The 29th Meeting on Image Recognition and Understanding

HOIBlender: Efficient Multi-Prior Fusion with HOI Decoders for Query-Free Human-Object Interaction Detection

Junwen CHEN[1,a)] Keiji YANAI[1,b)]

Abstract

Recent HOI detection research has progressively adopted state-of-the-art object detectors as stronger foundations. In particular, Hybrid-SOV, built on RT-DETR, has demonstrated that high HOI accuracy can be achieved together with faster inference. However, its complex denoising query design and high dependence on an efficient encoder still leave room for improvement in simplicity, scalability, and deployment efficiency. Motivated by the rapid evolution of modern detectors, we further introduce the latest RF-DETR-based detection architecture into HOI detection and establish a new generation of strong baselines. Experiments on HICO-DET show that our method consistently improves over previous approaches while maintaining competitive efficiency.

1. Introduction

Human-object interaction (HOI) detection aims to localize a human, localize an interacted object, and classify their interaction in a unified triplet representation. Since the release of HICO-DET [3], HOI detection has become a standard benchmark for scene understanding beyond object recognition. However, practical HOI detection remains challenging because interaction labels are long-tailed, rare HOI categories have weak supervision, and visual evidence is often ambiguous for fine-grained predicates.

Early HOI methods mainly relied on two-stage pipelines and manually designed pair modeling [10], [11], [29]. With the success of transformers [2], [34], DETR-style end-to-end frameworks such as HOTR [14], QPIC [33], and subsequent variants [21], [37], [40] significantly simplified association learning. More recently, vision-language priors [19], [30] have shown clear potential for semantic generalization in HOI [27], [35]. Nevertheless, two gaps remain: *(i)* efficiency-oriented HOI modeling is still underexplored compared with modern real-time detectors, and *(ii)* external semantic priors are often introduced as auxiliary heads rather than deeply integrated into interaction decoding.

In this work, we present **HOIBlender**, an efficient HOI framework built upon RF-DETR-style detection design [9], [32]. Starting from a strong and efficient detector backbone, HOIBlender introduces a dual-stage HOI decoding strategy: a spatial interaction stage first builds object/subject-aware query states, and a semantic interaction stage then performs unified verb/HOI decoding with cross-modal prior fusion. To improve long-tail robustness, we inject external priors from BLIP-2 [19] into decoding and classifier initialization, so that interaction logits are aligned with rich vision-language semantics. In addition, we transfer Group-DETR-style grouped query training [8] into HOI learning, which improves optimization efficiency without introducing complicated denoising-query design.

Our contributions are summarized as follows.

- We propose HOIBlender, an efficient HOI detector that unifies RF-DETR-style detection efficiency and interaction-aware decoding in a single end-to-end framework.
- We introduce multi-prior semantic fusion decoding for both HOI/verb prediction, which improves visual-semantic alignment.
- We show consistent gains on HICO-DET over strong baselines, validating the effectiveness of combining efficient detection design and grouped-query training.

2. Related Work

HOI detection. Traditional HOI detectors usually follow a detect-then-classify paradigm with handcrafted pair representations [10], [11], [12], [29]. Transformer-based methods reformulated HOI as set prediction and explicit human-object association, substantially improving end-to-end learning [7], [14], [33], [42]. Later approaches further enhanced context modeling, query mining, and compositional generalization [13], [21], [37], [39]. Our work is aligned with this line but focuses on a complementary direction: improving efficiency and semantic robustness simultaneously.

Efficient DETR-style detectors for HOI. DETR [2] and Deformable DETR [41] established end-to-end detection as a scalable paradigm, while newer variants targeted training and inference efficiency [8], [18], [22], [38]. RF-DETR/LW-DETR-style designs [9], [32] provide a strong foundation with efficient feature extraction structures

[1] The University of Electro-Communications
[a)] chen-j@mm.inf.uec.ac.jp
[b)] yanai@mm.inf.uec.ac.jp

