STIP+: Context-Aware Human–Object Interaction Detection with Adaptive Relation Sampling

Yanxi Lin[1], Noha Sarhan[1], and Simone Frintrop[1]

University of Hamburg, Department of Informatics, Hamburg, Germany
{yanxi.lin, noha.sarhan, simone.frintrop}@uni-hamburg.de

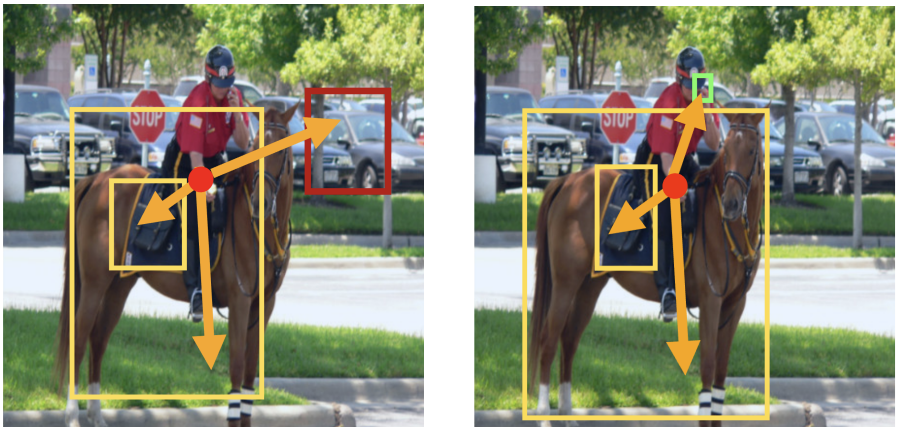
Abstract. While advanced transformer-based detectors have revolutionized object detection, their potential remains largely unexplored in Human–Object Interaction (HOI) tasks, where conventional two-stage frameworks still struggle with small objects and noisy interaction proposals. To bridge this gap, we propose **STIP+**, an enhanced framework building upon the structure-aware formulation of STIP [28] that integrates a strong transformer-based detector with a context-aware interaction proposal mechanism. First, we adopt DINO with a Swin-L backbone to provide robust instance localization for crowded scenes and small-object scenarios. Second, we introduce an Adaptive Relation Sampling strategy that incorporates spatial proximity, detection confidence, and semantic cues to filter interaction candidates, effectively replacing indiscriminate hard mining with a plausibility-based selection. Finally, a Structure-Aware Transformer aggregates these refined proposals for classification. Extensive experiments on V-COCO and HICO-DET demonstrate that STIP+ consistently outperforms state-of-the-art methods, with notable improvements in rare and small-object interactions.

Keywords: Human–Object Interaction · Object Detection · Transformer · Adaptive Sampling

1 Introduction

Human–object interaction (HOI) detection aims to recognize not only objects in an image but also the semantic relationships between humans and objects, identifying interaction triplets such as ⟨human, ride, bike⟩ [6]. By explicitly modeling how humans interact with surrounding objects, HOI detection serves as a critical building block for high-level visual reasoning in applications ranging from autonomous driving to human-robot collaboration.

Recent progress in transformer-based object detection [18] has significantly improved object localization and recognition, providing stronger visual foundations for HOI reasoning [1]. However, despite these gains, two critical bottlenecks remain. First, standard DETR (DEtection TRansformer)-based detectors [2] often struggle to resolve small objects and distinct instances in crowded scenes,



(a) Baseline (STIP [28]): Missed Object (b) **STIP+ (Ours)**: Robust

Fig. 1: **Failure Case Analysis: Small Objects.** Visual comparison of detection sensitivity. **(a)** The baseline method (STIP [28]) fails to resolve the small object, missing the phone and incorrectly associating the human with background clutter (red box). **(b)** STIP+ leverages robust representations to correctly localize the challenging "cell phone" (green box) and detect the ⟨human, talk_on, phone⟩ interaction.

leading to missed interactions (e.g., small objects such as cell phones, as illustrated in Figure 1) [30]. Second, as detection quality improves, the number of candidate pairs grows rapidly, introducing severe noise [22]. To manage this, recent methods such as STIP [28] employ "hard mining" strategies [22], which prioritize training on samples with the highest loss magnitudes, to focus training on difficult pairs. While effective in some cases, hard mining implicitly assumes that the hardest samples are the most informative. In HOI detection, this assumption often fails: hard pairs frequently arise from annotation errors, visually ambiguous configurations, or semantically implausible human–object combinations, especially in crowded scenes and long-tail interaction categories. Consequently, hard mining can amplify noise and induce spurious interaction predictions (hallucinations) [29].

