# B-RIGHT: Benchmark Re-evaluation for Integrity in Generalized Human-Object Interaction Testing

Yoojin Jang*[1]
softjin@unist.ac.kr
Junsu Kim*[1]
jjunsssk@unist.ac.kr
Hayeon Kim[1]
rlagkdus705@unist.ac.kr
Eunki Lee[2]
alexleee@hanyang.ac.kr
Eunsol Kim[2]
eunsolkim@hanyang.ac.kr
Seungryul Baek[1]
srbaek@unist.ac.kr
Jaejun Yoo†[1]
jaejun.yoo@unist.ac.kr

[1] Ulsan National Institute of Science and Technology
Ulsan, South Korea
[2] Hanyang University
Seoul, South Korea

**Abstract**

Human-object interaction (HOI) detection aims to understand complex relationships between humans and objects, but existing benchmarks like HICO-DET suffer from severe class imbalance and inconsistent train-test splits, compromising reliable evaluations. To address these issues, we introduce **B-RIGHT**, a systematically balanced dataset constructed via a novel balancing algorithm and an automated image generation-and-filtering pipeline. By ensuring uniform representation across HOI classes, B-RIGHT significantly reduces performance variance. Our re-evaluation of state-of-the-art methods reveals substantial shifts in model rankings compared to HICO-DET, highlighting previously hidden biases. These results underscore the critical importance of balanced benchmarks for fair and insightful model comparisons. The dataset is publicly available at B-RIGHT.

## 1 Introduction

Human-object interaction (HOI) detection plays a crucial role in enabling artificial intelligence to interpret the visual world, as it captures the complex relationships between humans and objects. Formally, HOI is defined by a triplet (subject, object, interaction), encompassing diverse and subtle actions.

*Equal contributions. †Corresponding author.

For instance, as clearly visualized in Fig. 1a, illustrates a scenario where “a person is sitting on a motorcycle," requiring an HOI detector not only to localize the “person" and the “motorcycle" but also to classify interactions such as “sit on," “ride," or “straddle." Advancing HOI detection can benefit a wide range of computer vision tasks, including image captioning, reasoning [46] and other applications [18, 35].

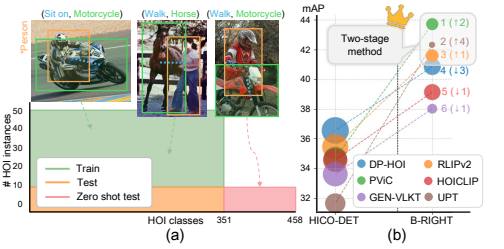


Figure 1: (a) Our balanced distribution of HOI classes in train, test, and zero-shot. (b) Clear ranking shifts between HICO-DET and B-RIGHT, highlighting reduced variance (circle size) in class-wise AP. The arrows indicate shifted ranking.

Recent efforts in HOI detection have heavily depended on Transformer-based [44] architectures [26, 27, 32, 42, 47, 52, 53, 54, 55], leading to notable improvements in accuracy and efficiency. These methods are mainly evaluated on established benchmarks like HICO-DET [5], which offers a widely used standardized framework for model comparisons.

Despite their broad use, there has been limited systematic investigation into the benchmarks themselves, particularly regarding how inherent dataset limitations, such as significant class-imbalance and inconsistent data splits, may skew performance assessments and hinder reliable model comparisons.

In particular, HICO-DET exhibits pronounced class-imbalance across its train and test splits, with some classes having disproportionately fewer or greater instances, leading to unreliable and potentially misleading performance metrics. Addressing these imbalances by uniformly expanding the dataset for each HOI class is impractical, given that certain interactions (e.g.,(dry, cat) or (tag, person) in Fig. 2a) are inherently rare or context-dependent, posing challenges for collection and annotation due to scarcity and copyright constraints.

To overcome these challenges, we introduce **B**enchmark **R**e-evaluation for **I**ntegrity in **G**eneralized **H**uman-object Interaction **T**esting (**B-RIGHT**). Rather than pursuing perfect balance across all HOI classes, B-RIGHT strategically complements existing benchmarks by selecting a representative subset of classes that can feasibly achieve balance. We then augment these classes using a rigorous pipeline involving high-quality data generation and web-crawled images, thoroughly filtered through automated processes leveraging large language models (LLMs) and vision-language models (VLMs), followed by meticulous manual verification. This systematic combination of automated and manual procedures ensures uniform class representation with high-quality, copyright-free images, thus establishing a reliable benchmark that uncovers inherent characteristics previously hidden by data imbalance.