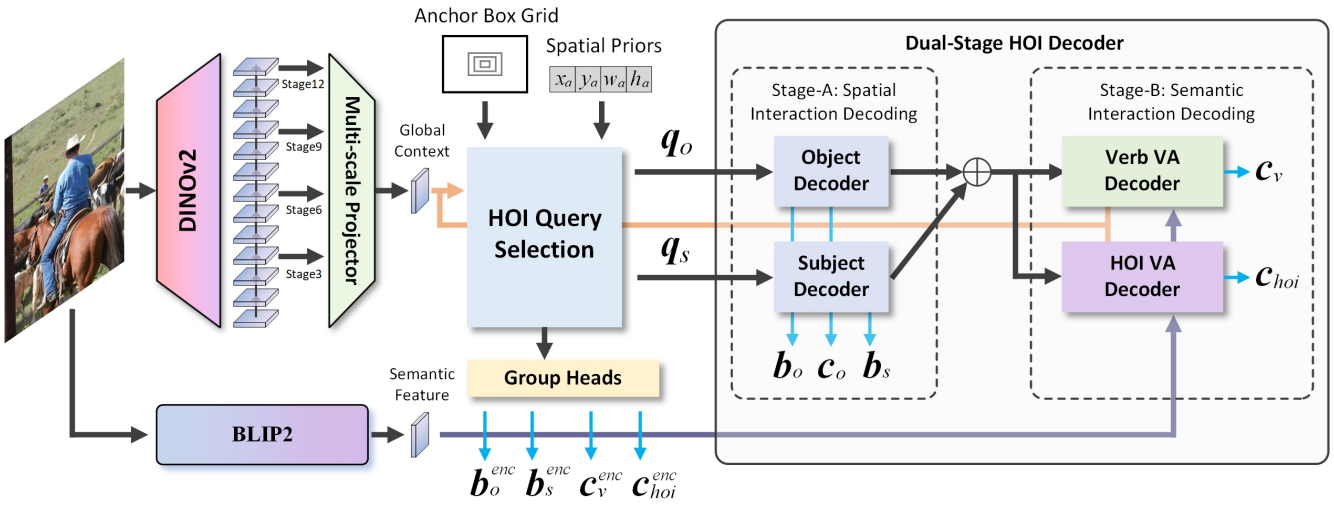


Fig. 1 Overall architecture of HOIBlender.

and practical training recipes. HOIBlender inherits this efficiency-oriented backbone without an additional transformer encoder and adapts grouped-query training to HOI-specific matching and losses.

Vision-language priors for HOI understanding. Vision-language pre-training has become an effective solution for long-tail and open-vocabulary semantics [19], [20], [30]. In HOI, CLIP-based transfer and relational language-image pre-training have demonstrated clear benefits [27], [35], [36]. Different from methods that only use external priors as post-hoc calibration, follow recent approaches [1], [4], [5], HOIBlender integrates BLIP-2 [19] priors directly into interaction decoding and classifier initialization, leading to stronger category-aware HOI prediction under an efficient end-to-end architecture.

3. Method

3.1 Problem Setup

Given an image I, HOI detection predicts a set of interaction instances (b_h, b_o, c_o, c_{hoi}), where b_h and b_o denote human and object boxes, c_o denotes object classes, and c_{hoi} denotes an HOI interaction class (or equivalently a verb–object composition). Our HOIBlender jointly optimizes localization and interaction classification in one end-to-end framework. HOIBlender outputs object logits, object boxes, subject (human) boxes, verb logits, and HOI logits, so that geometric grounding and semantic interaction reasoning are coupled during training.

Compared with prior query-based HOI detectors that rely on heavy pairwise construction, our formulation keeps a detector-first structure and upgrades only interaction-specific decoding and supervision. This makes the method easy to scale across model sizes (Nano/Small/2XLarge) while preserving a consistent training objective.

3.2 Framework of HOIBlender

As shown in Fig. 1, HOIBlender is built on an efficient RF-DETR/LW-DETR-style detector [9], [32]. The DINOv2 [28] visual encoder extracts multi-scale features, and a multi-scale projector aggregates them into the global context features. Different from previous state-of-the-art HOI detectors that employ a transformer encoder to refine the global context features, HOIBlender directly uses the features from the multi-scale projector for the subsequent decoding stages. From our experiment results, this design simplifies the architecture and reduces computational overhead, while still providing strong feature representations.

Similar to Hybrid-SOV [5], we adopt the HOI Query Selection Module to select top-K global feature vectors as subject and object queries. Different from Hybrid-SOV, during training, we adopt grouped prediction heads (Group-DETR style [8]) to improve optimization efficiency; during inference, a single group of heads is used.