In this paper, we argue that effective HOI learning requires selecting interaction candidates that are not merely hard, but structurally plausible. To directly address the critical bottlenecks of indiscriminate hard mining and small-object failure, we introduce a targeted architectural refinement for two-stage pipelines. Specifically, we propose an Adaptive Relation Sampling strategy that explicitly models the plausibility of human–object pairs using complementary cues from detection confidence, semantic compatibility, and spatial proximity. By prioritizing candidate relations that are both informative and semantically coherent, our approach systematically reduces noisy supervision while preserving meaningful interaction patterns.

We integrate this strategy into **STIP+**, a unified framework featuring a Context-Aware Interaction Proposal Network (CIPN) and a state-of-the-art, transformer-based detector, DINO (with Swin-L) [27]. As shown in Figure 1 and empirically validated in Section 4, these components enable STIP+ to recover subtle interactions while suppressing spurious predictions. By leveraging Contrastive Denoising, the detector recovers fine-grained spatial details for small objects, while our adaptive sampling refines interaction candidates for robust HOI reasoning. Extensive experiments on V-COCO [6] and HICO-DET [3] benchmarks demonstrate that STIP+ generates higher-quality proposals and achieves consistent improvements over state-of-the-art methods.

Our main contributions are summarized as follows:

- **Adaptive Relation Sampling:** We propose a sampling mechanism driven by plausibility that mitigates spurious human–object pairs using detection confidence, semantic compatibility, and spatial proximity, addressing the limitations of conventional hard mining in HOI detection.
- **Context-Aware Interaction Framework:** We incorporate this mechanism into an enhanced architecture (**STIP+**) that integrates the DINO detector to robustly capture small objects and fine-grained spatial features.
- **State-of-the-Art Performance:** STIP+ achieves consistent gains across V-COCO and HICO-DET benchmarks, establishing competitive performance among two-stage methods by robustly recovering rare and small-object interactions.

2 Related Work

This section contextualizes our proposed STIP+ framework within the broader landscape of scene understanding. We first trace the evolution of foundational HOI detection paradigms and modern transformer-based object detectors, followed by an analysis of recent structure-aware architectures and the growing trend of vision-language models.

Evolution of HOI Detection. Early HOI detection methods largely followed a two-stage paradigm, extracting region proposals before relation classification [5,4]. To reduce inference latency, one-stage approaches later emerged, predicting human-object interactions directly from global feature maps or point-based representations [12,29]. Recently, end-to-end transformer-based models [10,31] have further streamlined the architecture by treating HOI as a set prediction problem. However, two-stage frameworks remain highly advantageous because they decouple instance localization from interaction reasoning, allowing the integration of powerful, pre-trained object detectors without catastrophic forgetting [1].

Transformer-Based Object Detectors. The evolution of object detection has profoundly impacted two-stage HOI performance. While earlier models relied heavily on CNN backbones like ResNet [7], the introduction of DETR [2] and its deformable variants [30] demonstrated the efficacy of self-attention for global scene understanding. More recently, advanced architectures such as DINO [27]

paired with Swin Transformers [16] have set new standards for localizing challenging, small-scale instances. STIP+ deliberately harnesses these robust, fine-grained visual representations to prevent interaction failures caused by upstream detection errors, particularly in crowded scenes.

Structure-Aware Two-Stage Architectures. Building upon robust localizers, modern two-stage HOI detectors focus heavily on refining the interaction classification module. Methods like QPIC [23], UPT [26], and STIP [28] have established strong baselines by modeling intra- and inter-interaction spatial structures. Nevertheless, as the sensitivity of the underlying detector increases, these frameworks become vulnerable to a surge of noisy interaction proposals, which standard hard-mining techniques [22] often amplify. Our framework directly addresses this bottleneck by replacing indiscriminate hard mining with a plausibility-driven adaptive sampling mechanism. We note that this approach is highly complementary to recent methods like PVIC [13], which inject learned predicate visual context into frozen detectors to achieve highly competitive HICO-DET performance; STIP+ explores an orthogonal direction by focusing on the proposal filtering stage.

