, "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*, 2020, pp. 482–490.

[4] Junwen Chen and Keiji Yanai, "Qahoi: query-based anchors for human-object interaction detection," *arXiv preprint arXiv:2112.08647*, 2021.

[5] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim, "Hotr: End-to-end human-object interaction detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 74–83.

[6] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim, "Uniondet: Union-level detector towards real-time human-object interaction detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 498–514.

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

[8] 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*, 2021, pp. 13234–13243.

[9] 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)*. IEEE, 2018, pp. 381–389.

[10] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He, "Detecting and recognizing human-object interactions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8359–8367.

[11] Tanmay Gupta, Alexander Schwing, and Derek Hoiem, "No-frills human-object interaction detection: Factorization, layout encodings, and training techniques," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9677–9685.

[12] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon, "Detecting human-object interactions with action co-occurrence priors," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 2020, pp. 718–736.

[13] Eastman ZY Wu, Yali Li, Yuan Wang, and Shengjin Wang, "Exploring pose-aware human-object interaction via hybrid learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17815–17825.

[14] 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*, 2021, pp. 10410–10419.

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

[16] Ye Liu, Junsong Yuan, and Chang Wen Chen, "Consnet: Learning consistency graph for zero-shot human-object interaction detection," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4235–4243.

[17] 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*, vol. 35, pp. 37416–37431, 2022.

[18] 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*, 2023, pp. 23507–23517.

[19] Ting Lei, Shaofeng Yin, Yuxin Peng, and Yang Liu, "Exploring conditional multi-modal prompts for zero-shot hoi detection," in *European Conference on Computer Vision*. Springer, 2025, pp. 1–19.

[20] 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*, 2023, pp. 6480–6490.

[21] Sandipan Sarma, Pradnesh Kalkar, and Arijit Sur, "Boosting zero-shot human-object interaction detection with vision-language transfer," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 6355–6359.

[22] Jianjun Gao, Kim-Hui Yap, Kejun Wu, Duc Tri Phan, Kratika Garg, and Boon Siew Han, "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)*. IEEE, 2024, pp. 13436–13440.

[23] Yichao Cao, Qingfei Tang, Xiu Su, Song Chen, Shan You, Xiaobo Lu, and Chang Xu, "Detecting any human-object interaction relationship: Universal hoi detector with spatial prompt learning on foundation models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 739–751, 2023.

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19730–19742.

[25] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al., "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

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

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[28] Frederic Z Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould, "Exploring predicate visual context in detecting of human-object interactions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10411–10421.

[29] Qinqian Lei, Bo Wang, and Tan Robby T., "Ez-hoi: Vlm adaptation via guided prompt learning for zero-shot hoi detection," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.