The selected queries are fed into a dual-stage HOI decoder that first refines spatial interaction features and then performs semantic interaction decoding with BLIP-2 [19] prior fusion for better long-tail generalization. HOIBlender contains three functional blocks: (1) efficient query proposal generation from backbone and multi-scale projector, (2) dual-stage spatial interaction decoding, and (3) semantic interaction decoding with external prior fusion. The first block provides strong and stable initialization for interaction reasoning; the second block decouples geometry and semantics; and the third block injects category-level prior knowledge for tail interactions. This follows a practical principle: *train with richer query diversity, infer with a compact query budget.*

3.3 Dual-stage HOI Decoding

Stage-A: spatial interaction decoding. Starting from top-K encoder proposals, HOIBlender uses two coupled streams to refine object-centric and subject-centric geometry. One stream predicts object-aware states and boxes, and the other predicts subject-aware states and boxes. We

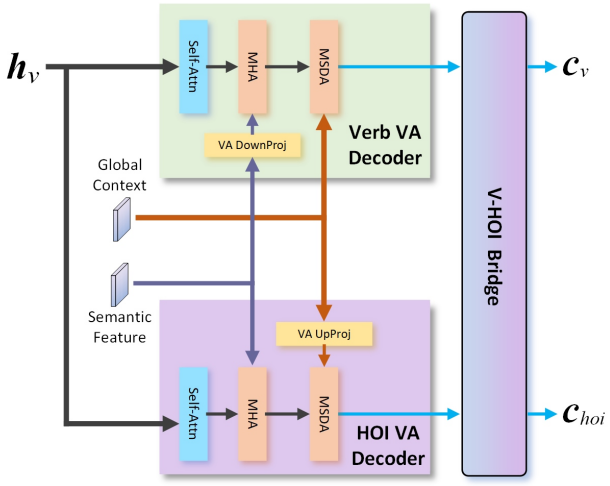


Fig. 2 Stage-B (verb/HOI decoding with fused priors).

keep these streams synchronized by sharing proposal origins and decoder depth, which stabilizes human–object alignment without explicit pair enumeration.

Let q_o^l and q_s^l denote object and subject query states at decoder layer l. Stage-A produces final spatial states q_o^L, q_s^L and box predictions $\hat{b}_o, \hat{b}_s$, which are then consumed by semantic decoding.

Stage-B: semantic interaction decoding. Based on Stage-A outputs, we build interaction states for two semantic branches: a verb branch and an HOI branch. The verb branch emphasizes action discrimination, while the HOI branch performs category-unified interaction classification with language-aligned features. To facilitate the verb prediction with vision-language priors, we inject BLIP-2 features into the verb branch via the vision advisor decoder design with a dimension reduction Linear layer. Then, the V-HOI Bridge module [4] is applied to fuse verb-aware and HOI-aware embeddings, which further improves the visual-semantic alignment for interaction classification. We project and normalize the output verb and hoi embeddings before classification and apply a learnable logit scale to stabilize similarity-based prediction.

In simplified form,

$$h_v = (q_o^L + q_s^L)/2, \quad z_{hoi} = s \cdot W_{hoi}\left(\frac{h_{hoi}}{\|h_{hoi}\|_2}\right), \quad (1)$$

where s is a learnable temperature. This separation of spatial and semantic roles improves optimization compared with single-stage interaction heads.

3.4 Multi-prior Semantic Fusion

To improve long-tail HOI generalization, HOIBlender injects external priors at both image and label levels. As shown in Fig. 2, we adopt the vision advisor decoder from SOV-STG-VLA [4] and Hybrid-SOV [5] as our verb and HOI decoders. Different from SOV-STG-VLA and Hybrid-SOV, which only use the vision advisor for HOI classification, we further introduce the vision advisor for verb classification, which provides more fine-grained semantic guidance for verb prediction.

Image-level prior. BLIP-2 features are fed as additional memory in VA decoders. This provides global semantic context complementary to local detector features.

Label-level prior. According to SOV-STG-VLA, we adopt V-HOI Bridge and precompute HOI text embeddings to initialize HOI classifier weights, aligning interaction logits with language semantics from the beginning of training.