Vision-Language Models in HOI. Finally, recent advancements have increasingly leveraged Vision-Language Models (VLMs) and diffusion priors to improve generalization. Methods such as HOICLIP [19], RLIP-v2 [25], and DiffHOI [24] utilize massive external text corpora to guide interaction discovery. While these approaches achieve impressive zero-shot capabilities, they introduce substantial computational overhead. In contrast, STIP+ demonstrates that highly competitive, state-of-the-art results can be achieved through a purely visual, spatial-first architecture, avoiding the dependency on external language priors while maintaining high efficiency in complex environments.

3 Method

We propose **STIP+**, a two-stage framework for Human–Object Interaction (HOI) detection that addresses both the representational limitations of prior DETR-based architectures and the noise sensitivity of existing interaction proposal strategies. As illustrated in Figure 2, STIP+ consists of three components: (1) a robust object detector based on DINO for accurate instance localization, (2) a Context-Aware Interaction Proposal Network (CIPN) with Adaptive Relation Sampling, and (3) a Structure-Aware Transformer for interaction classification.

3.1 Object Detector Integration

To overcome the limitations of standard DETR-based detectors on small objects and crowded scenes, we employ the DINO detector equipped with a Swin-L backbone [27]. Given an input image, DINO produces a set of object instances $\mathcal{D} = \{(b_i, c_i, s_i)\}$, where b_i, c_i, and s_i denote the bounding box, category label, and detection confidence, respectively. By leveraging Contrastive DeNoising, DINO stabilizes training and preserves fine-grained spatial details essential for

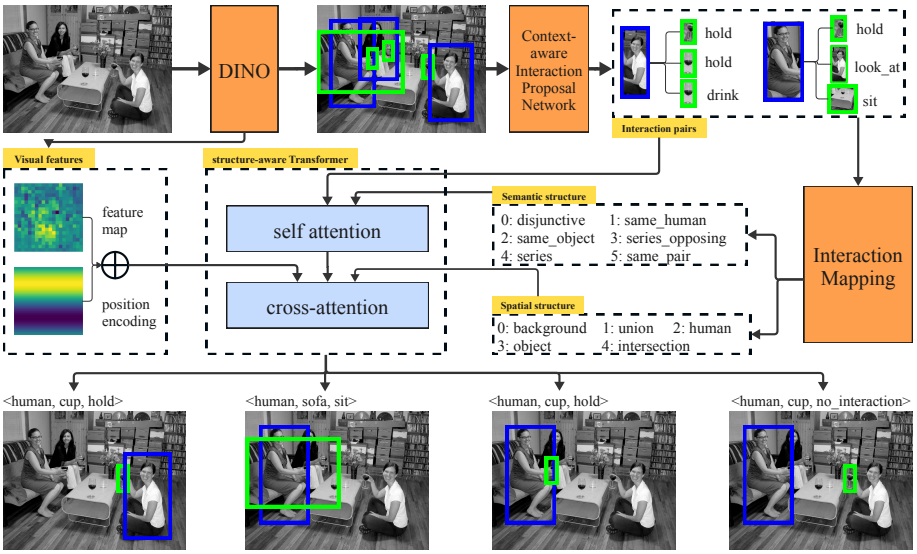


Fig. 2: Overview of the STIP+ framework. Given an input image, the DINO detector produces human and object instances. A Context-Aware Interaction Proposal Network (CIPN) applies Adaptive Relation Sampling to select semantically and spatially plausible human–object pairs. A Structure-Aware Transformer integrates visual features with explicit spatial and semantic structures to generate the final HOI predictions.

identifying small interacting objects. Crucially, we freeze the entire detector during training to leverage these robust representations without the computational cost of fine-tuning.