***Importantly, our goal is not to replace existing datasets but to complement them with a class-balanced, bias-reduced benchmark, paving the way for deeper analysis of architectural and training factors under controlled conditions.*** Specifically, B-RIGHT aims to validate whether model improvements and architectural innovations translate effectively under balanced conditions, highlighting genuine performance gains beyond dataset-induced biases. Additionally, we introduce a balanced zero-shot test set to systematically assess model generalization to unseen HOI interactions.

When re-evaluating existing models with B-RIGHT, Fig. 1b demonstrates a clear reduction in class-wise performance variance and substantial ranking shifts compared to HICO-DET. These findings confirm that our balanced conditions effectively mitigate previously inflated or deflated class scores caused by data imbalance, thus providing a reliable foundation for detailed analyses of model architectures and training strategies.

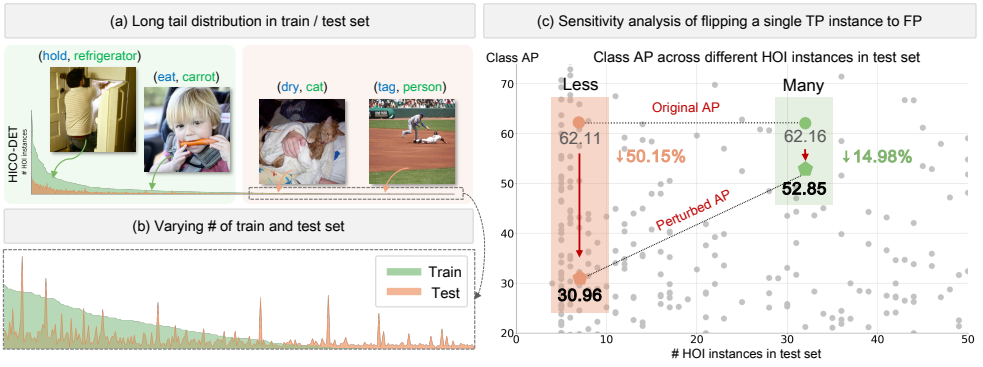


Figure 2: **Analysis of dataset issues in HICO-DET.** Left: (a) Severe long-tail distribution of HOI classes (examples: `(hold, refrigerator)`, `(eat, carrot)`, `(dry, cat)`, `(tag, person)`); (b) inconsistent distribution between train and test splits. Right: (c) Sensitivity analysis showing the impact of flipping a single true positive (TP) instance to false positive (FP). The class with fewer test instances (Less) experiences a drastic AP drop compared to the class with more instances (Many). Original AP scores (circles, ●) and perturbed AP scores (pentagon, ⬟) are shown with percentage decreases.

Our contributions are as follows: (1) We conduct a detailed analysis of previously overlooked dataset limitations in existing HOI benchmarks, emphasizing issues arising from long-tail distributions and class-imbalance. (2) We propose **B-RIGHT**, the first systematically balanced HOI benchmark, constructed through a rigorous pipeline that integrates novel balancing algorithms, automated image generation-and-filtering mechanisms, and meticulous human verification to guarantee dataset reliability. (3) We re-trained and systematically re-evaluated state-of-the-art HOI models using **B-RIGHT**, offering comprehensive insights into their structural and generalization performance under balanced data conditions.

## 2 Related works

**HOI benchmarks.** Various benchmarks for HOI detection have emerged over the years, including V-COCO [14] and HICO-DET [5]. While both datasets derive from the MS-COCO dataset, HICO-DET provides more HOI-specific data, involving a broader range of interaction classes and a larger number of images compared to V-COCO. Building upon HICO-DET's broad applicability, recent studies have addressed advanced evaluation scenarios including zero-shot learning [34, 52]. However, most of these scenarios are being managed on imbalanced datasets like HICO-DET; thus we take the first step toward introducing and evaluating a balanced benchmark, aiming to enable more reliable model evaluations.