3.5 Matching and Training Objectives

We adopt Hungarian matching [16] to assign predictions to ground-truth interaction pairs. The matching cost combines object classification, box regression, generalized IoU [31], and interaction-related classification terms.

The overall objective is

$$\begin{aligned}\mathcal{L} = \lambda_{vfl}\mathcal{L}_{vfl} + \lambda_{box}(\mathcal{L}_{sub}^{box} + \mathcal{L}_{obj}^{box}) \\ +\lambda_{giou}(\mathcal{L}_{sub}^{giou} + \mathcal{L}_{obj}^{giou}) + \lambda_{verb}\mathcal{L}_{verb} + \lambda_{hoi}\mathcal{L}_{hoi},\end{aligned} \quad (2)$$

where $\mathcal{L}_{vfl}$ is object-focused varifocal-style classification [38], $\mathcal{L}_{verb}$ and $\mathcal{L}_{hoi}$ are multi-label interaction losses, and box/GIoU terms supervise human-object localization.

We further apply auxiliary supervision on intermediate decoder layers and encoder outputs, which improves convergence stability. In practice, AdamW [23], mixed precision, gradient accumulation, and grouped-query training [8] are jointly used for efficient optimization.

4. Experiments

4.1 Dataset and Evaluation Protocol

We evaluate HOIBlender on HICO-DET [3], a widely used benchmark for HOI detection with long-tailed interaction labels. Following standard protocols, we report mean Average Precision (mAP) on both *Default* and *Known Object* settings, each with *Full*, *Rare*, and *Non-Rare* splits. This protocol allows us to evaluate both overall interaction quality and tail-category robustness.

4.2 Implementation Details

Following RF-DETR, we build HOIBlender variants (Nano/Small/2XL) by scaling the backbone and decoders. The numbers of decoder layers are 2, 3, and 5 for HOIBlender-N, -S, and -2XL, respectively, fewer than the 6 layers in Hybrid-SOV. All variants use 13 query groups during training. We train 15 epochs for Nano and Small, and 9 for 2XL; training from stronger pretrained detectors also converges faster.

4.3 Comparison with State of the Art

Table 1 summarizes the comparison with representative methods. Compared with prior transformer-based and vision-language-enhanced HOI approaches, HOIBlender achieves competitive or superior mAP with fewer training sessions.

4.4 Ablation Study

Grouped queries. We vary the number of query groups during training on HOIBlender-N. As shown in Table 3, 13

Method	Epoch	Backbone	Default			Known Object		
			Full	Rare	Non-Rare	Full	Rare	Non-Rare
QPIC [33]	150	ResNet-50	29.07	21.85	31.23	31.68	24.14	33.93
MSTR [15]	50	ResNet-50	31.17	25.31	32.92	34.02	28.83	35.57
GEN-VLKT-S [21]	90	ResNet-50	33.75	29.25	35.10	36.78	32.75	37.99
HOICLIP [27]	90	ResNet-50	34.69	31.12	35.74	37.61	34.47	38.54
CLIP4HOI [26]	100	ResNet-50	35.33	33.95	35.74	37.19	35.27	37.77
HOLa [17]	12	ResNet-50	39.05	38.66	39.17	-	-	-
SICHOI [24]	30	ResNet-50	41.79	42.38	41.61	44.27	43.64	44.46
QAHOI [6]	150	Swin-L	35.78	29.80	37.56	37.59	31.36	39.36
FGAHOI [25]	190	Swin-L	37.18	30.71	39.11	38.93	31.93	41.02
PViC [37]	30	Swin-L	44.32	**44.61**	44.24	**47.81**	**48.38**	**47.64**
SOV-STG-VLA [4]	15	ResNet-50	41.16	39.48	41.67	43.81	42.63	44.17
Hybrid-SOV-VLA [5]	15	ResNet-50	43.10	43.04	43.12	46.02	46.14	45.98
HOIBlender-N	15	DINOv2-S	41.36	41.64	41.41	44.26	45.71	43.83
HOIBlender-S	15	DINOv2-S	43.19	43.07	43.23	46.17	47.02	45.92
HOIBlender-2XL	9	DINOv2-B	**44.49**	43.21	**44.86**	47.45	47.76	47.36

Table 1 Comparison on HICO-DET.