Unless otherwise specified, we initialize the DINO detector with publicly available pre-trained weights and freeze the entire detection module (including the backbone and detection heads) during HOI training, updating only the HOI-specific components. This design leverages robust pre-trained visual representations while avoiding the computational overhead and instability of detector fine-tuning. Unlike the original STIP [28], which relies on a ResNet backbone and hard mining, STIP+ integrates the DINO detector (Swin-L) and replaces hard mining with plausibility-based adaptive sampling.

3.2 Context-Aware Interaction Proposal Network (CIPN)

Standard proposal networks often rely on hard mining, which indiscriminately emphasizes difficult samples, including outliers and annotation errors. As shown in Figure 2, we replace this paradigm with a CIPN incorporating principled feature encoding and plausibility-driven adaptive sampling.

Heterogeneous Feature Encoding. For each candidate human–object pair (h, o), we construct a feature representation f_{pair} by fusing four complementary modalities: (1) *Appearance:* ROI Align extracts visual features for the human, object, and their union region; (2) *Spatial:* geometric relations including relative distances, angles, and Intersection-over-Union (IoU); (3) *Semantic:* object class embeddings capture label-level dependencies; (4) *Context:* union-region features provide interaction-level cues. These features are fused via a Multi-Layer Perceptron (MLP) to predict a preliminary interaction score S_{rel}. To handle class imbalance, we employ Focal Loss [14]:

$$\mathcal{L}_{prop} = \frac{1}{N_{pos}} \sum_{i} FL(S_{rel}^{(i)}, y_i), \tag{1}$$

where $y_i \in \{0, 1\}$ indicates pair validity and N_{pos} denotes the number of positive pairs. Because this multi-modal encoding strategy rigorously follows the established structural baseline of STIP [28], we maintain this formulation to specifically isolate and evaluate the impact of our novel sampling mechanism.

Adaptive Relation Sampling. To mitigate noisy interaction proposals, we introduce an Adaptive Relation Sampling strategy—termed "adaptive" because it dynamically adjusts the plausibility prior for each candidate pair based on its unique spatial and semantic context, rather than relying on a static, dataset-wide loss threshold. Instead of selecting pairs based solely on loss (difficulty), we compute a composite plausibility score $S(h, o)$ for each pair:

$$S(h, o) = S_{rel} + \lambda_1 S_o + \lambda_2 S_h + \lambda_3 S_{cnt} + \lambda_4 S_{prox}, \tag{2}$$

where S_{rel} is the predicted relation score from the MLP, S_o and S_h are the object and human detection confidences from DINO, S_{cnt} denotes human centrality (normalized distance from image center), and S_{prox} is the spatial proximity. This formulation encourages the proposal network to prioritize pairs that are jointly supported by semantic, spatial, and detection evidence, thereby reducing the influence of noisy or implausible interactions. All scores are normalized to comparable ranges, and the weighting coefficients $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are fixed across experiments. Based on $S(h, o)$, we select the top-K candidate pairs for subsequent interaction classification.

Agent-First Design Principle. A core challenge in conventional interaction filtering is the over-reliance on raw, uncalibrated object detections. In our formulation, we establish an explicit, hierarchical prior driven by human agency. Specifically, we assign the highest weight to human detection confidence ($\lambda_2 = 0.4$), followed by spatial proximity ($\lambda_4 = 0.3$), while object confidence and centrality are weighted equally ($\lambda_1 = \lambda_3 = 0.2$). This configuration is derived from the fundamental nature of HOI: humans are the primary initiators of interactions. A low-confidence human prediction inherently corrupts the interaction tuple, making it the most critical filtering signal, while spatial proximity serves as a near-universal prerequisite for contact-based interactions.

3.3 Structure-Aware Interaction Classification

The selected interaction proposals are classified using a Structure-Aware Transformer (Figure 2). Following STIP [28], we explicitly encode both *intra-interaction spatial structure* (relative layouts between humans and objects) and *inter-interaction semantic structure* (dependencies among multiple interactions within an image). A transformer decoder aggregates visual, spatial, and semantic embeddings via cross-attention and outputs the final HOI action predictions.

4 Experiments

In this section, we comprehensively evaluate the proposed STIP+ framework to validate its effectiveness, particularly in challenging scenarios involving small objects and crowded scenes. We begin by detailing the benchmark datasets and implementation specifics, followed by a quantitative comparison against state-of-the-art methods. Finally, we present qualitative robustness analyses and targeted ablation studies to isolate the impact of our architectural design choices.