**HOI methods.** HOI detection typically comprises two main subtasks: detection and interaction classification. Accordingly, the HOI detectors are categorized into the *one-stage* and *two-stage* architectures. One-stage models integrate detection and interaction classification into a single pipeline, whereas two-stage models first detect objects and then classify the interactions. Recently, most methods incorporate query-based Transformer [44] architectures into their schemes, and adopt the DETR family [2, 31] as a baseline [42]. Further refinements have explored specialized query designs [47, 53] to better capture interaction classes. Efforts to improve interaction understanding for rare classes also include the utilization of foundation models such as CLIP [27, 34], and external knowledge from large-scale datasets [26, 52].

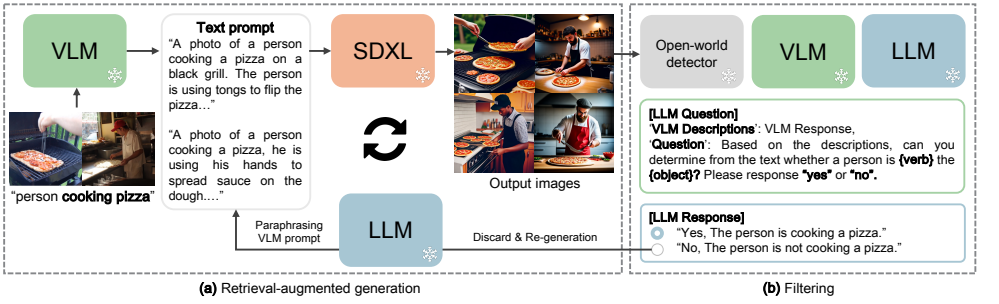


Figure 3: **Generation–filtering pipeline for collecting class-balanced HOI images.** (a) *Retrieval-augmented generation*: A real image from HICO-DET is fed to a frozen VLM to obtain a descriptive prompt; an LLM paraphrases the prompt, and SDXL synthesizes candidate images. (b) *Filtering*: A frozen open-world detector localizes people and objects, after which a VLM + LLM pair answers a yes/no query to verify whether the target HOI is present. Images that fail verification are discarded and regenerated with paraphrased prompts until the class quota is met or the generation budget is reached.

# 3 Motivation

**Potential issues of HOI dataset.** HICO-DET has gained popularity due to its diverse classes and scenes, but it has critical limitations that compromise reliable evaluation. As illustrated in Fig. 2a, HICO-DET suffers from a severe long-tail distribution, with instance counts per class ranging dramatically from over 4,000 down to just 1, far exceeding the imbalance observed in standard object detection datasets [10, 29]. Furthermore, Fig. 2b shows the significant variation in the ratio of train to test instances for each class, which is not limited to rare classes but also appears among non-rare ones, leading to unfair class-wise evaluation. As a result, many HOI classes remain substantially underrepresented, skewing model training toward dominant classes and undermining fair comparisons across different approaches.

**Impact of long-tail data on model prediction.** Class-imbalanced train set is known to bias a model toward frequent classes [45], thereby degrading HOI performance [20, 57]. When the test set is also imbalanced, the problem intensifies: classes with few test instances become disproportionately sensitive to each prediction. Specifically, having fewer test instances in certain classes amplifies the impact of each prediction on overall performance metrics. To quantify this impact, we follow PViC [55] and measure the change in average precision (AP) when a single true-positive (TP) is converted to a false-positive (FP) heuristically. As depicted in Fig. 2c, we compare two classes with comparable training counts and similar original AP: `Many` (more test instances) and `Less` (fewer). After aligning confidence scores, replacing the highest-confidence TP with an FP reduces AP far more in `Less` than in `Many`, revealing the sensitivity of mean AP (mAP) to test-set imbalance. Hence, evaluations on imbalanced test sets can inflate or deflate the reported performance, leading to high variance and unreliable metrics [43]. A class-balanced test set is therefore essential for faithfully assessing generalization across diverse HOI instances.

# 4 B-RIGHT: Balanced HOI dataset construction

**Constructing balanced datasets with $K$ classes.** We construct a unified dataset by combining train and test sets to alleviate the skewed representation of the initial train and test sets.