Method	Default (Hybrid-SOV vs. HOIBlender)					
	Full		*Rare*		*Non-Rare*	
baseline	43.10	**44.49**	43.04	**43.21**	43.12	**44.86**
w/o Verb Dec.	43.05	43.92	42.09	42.11	43.33	44.46
w/o Verb & VLA	32.36*	34.26	23.92*	24.69	34.88*	37.12
w/o Subject & VLA	35.50	36.97	30.59	30.65	36.96	38.85

Table 2 Component ablations under the Default protocol. Each split shows Hybrid-SOV-VLA-R50 (left) and HOIBlender-2XL (right). * indicates results reproduced by our implementation.

Group Num.	Full	Rare	Non-Rare
13	**41.36**	**41.64**	**41.41**
6	40.69	40.55	40.73
1	37.27	39.04	36.74

Table 3 Ablation on the number of group heads (HOIBlender-N). "1" denotes training without additional group heads.

groups yield the best result across all splits, while 6 or 1 group leads to noticeable drops, especially on rare categories, validating that grouped-query training improves optimization without complex denoising-query design.

Module contribution. Table 2 ablates the verb decoder, the BLIP-2 vision-language advisor (VLA), and the subject decoder on Hybrid-SOV-VLA-R50 and HOIBlender-2XL. Removing the verb decoder yields only a small Full drop on HOIBlender-2XL (44.49 → 43.92) but a larger Rare drop (43.21 → 42.11), confirming that the verb branch is most useful for tail interactions. Further removing the VLA collapses Rare mAP to 24.69, isolating multi-prior semantic fusion as the dominant source of long-tail robustness. Disabling the subject decoder together with VLA yields 36.97 Full and 30.65 Rare, showing that the dual-stream spatial design is also load-bearing. The same trend holds on Hybrid-SOV-VLA-R50, indicating that dual-stage decoding and prior fusion remain effective after removing the encoder and upgrading to DINOv2.

4.5 Efficiency Analysis

Figure 3 visualizes the tradeoff among detection quality, inference speed, and model size. We use Default Full mAP and per-image latency (ms, averaged over 100 runs on a single NVIDIA A100 GPU with batch size 1) as axes and

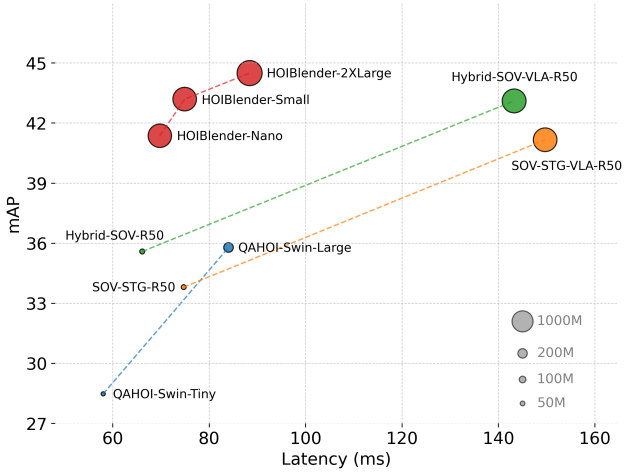


Fig. 3 Performance-efficiency-parameter tradeoff on HICO-DET. The x-axis is latency time (ms), the y-axis is mAP, and bubble size indicates trainable parameters.

encode trainable parameters by bubble size. HOIBlender-S achieves competitive mAP at lower latency than SOV-STG-VLA and Hybrid-SOV-VLA, while HOIBlender-2XL pushes the accuracy frontier with a moderate latency increase, showing that efficient detection and interaction-aware decoding can be combined effectively.

5. Conclusion

We presented HOIBlender, an efficient HOI detection framework that combines an RF-DETR-style detection backbone, dual-stage HOI decoding, and multi-prior semantic fusion. By coupling spatial interaction construction with semantic interaction reasoning and injecting external vision-language priors, HOIBlender improves visual-semantic alignment and strengthens HOI recognition under a practical training and inference pipeline. Experiments on HICO-DET show that this design yields a strong efficiency–accuracy tradeoff compared with recent baselines. Future work includes exploring more advanced vision-language priors, extending to video-based HOI detection, and further optimizing the architecture for real-time applications.