4.1 Datasets and Metrics

We evaluate STIP+ on two standard HOI benchmarks: V-COCO [6], a subset of the widely studied COCO ecosystem [15,20], and HICO-DET [3].

V-COCO comprises 10,346 images in total, split into 2,533 images for training, 2,867 for validation, and 4,946 for testing. It includes annotations for 24 distinct action classes. Following the standard V-COCO protocol, we report $\mathrm{AP}^{\mathrm{S1}}_{\mathrm{role}}$ and $\mathrm{AP}^{\mathrm{S2}}_{\mathrm{role}}$. As defined in [6], $\mathrm{AP}_{\mathrm{role}}$ evaluates the correctness of the human–object–action triplet. Scenario 1 ($S1$) requires the model to correctly predict occluded objects as empty, whereas Scenario 2 ($S2$) ignores interactions with occluded objects.

HICO-DET is a large-scale dataset containing 47,776 images, split into 38,118 for training and 9,658 for testing. It features 600 specific HOI triplet categories constructed from 117 distinct verbs and 80 COCO object classes. For HICO-DET, we report mAP under both *Default* and *Known Object* settings, each further divided into *Full*, *Rare* (fewer than 10 training instances), and *Non-Rare* subsets to comprehensively evaluate long-tail robustness.

4.2 Implementation Details

We initialize the detector with Objects365-pretrained DINO-SwinL [21], freezing its weights. We optimize the model using AdamW [17] with batch size 4 for 40 epochs. The initial learning rate is 5×10^{-5} and is reduced by a factor of 0.1 if the validation metric does not improve for two consecutive epochs (V-COCO: $\mathrm{AP}^{\mathrm{S1}}_{\mathrm{role}}$; HICO-DET: Default Full mAP). Following standard practice, we select the top-K ($K = 32$) interaction proposals from the CIPN for final classification. The adaptive sampling weights are set to $\lambda_1 = 0.2$, $\lambda_2 = 0.4$, $\lambda_3 = 0.2$, and $\lambda_4 = 0.3$. All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU (24GB).

Table 1: Comparison with state-of-the-art methods on the V-COCO benchmark. Results are reported using AP^{S1}_{role} and AP^{S2}_{role} following the standard evaluation protocol.

Method	**Backbone**	AP^{S1}_{role}	AP^{S2}_{role}
InteractNet [5]	ResNet-50-FPN	40.0	48.0
HOI-Trans [31]	ResNet-101	52.9	–
QPIC [23]	ResNet-101	58.3	60.7
UPT [26]	ResNet-50	59.0	64.5
HOICLIP [19]	ResNet-50	63.5	–
DiffHOI-L [24]	Swin-L	65.7	68.2
STIP [28]	ResNet-50	66.0	70.7
DP-HOI [11]	ResNet-50	66.6	–
STIP+ (Ours)	**Swin-L**	**68.7**	**72.4**

Table 2: Comparison with state-of-the-art methods on the HICO-DET benchmark (mAP). Results are reported under the *Default* and *Known Object* settings, each evaluated on the *Full*, *Rare*, and *Non-Rare* subsets.

Method	**Backbone**	**Default**			**Known Object**		
		Full	Rare	Non-Rare	Full	Rare	Non-Rare
QPIC [23]	ResNet-101	29.09	23.69	31.69	32.38	26.06	34.27
STIP [28]	ResNet-50	28.81	27.55	29.18	32.28	32.07	32.64
STIP+ (Ours)	**Swin-L**	**32.31**	**28.55**	**33.44**	**35.51**	**33.57**	**36.09**