Then, we choose the value of $K$ classes to (1) cover a diverse objects and verbs, (2) focus on classes with instance counts near the overall median, and (3) approximate 85:15 ratio train-test split, with the test set containing only original (i.e., non-synthetic) instances. We sample images for these classes, and provide details of the sampling and the $K$ selection process in C. Since the original data follows a severe long-tail distribution, there may still be classes in the train set with only a few instances even after image sampling. Additional data collection is necessary for these underrepresented classes. Within the defined $K$ range, we leverage a generative model to produce sufficient data through advanced prompt construction.

### 4.1 Retrieval-augmented generation

We adopt a retrieval-augmented approach for constructing prompts to generate high-fidelity images. Given a HOI class $\mathbf{c}$, we retrieve a real image $x$ from a subset $\mathcal{X}_\mathbf{c}$ of HICO-DET, which contains images annotated with the HOI class $\mathbf{c}$. We then leverage a vision-language model (VLM) [9] to produce context-rich textual descriptions of the reference image. To precisely control the VLM's interpretation of the image, we employ prompt-tuning [25], with detailed prompt examples provided in Appendix Sec. B.3. The resulting descriptive text $p_{\text{VLM}}$ is subsequently used as input for the text-to-image generation model [28]. After constructing the prompt $p_{\text{VLM}}$, we feed it to a text-to-image diffusion model, SDXL-lightning [28] to generate a high-fidelity image $x'$.

**Filtering process.** Despite careful prompt construction, generated images may still inaccurately represent target HOIs. To ensure accuracy, we introduce a multi-step filtering strategy utilizing both VLM [50] and LLM [19]. Initially, an open-world detector [32] identifies all persons and objects in each generated image. Subsequently, a region-based VLM [50] provides localized descriptions for each potential human-object pair, and these descriptions are verified by an open-source LLM [19] using carefully designed query (*e.g.*, "Based on the description, can you check if the person is {verb} the {object}? Answer 'Yes' or 'No'"). Images *confirmed* by the LLM as accurately depicting the intended HOI are retained, while *rejected* images are either discarded or regenerated with paraphrased prompts. This loop continues until we obtain the target quota of valid images or exhaust the generation budget. If the quota is still unmet, we perform a minimal copyright-free web crawl and run the retrieved images through the same VLM–LLM filter. The overall pipeline is illustrated in Fig. 3, and Fig. 4 shows representative examples of both successfully generated and crawled images that passed all filters. Additional details on data crawling and automated pseudo-labeling are provided in Appendix Sec. D.


Figure 4: Example of the generated and crawled images with `(verb,object)` pairs.

## 5 Experiments

### 5.1 Experimental settings

**Dataset.** We construct B-RIGHT, a balanced subset consisting of 351 HOI classes, each with exactly 50 train and 10 test instances, as described in Sec. 4. Further details on the construction process are provided in Appendix Sec. C. Additionally, we construct a balanced

| Dataset | Split | Balanced | #Images | #Instances / #Max / #Min | #Classes | #Object | #Verb |
|---|---|---|---|---|---|---|---|
| HICO-DET | Train | ✘ | 38,118 | 117,871 / 4,051 / 1 | 600 | 80 | 117 |
| HICO-DET | Test | ✘ | 9,658 | 33,405 / 898 / 2 | 600 | 80 | 117 |
| B-RIGHT | Train | ✔ | 6,792 | 17,550 / 50 / 50 | 351 | 78 | 87 |
| B-RIGHT | Test | ✔ | 1,605 | 3,510 / 10 / 10 | 351 | 78 | 87 |
| B-RIGHT | ZS-Test | ✔ | 750 | 1,070 / 10 / 10 | 107 | 58 | 40 |

Table 1: **Statistics of HICO-DET and B-RIGHT.** "#Instances," "#Max," and "#Min" represent the total number of instances across all classes, the maximum number of instances within a single class, and the minimum number of instances within a single class, respectively. "Balanced" indicates whether the dataset split is balanced (✔ = Yes, ✘ = No). "ZS-Test" refers to zero-shot test set. Remaining splits are provided in Appendix Sec. B.6.

zero-shot test set by creating new HOI classes based on novel combinations of objects and verbs that were seen during training but not combined in the 351 selected classes. To reduce the impact of randomness in training, we evaluated the model over four distinct data splits and reported the mean performance across them. Tab. 1 compares core statistics between HICO-DET and B-RIGHT, illustrating how B-RIGHT eliminates the extreme class-imbalances.