References

[1] Cao, Y., Tang, Q., Su, X., Chen, S., You, S., Lu, X. and Xu, C.: Detecting any human-object interaction relationship: Universal hoi detector with spatial prompt learning on foundation models, *Proc. Adv. Neural Inform. Process. Syst. (NIPS)* (Dec. 2024).

[2] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S.: End-to-end object detection with transformers, *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 213–229 (Aug. 2020).

[3] Chao, Y.-W., Liu, Y., Liu, X., Zeng, H. and Deng, J.: Learning to Detect Human-Object Interactions, *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 381–389 (Mar. 2018).

[4] Chen, J., Wang, Y. and Yanai, K.: Focusing on what to Decode and what to Train: SOV Decoding with Specific Target Guided DeNoising and Vision Language Advisor, *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 9416–9425 (Feb. 2025).

[5] Chen, J. and Yanai, K.: Bridging Detection Architectures With Foundation Models: A Unified Framework for Human–Object Interaction Detection, *IEEE Access*, Vol. 14, pp. 23299–23310 (2026).

[6] Chen, J. and Yanai, K.: QAHOI: Query-based anchors for human-object interaction detection, *Proc. Int. Conf. Mach. Vis. Appl. (MVA)*, pp. 1–5 (Jul. 2023).

[7] Chen, M., Liao, Y., Liu, S., Chen, Z., Wang, F. and Qian, C.: Reformulating HOI detection as adaptive set prediction, *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* (Jun. 2021).

[8] Chen, Q., Chen, X., Wang, J., Zhang, S., Yao, K., Feng, H., Han, J., Ding, E., Zeng, G. and Wang, J.: Group detr: Fast detr training with group-wise one-to-many assignment, *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)* (2023).

[9] Chen, Q., Su, X., Zhang, X., Wang, J., Chen, J., Shen, Y., Han, C., Chen, Z., Xu, W., Li, F. et al.: LW-DETR: A Transformer Replacement to YOLO for Real-Time Detection, *arXiv preprint arXiv:2406.03459* (2024).

[10] Gao, C., Zou, Y. and Huang, J.-B.: iCAN: Instance-centric attention network for human-object interaction detection, *Proc. Brit. Mach. Vis. Conf. (BMVC)* (Sep. 2018).

[11] Gkioxari, G., Girshick, R., Dollár, P. and He, K.: Detecting and Recognizing Human-Object Interactions, *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 8359–8367 (Jun. 2018).

[12] Gupta, T., Schwing, A. and Hoiem, D.: No-Frills Human-Object Interaction Detection: Factorization, Layout Encodings, and Training Techniques, *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 9676–9684 (Oct. 2019).

[13] Iftekhar, A., Chen, H., Kundu, K., Li, X., Tighe, J. and Modolo, D.: What to look at and where: Semantic and Spatial Refined Transformer for detecting human-object interactions, *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* (2022).

[14] Kim, B., Lee, J., Kang, J., Kim, E.-S. and Kim, H. J.: HOTR: End-to-end human-object interaction detection with transformers, *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* (Jun. 2021).

[15] Kim, B., Mun, J., On, K.-W., Shin, M., Lee, J. and Kim, E.-S.: MSTR: Multi-Scale Transformer for End-to-End Human-Object Interaction Detection, *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 19556–19565 (Jun. 2022).

[16] Kuhn, H. W.: The Hungarian method for the assignment problem, *Naval Res. Logist. Quart*, pp. 83–97 (Mar. 1955).

[17] Lei, Q., Wang, B. and Tan, R. T.: HOLa: Zero-Shot HOI Detection with Low-Rank Decomposed VLM Feature Adaptation, *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 1825–1835 (Oct. 2025).

[18] Li, F., Zhang, H., Liu, S., Guo, J., Ni, L. M. and Zhang, L.: DN-DETR: Accelerate DETR Training by Introducing Query DeNoising, *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2239–2251 (Jun. 2022).

[19] Li, J., Li, D., Savarese, S. and Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 19730–19742 (Jul. 2023).

[20] Li, J., Li, D., Xiong, C. and Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 12888–12900 (Jul. 2022).

[21] Liao, Y., Zhang, A., Lu, M., Wang, Y., Li, X. and Liu, S.: GEN-VLKT: Simplify Association and Enhance Interaction Understanding for HOI Detection, *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 20091–20100 (Jun. 2022).