4.3 Comparison with State-of-the-Art

Tables 1 and 2 compare STIP+ with existing two-stage methods. Note that while most baselines utilize ResNet backbones, STIP+ adopts a Swin-L backbone, similar to DiffHOI-L [24], to provide higher-capacity visual representations. On **V-COCO**, STIP+ achieves **68.7** AP^{S1}_{role} and **72.4** AP^{S2}_{role}. Crucially, STIP+ not only improves upon its direct structural baseline (STIP) by **+2.7** (S1) and **+1.7** (S2), but it also outperforms recent, heavy-weight generative and Vision-Language approaches such as HOICLIP (CVPR 2023) and DP-HOI (CVPR 2024). Despite lacking the rich, zero-shot semantic space provided by massive CLIP pre-training, STIP+ surpasses DP-HOI's 66.6 S1 score, proving that our refined, purely visual spatial prior can establish state-of-the-art performance without the computational overhead of VLM feature extraction. On **HICO-DET**, STIP+ reaches **32.31** mAP (Default Full) and yields a clear gain on the long-tail **Rare** subset (**28.55**, i.e., **+1.00** over STIP), supporting the effectiveness of our context-aware proposal filtering.

4.4 Qualitative Results

We analyze STIP+ in two failure scenarios common to two-stage detectors: small object interactions and crowded scenes.

Small Objects (Fig. 1). As illustrated in Figure 1, baselines often fail to resolve small objects (e.g., cell phones) due to low detection confidence. STIP+ leverages the high-fidelity DINO backbone to successfully localize these targets and recover the *talk on phone* interaction (Figure 1(b)).

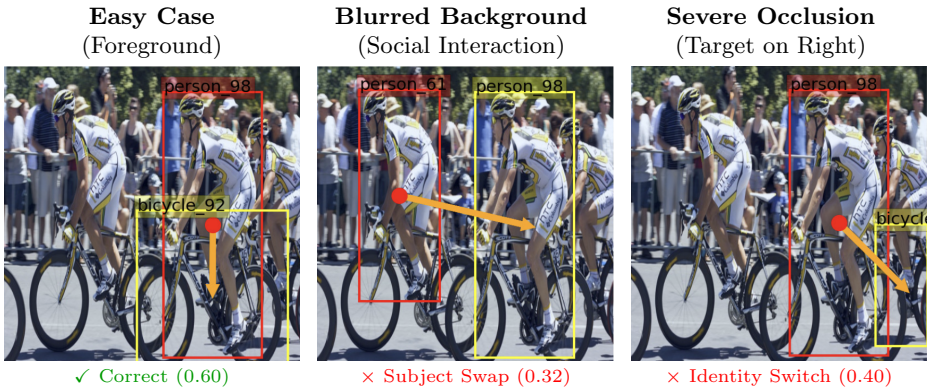
Crowded Scenes (Fig. 3). Figure 3 analyzes robustness across three difficulty levels. First, in the easy foreground case (Column 1), all models correctly predict the *ride* interaction, though STIP+ yields the highest confidence (0.68). Second, in the social interaction scenario (Column 2), STIP fails due to a semantic subject–object swap, whereas UPT and STIP+ correctly identify the *look* relation. However, STIP+ detects this with significantly higher confidence (0.66 vs. 0.23). Third, under severe occlusion (Column 3), baseline methods collapse: STIP commits an identity switch, while UPT fails to localize the interaction entirely. In contrast, STIP+ successfully resolves the occlusion, correctly associating the obscured rider with their bicycle.

4.5 Ablation Studies

To rigorously validate the architectural decisions within STIP+, we conduct targeted ablation experiments on the V-COCO dataset. Specifically, we first evaluate the impact of the underlying object detector to justify our integration of the high-capacity DINO-SwinL model. Subsequently, we isolate our proposed Adaptive Relation Sampling mechanism to demonstrate its distinct advantage over conventional hard-mining strategies.

Detector Choice. We first investigate the impact of the object detector on overall HOI performance. To isolate this factor, we fix the HOI head and sampling strategy (hard mining) while varying the detector architecture. Table 3 compares the performance of RT-DETRv2-L against two variants of DINO. Despite being a state-of-the-art real-time detector, RT-DETRv2-L yields substantially lower HOI performance (≈13% drop in AP). We hypothesize that its optimization for inference speed compromises the fine-grained feature fidelity required for complex interaction reasoning. In contrast, DINO-based models perform significantly better. Specifically, DINO-SwinL outperforms DINO-Res50 by **+1.2 AP** on $\text{AP}^{\text{S1}}_{\text{role}}$, confirming that high-capacity backbones are critical for resolving spatially ambiguous interactions in crowded scenes.