**HOI Detectors.** We evaluate several state-of-the-art Transformer-based HOI detectors. We employ a set of one-stage models, including DP-HOI [26], HOICLIP [34], GEN-VLKT [27], CDN [53], CQL [47], and QPIC [42]. Most one-stage models rely primarily on COCO for pre-training; However, DP-HOI utilizes additional datasets such as HAA500 [7], Kinetics700 [3], Flickr30K [51], Visual Genome (VG)[24], and Object365 (O365)[40]. For two-stage models, we consider RLIPv2 [52], PViC [55], and UPT [54], which employ large-scale datasets such as VG, COCO, O365, and HICO-DET during pre-training. Most evaluated methods adopt the DETR [2] architecture; RLIPv2 specifically utilizes the DAB-Deformable DETR (DAB-DDETR) [31]. Detailed model configurations are summarized in Appendix Sec. E. This comprehensive evaluation covers diverse training strategies and architectures within the balanced setting of B-RIGHT.

**Evaluation metrics.** We follow the standard practice of using mean average precision (mAP) to measure HOI detection accuracy. Our evaluation framework consists of two distinct scenarios. In the default scenario, models for HICO-DET are trained and evaluated on all classes, reflecting the dataset's original design. In contrast, for B-RIGHT, models are trained and evaluated on 351 classes, which aligns with the dataset's focus on balanced evaluation. Furthermore, we evaluate zero-shot performance scenario using two different splits: HICO-DET's original rare first unseen composition (RF-UC) [5], which exhibits class-imbalance, and our newly proposed B-RIGHT zero-shot partition.

## 5.2 Comprehensive Analysis

Here, we analyze how state-of-the-art HOI detectors perform on our balanced B-RIGHT compared to HICO-DET, which is imbalanced. We also examine zero-shot performance across both benchmarks.

**Variance in class-wise AP.** In Sec. 3, we highlight that imbalanced test sets can lead to high variance in class-wise performance and undermine the reliability of evaluation metrics. To address this issue, we demonstrate that balanced datasets support more consistent model performance across classes, thereby reducing variance and improving metric reliability. Fig. 5

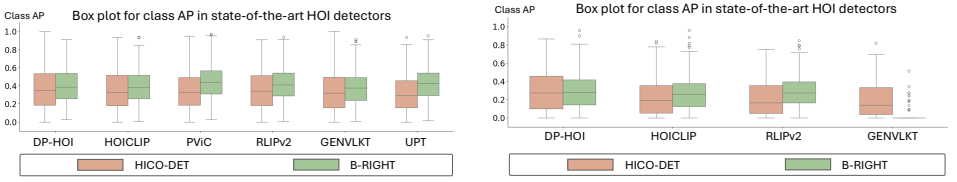


Figure 5: Box plots of class AP for HOI detectors trained/evaluated on HICO-DET and B-RIGHT. (Left: default / Right: zero-shot setting)

| | | Default | | | | Zero-shot (RF-UC) | |
|---|---|---|---|---|---|---|---|
| | | HICO-DET | | B-RIGHT | | HICO-DET | B-RIGHT |
| Model | Architecture | mAP | Rank | mAP | Rank | mAP | mAP |
| DP-HOI [26] | one stage | 36.56 | 1 | 40.85 | 4 (↓ 3) | 30.49 | 31.81 |
| RLIPv2 [62] | two stage | 35.46 | 2 | 41.61 | 3 (↓ 1) | 21.45 | 29.97 |
| PViC [53] | two stage | 34.69 | 3 | 43.73 | 1 (↑ 2) | – | 37.07 |
| HOICLIP [34] | one stage | 34.56 | 4 | 39.14 | 5 (↓ 1) | 25.53 | 28.36 |
| GEN-VLKT [27] | one stage | 33.61 | 5 | 38.02 | 6 (↓ 1) | 21.36 | 4.38 |
| UPT [64] | two stage | 31.65 | 6 | 42.34 | 2 (↑ 4) | – | 35.95 |
| CQL [47] | one stage | 31.58 | 7 | 33.59 | 9 (↓ 2) | – | 24.68 |
| CDN [53] | one stage | 31.36 | 8 | 35.24 | 7 (↑ 1) | – | 27.05 |
| QPIC [42] | one stage | 29.11 | 9 | 34.00 | 8 (↑ 1) | – | 24.05 |