[22] Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J. and Zhang, L.: DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR, *Proc. Int. Conf. Learn. Represent. (ICLR)* (Apr. 2022).

[23] Loshchilov, I. and Hutter, F.: Decoupled Weight Decay Regularization, *Proc. Int. Conf. Learn. Represent. (ICLR)* (2018).

[24] Luo, J., Ren, W., Jiang, W., Chen, X., Wang, Q., Han, Z. and Liu, H.: Discovering syntactic interaction clues for human-object interaction detection, *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 28212–28222 (Jun. 2024).

[25] Ma, S., Wang, Y., Wang, S. and Wei, Y.: FGAHOI: Fine-Grained Anchors for Human-Object Interaction Detection, *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 2415–2429 (Apr. 2024).

[26] Mao, Y., Deng, J., Zhou, W., Li, L., Fang, Y. and Li, H.: CLIP4HOI: Towards Adapting CLIP for Practical Zero-Shot HOI Detection, *Proc. Adv. Neural Inform. Process. Syst. (NIPS)* (Dec. 2024).

[27] Ning, S., Qiu, L., Liu, Y. and He, X.: HOICLIP: Efficient Knowledge Transfer for HOI Detection with Vision-Language Models, *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 23507–23517 (Jun. 2023).

[28] Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A. et al.: DINOv2: Learning Robust Visual Features without Supervision, *Proc. Int. Conf. Learn. Represent. (ICLR)* (Apr. 2025).

[29] Qi, S., Wang, W., Jia, B., Shen, J. and Zhu, S.-C.: Learning Human-Object Interactions by Graph Parsing Neural Networks, *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 407–423 (Sep. 2018).

[30] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al.: Learning Transferable Visual Models From Natural Language Supervision, *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 8748–8763 (Jul. 2021).

[31] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I. and Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression, *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* (2019).

[32] Robinson, I., Robicheaux, P., Popov, M., Ramanan, D. and Peri, N.: RF-DETR: neural architecture search for real-time detection transformers, *arXiv preprint arXiv:2511.09554* (2025).

[33] Tamura, M., Ohashi, H. and Yoshinaga, T.: QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information, *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 10405–10414 (Jun. 2021).

[34] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I.: Attention is all you need, *Proc. Adv. Neural Inform. Process. Syst. (NIPS)* (Dec. 2017).

[35] Yuan, H., Jiang, J., Albanie, S., Feng, T., Huang, Z., Ni, D. and Tang, M.: RLIP: Relational language-image pre-training for human-object interaction detection, *Proc. Adv. Neural Inform. Process. Syst. (NIPS)* (Nov. 2022).

[36] Yuan, H., Zhang, S., Wang, X., Albanie, S., Pan, Y., Feng, T., Jiang, J., Ni, D., Zhang, Y. and Zhao, D.: RLIPv2: Fast Scaling of Relational Language-Image Pre-training, *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 21592–21604 (Oct. 2023).

[37] Zhang, F. Z., Yuan, Y., Campbell, D., Zhong, Z. and Gould, S.: Exploring Predicate Visual Context in Detecting Human-Object Interactions, *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 10377–10387 (Oct. 2023).

[38] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y. and Chen, J.: Detrs beat yolos on real-time object detection, *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 16965–16974 (Jun. 2024).

[39] Zhong, X., Ding, C., Li, Z. and Huang, S.: Towards hard-positive query mining for detr-based human-object interaction detection, *Proc. Eur. Conf. Comput. Vis. (ECCV)*, p. 444–460 (Oct. 2022).

[40] Zhou, D., Liu, Z., Wang, J., Wang, L., Hu, T., Ding, E. and Wang, J.: Human-Object Interaction Detection via Disentangled Transformer, *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* (2022).

[41] Zhu, X., Su, W., Lu, L., Li, B., Wang, X. and Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection, *Proc. Int. Conf. Learn. Represent. (ICLR)* (Apr. 2020).

[42] Zou, C., Wang, B., Hu, Y., Liu, J., Wu, Q., Zhao, Y., Li, B., Zhang, C., Zhang, C., Wei, Y. et al.: End-to-end human object interaction detection with hoi transformer, *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 11820–11829 (Jun. 2021).