Impact of Adaptive Relation Sampling. We next isolate our proposed sampling mechanism, fixing the detector to DINO-SwinL and comparing our plausibility-based sampling against the standard hard mining baseline. As shown in Table 4, Adaptive Relation Sampling improves $\text{AP}^{\text{S1}}_{\text{role}}$ by **+0.9** and $\text{AP}^{\text{S2}}_{\text{role}}$ by **+0.2**. Unlike hard mining, which blindly focuses on high-loss outliers, our adaptive strategy leverages complementary cues to identify high-quality training targets. The spatial proximity term (S_{prox}) proved particularly effective in suppressing physically implausible pairs in crowded scenes, while the detection



(a) STIP [28]: Fails on background/occlusion

(b) UPT [26]: Low confidence / Inconsistent

(c) **STIP+ (Ours)**: High confidence & Robustness

Fig. 3: **Crowded Scene Robustness.** Scores in parentheses indicate prediction confidence. **Col. 1:** All methods detect the easy case, but STIP+ yields the highest confidence (0.68). **Col. 2:** STIP+ correctly identifies the social interaction with significantly higher confidence (0.66) than UPT (0.23). **Col. 3:** Only STIP+ (c) resolves the severe occlusion, while baselines fail.

Table 3: Ablation on V-COCO: effect of object detector choice (hard mining).

Detector	AP^{S1}_{role}	AP^{S2}_{role}
RT-DETRv2-L	54.4	58.8
DINO-Res50	66.6	70.9
DINO-SwinL	**67.8**	**72.2**

Table 4: Ablation on V-COCO: effect of Adaptive Relation Sampling (DINO-SwinL).

Sampling Strategy	AP^{S1}_{role}	AP^{S2}_{role}
Hard Mining	67.8	72.2
Adaptive Relation Sampling	**68.7**	**72.4**

confidence terms (S_o, S_h) helped filter out hallucinated interactions arising from weak object proposals. This highlights that even with strong detectors, principled interaction modeling is essential. Crucially, because both configurations in Table 4 utilize the exact same frozen DINO-SwinL detector, this +0.9 AP gain is entirely attributable to the Adaptive Relation Sampling mechanism itself, explicitly proving that the performance improvements stem from our proposed methodological formulation rather than merely scaling the detector.

4.6 Limitations

While STIP+ demonstrates strong robustness in standard crowded environments, its reliance on 2D bounding box proximity (S_{prox}) introduces a specific failure mode. In scenarios with extreme depth ambiguity—such as a person in the foreground appearing adjacent to a small object far in the background—the 2D spatial prior may incorrectly elevate the plausibility score of a spurious interaction, or conversely, suppress a valid but visually distant interaction. Resolving these forced-perspective ambiguities remains an open challenge for purely visual, 2D HOI frameworks and points toward the future integration of explicit depth estimation or 3D scene modeling.

5 Conclusion

In this paper, we presented STIP+, a unified framework for Human–Object Interaction detection that addresses the dual challenges of feature representation and proposal noise. By replacing conventional hard mining with an adaptive relation sampling mechanism, our method prioritizes interaction candidates based on a composite plausibility score, effectively suppressing spurious predictions while preserving rare interactions. Coupled with the robust DINO detector (Swin-L), STIP+ overcomes the limitations of prior methods in resolving small objects

and crowded scenes. Extensive experiments on V-COCO and HICO-DET confirm that these contributions yield state-of-the-art performance, with significant gains in detecting small-object interactions and long-tail categories where standard baselines often fail.

Looking forward, several directions promise to further enhance HOI understanding. First, while our current frozen-detector strategy ensures stability, applying parameter-efficient fine-tuning (e.g., LoRA [9] or Adapters [8]) could better align detection features with interaction tasks without the cost of full training. Second, extending the framework to an open-vocabulary setting by leveraging vision–language models (e.g., CLIP [19]) would allow for scaling to novel, natural language-described interactions. Finally, incorporating temporal cues from video data or exploring end-to-end optimization strategies could bridge the gap between detection and interaction reasoning, enabling more holistic scene understanding.