Table 2: **Comparison of rankings of HOI detectors on HICO-DET and B-RIGHT.** 'Rank' indicates the ranking of models based on their mAP scores, while arrows and numbers denote ranking shifts. The ↑ indicates an improvement in ranking compared to HICO-DET, while the ↓ indicates a decrease in ranking. The mAP scores for HICO-DET are referenced from previously reported results in their paper. The symbol (–) indicates models for which official implementation is unavailable.

visualizes the class AP statistics for various models on both HICO-DET and B-RIGHT. On HICO-DET, all evaluated models exhibit substantial class-wise AP variance, indicating inconsistent performance across classes. In contrast, B-RIGHT reduces variance and enhances overall AP, enabling more uniform and reliable performance evaluations. Similar trends are consistently observed under the zero-shot scenario.We acknowledge that some performance improvements may result partially from the reduced number of samples per class. However, the observed decrease in class-wise AP variance across all evaluated models confirms that our proposed benchmark significantly reduces evaluation bias, enabling more reliable and fair comparisons.

**Ranking shifts on B-RIGHT.** Tab. 2 shows that balancing the test set radically reshapes the leaderboard. On HICO-DET, DP-HOI and RLIPv2 originally occupied the top two positions. However, when evaluated on B-RIGHT, DP-HOI drops significantly from 1st to 4th (↓3), suggesting that its reliance on frequent-class priors undermines performance in balanced scenarios. RLIPv2, despite its robust object-proposal stage, also loses one rank, shifting from 2nd to 3rd (↓1). Conversely, two-stage methods, such as PViC and UPT, notably rise to 1st (↑2) and 2nd (↑4) positions respectively, indicating that their verb–object disentanglement strategies were previously underappreciated due to the influence of class-imbalance noise.

Among one-stage methods, HOICLIP and GEN-VLKT both drop by one rank (HOICLIP from 4th to 5th, GEN-VLKT from 5th to 6th, each ↓1) despite increased absolute mAP

scores, demonstrating that improvements from removing class-imbalance are broadly shared across one-stage architectures, with limited relative gains. Notably, CQL drops sharply from 7th to 9th (↓2), as its bias toward frequent classes no longer compensates for weaknesses on rare classes. Meanwhile, CDN and QPIC each experience minor rank improvements of one position (↑1), moving from 8th to 7th and 9th to 8th respectively.

Zero-shot evaluation columns (RF-UC) reinforce this trend: detectors exhibiting substantial rank improvements in the balanced default scenario (*e.g.*, PViC and UPT) also achieve substantial gains without supervision, underscoring that balanced evaluation is essential for accurately assessing both generalization and absolute accuracy. Overall, B-RIGHT clearly exposes architecture-specific strengths previously obscured by HICO-DET, transforming seemingly incremental differences into clear, meaningful distinctions among methods.

**Impact of architectural choices.** Ranking shifts on B-RIGHT (Tab. 2) clearly expose architectural contrasts. Two-stage detectors such as PViC and UPT climb sharply to the top ranks (↑2, ↑4), implying that explicitly separating object detection from interaction classification yields more balanced verb–object representations. RLIPv2, although a two-stage detector, falls by only one rank (↓1). Unlike PViC and UPT, whose gains stem from stronger visual pairwise modelling, RLIPv2 already exploits large-scale language-image pre-training, so its head-class bias is weaker to begin with and the relative advantage on a balanced test set is smaller. In contrast, integrated one-stage methods like DP-HOI, HOICLIP, GEN-VLKT and CQL all lose the original ranking, confirming their greater susceptibility to overfitting frequent interactions and their difficulty in generalising when each class is equally represented.

**Impact of pre-training datasets.** DP-HOI and RLIPv2 are pre-trained on extensive pre-training datasets, as detailed in the default setting of Appendix Sec. E. Despite this, both models experience significant performance drops on B-RIGHT compared to HICO-DET: DP-HOI falls from first to fourth rank, and RLIPv2 drops from second to third rank, as shown in Tab. 2. These results indicate that diverse pre-training alone does not ensure improved performance in balanced scenarios. Instead, extensive pre-training may amplify biases toward frequent classes, with balanced benchmarks like B-RIGHT exposing generalization gaps in rare categories. This highlights the importance of aligning pre-training strategies on balanced setting.