References

1. Antoun, M., Asmar, D.: Human object interaction detection: Design and survey. Image and Vision Computing **130**, 104617 (2023)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proc. ECCV. pp. 213–229. Springer (2020)
3. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: Proc. IEEE WACV. pp. 381–389 (2018)
4. Gao, C., Zou, Y., Huang, J.B.: iCAN: Instance-centric attention network for human-object interaction detection. In: Proc. BMVC (2018)
5. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: Proc. IEEE CVPR. pp. 8359–8367 (2018)
6. Gupta, S., Malik, J.: Visual semantic role labeling. In: Proc. IEEE CVPR. pp. 3191–3200 (2015)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE CVPR. pp. 770–778 (2016)
8. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. In: Proc. ICML. pp. 2790–2799 (2019)
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: Proc. ICLR (2022)
10. Kim, B., Lee, J., Kang, J., Kim, E.S., Kim, H.J.: HOTR: End-to-end human-object interaction detection with transformers. In: Proc. IEEE CVPR. pp. 74–83 (2021)
11. Li, Z., Li, X., Ding, C., Xu, X.: Disentangled pre-training for human-object interaction detection. In: Proc. IEEE CVPR. pp. 28191–28201 (2024). https://doi.org/10.1109/cvpr52733.2024.02663
12. Liao, Y., Liu, S., Wang, F., Chen, Y., Qian, C., Feng, J.: PPDM: Parallel point detection and matching for real-time human-object interaction detection. In: Proc. IEEE CVPR. pp. 482–490 (2020)
13. Liao, Y., Zhang, A., Lu, M., Wang, Y., Li, C., Liu, S.: Pvic: Predicate visual context for human-object interaction detection. In: Proc. IEEE ICCV (2023)

14. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proc. IEEE ICCV. pp. 2980–2988 (2017)
15. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Proc. ECCV, pp. 740–755. Springer (2014)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proc. IEEE ICCV. pp. 10012–10022 (2021)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proc. ICLR (2019)
18. Lv, W., Zhao, Y., Chang, Q., Huang, K., Wang, G., Liu, Y.: RT-DETRv2: Improved baseline with bag-of-freebies for real-time detection transformer. arXiv preprint arXiv:2407.17140 (2024)
19. Ning, S., Qiu, L., Liu, Y., He, X.: HOICLIP: Efficient knowledge transfer for HOI detection with vision-language models. In: Proc. IEEE CVPR. pp. 23507–23517 (2023)
20. Reddy, K.G., Basha, S.S.: Real time object identification: A study on COCO dataset. In: Proc. ICAMMSD. pp. 860–870. Atlantis Press (2025)
21. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proc. IEEE ICCV. pp. 8430–8439 (2019)
22. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proc. IEEE CVPR. pp. 761–769 (2016)
23. Tamura, M., Ohashi, H., Yoshinaga, T.: QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information. In: Proc. IEEE CVPR. pp. 10410–10419 (2021)
24. Yang, J., Li, B., Yang, F., Zeng, A., Zhang, L., Zhang, R.: Boosting human-object interaction detection with text-to-image diffusion model. In: Proc. IEEE CVPR. pp. 15309–15318 (2023)
25. Yuan, H., Zhang, S., Wang, X., Albanie, S., Pan, Y., Feng, T., Jiang, J., Ni, D., Zhang, Y., Zhao, D.: RLIP-v2: Fast scaling of relational language-image pre-training. In: Proc. IEEE ICCV. pp. 21649–21661 (2023)
26. Zhang, F.Z., Campbell, D., Gould, S.: Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In: Proc. IEEE CVPR. pp. 20104–20112 (2022)
27. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In: Proc. ICLR (2023)
28. Zhang, Y., Pan, Y., Yao, T., Huang, R., Mei, T., Chen, C.W.: Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In: Proc. IEEE CVPR. pp. 19548–19557 (2022)
29. Zhong, X., Qu, X., Ding, C., Tao, D.: Glance and Gaze: Inferring action-aware points for one-stage human-object interaction detection. In: Proc. IEEE CVPR. pp. 15163–15172 (2021)
30. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection. In: Proc. ICLR (2021)
31. Zou, C., Wang, B., Hu, Y., Liu, J., Wu, Q., Zhao, Y., Li, B., et al.: End-to-end human object interaction detection with HOI transformer. In: Proc. IEEE CVPR. pp. 11825–11834 (2021)