**Role of foundation models.** As shown in Tab. 2, vision-language foundation models like CLIP [37], which are integrated into networks such as HOICLIP and GEN-VLKT, align image and text representations to improve semantic understanding. However, these models show consistent ranking decreases on B-RIGHT, indicating that CLIP cannot resolve biases from imbalanced train data.

**Backbone architecture and feature representation.** While many HOI detectors adopt a basic DETR backbone, RLIPv2 stands out by integrating a more advanced variant, DAB-DDETR, which typically converges faster and yields richer spatial features. However, despite this sophisticated backbone and its two-stage design, RLIPv2 underperforms compared to other two-stage methods such as PViC and UPT that use standard DETR backbone on B-RIGHT (see Tab. 2). This suggests that a more powerful backbone alone may not be sufficient under balanced conditions. For backbone scale analysis, please refer to Appendix Sec. A.

**Comparison to models in zero-shot balanced test set.** In Tab. 2, two-stage architectures (*e.g.*, PViC and UPT) lead the rankings with 37.07 and 35.95 mAP respectively, demonstrating the advantages of separating detection from interaction classification. Their explicit two-stage pipeline proves effective at generalizing to unseen verb–object combinations, aligning

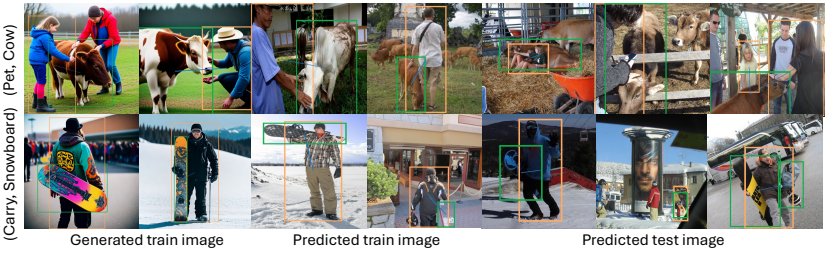


Figure 6: Visualization of HOICLIP [34] predictions on B-RIGHT.

with our findings in the *default setting*. Furthermore, DP-HOI and RLIPv2, which utilize specialized zero-shot or large-scale pre-training strategies, achieve 31.81 and 29.97 mAP, respectively. DP-HOI achieves a strong performance of 31.81 mAP, maintaining its competitiveness despite shifting from its previous top ranking in the RF-UC setting. Likewise, RLIPv2 demonstrates resilient performance (29.97 mAP) when transitioning from the RF-UC to the balanced setting, suggesting that its pre-training pipeline is particularly effective at handling balanced zero-shot scenarios.

Notably, HOICLIP and GEN-VLKT, despite both leveraging CLIP-based knowledge for enhanced zero-shot capabilities, achieve lower mAPs of 28.36 and 4.38, respectively. This indicates that while CLIP frameworks are effective in certain zero-shot splits (like the RF-UC setting), they may not consistently perform well across all zero-shot setting, particularly with balanced unseen class distributions. For a more analysis of HOICLIP and GEN-VLKT, please refer to Appendix Sec. A. These findings highlight how zero-shot performance depends on multiple factors, including architectural choices, specific pre-training approaches, and auxiliary network such as CLIP.

## 6 Discussion and Limitations

Here, we analyze the potential bias introduced by training with synthetic data and its impact on evaluating real HOI images, followed by a discussion of our method's limitations.

**Verification of utilizing synthetic data in train set.** Fig. 6 presents the prediction results of the HOI detector [34] trained on B-RIGHT and evaluated on real test images. The results indicate that the model performs well on real data, demonstrating no noticeable bias introduced by the use of synthetic images. Furthermore, we compare the performance of HOI detectors trained with and without synthetic data in B-RIGHT in Fig. 7. We observed that the inclusion of synthetic images leads to consistent performance improvements. These results align with prior works utilizing synthetic data for training [17, 48, 49], providing clear justification for incorporating synthetic images into our train dataset to enhance generalization in HOI detection.

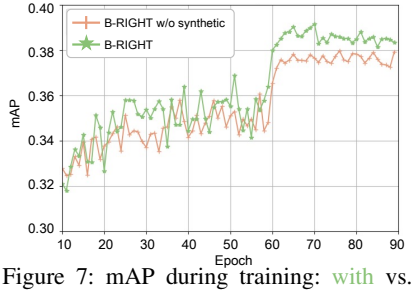


Figure 7: mAP during training: with vs. without synthetic data in B-RIGHT.

**Limitations.** Although B-RIGHT alleviates extreme class-imbalance, each HOI category is limited to a fixed number of instances, reducing overall coverage compared to larger-scale datasets. In addition, we rely partly on synthetic images and web-crawled samples, which

may not capture every nuanced real-world interaction. Nevertheless, this controlled balancing process offers a uniquely unbiased dataset for fair model comparisons, and we believe these limitations can be addressed by future expansions with more diverse data sources and improved generation techniques.

# 7 Conclusion

We addressed the challenge of imbalanced data in HOI detection benchmarks by introducing B-RIGHT, a new dataset that guarantees an equal number of instances for each interaction class. This balanced design exposes latent class biases overlooked in previous long-tailed benchmarks, enabling more transparent analysis of architectural innovations, training strategies, and foundational model. Our experiments show that B-RIGHT significantly reduces class-wise variance, revealing new insights such as the consistent advantages of two-stage methods under balanced conditions. By offering a more fair evaluation framework, we hope B-RIGHT will stimulate further progress in HOI detection, allowing researchers to better pinpoint areas for improvement and advance visual reasoning in more robust, equitable ways.

# 8 Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-II220264, Comprehensive Video Understanding and Generation with Knowledge-based Deep Logic Neural Network, No.RS-2020-II201336, Artificial Intelligence graduate school support (UNIST)), National Supercomputing Center with supercomputing resources including technical support (KSC-2025-CRE-0094), Korea Bio Data Station (K-BDS) with computing resources including technical support, and Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT (MSIT, Korea) & Gwangju Metropolitan City.

# References

[1] 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. *NIPS*, 2024.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[3] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv*, 2019.

[4] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. *CVPR*, 2024.

[5] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018.

[6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

[7] Jihoon Chung, Cheng-hsin Wuu, Hsuan-ru Yang, Yu-Wing Tai, and Chi-Keung Tang. Haa500: Human-centric atomic action dataset with curated videos. In *ICCV*, 2021.

[8] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. In *arXiv*, 2024.

[9] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv*, 2024.

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[11] Hugging Face. stable-diffusion-v1-5, 2022. URL https://huggingface.co/runwayml/stable-diffusion-v1-5. Accessed: 2024-05-19.

[12] Shuman Fang, Shuai Liu, Jie Li, Guannan Jiang, Xianming Lin, and Rongrong Ji. Improving human-object interaction detection via virtual image learning. In *ACMMM*, 2023.

[13] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv*, 2023.

[14] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv*, 2015.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NIPS*, 2020.

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv*, 2021.

[18] Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. Knowing where to focus: Event-aware transformer for video grounding. In *ICCV*, 2023.

[19] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv*, 2023.

[20] Mert Kilickaya and Arnold Smeulders. Diagnosing rarity in human-object interaction detection. In *CVPRW*, 2020.

[21] Junsu Kim, Hoseong Cho, Jihyeon Kim, Yihalem Yimolal Tiruneh, and Seungryul Baek. Sddgr: Stable diffusion-based deep generative replay for class incremental object detection. *CVPR*, 2024.

[22] Junsu Kim, Yunhoe Ku, Jihyeon Kim, Junuk Cha, and Seungryul Baek. Vlm-pl: Advanced pseudo labeling approach for class incremental object detection via vision-language model. *CVPRW*, 2024.

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv*, 2013.

[24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[25] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv*, 2021.

[26] Zhuolong Li, Xingao Li, Changxing Ding, and Xiangmin Xu. Disentangled pre-training for human-object interaction detection. In *CVPR*, 2024.

[27] 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 *CVPR*, 2022.

[28] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv*, 2024.

[29] 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 *ECCV*. 2014.

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NIPS*, 2024.

[31] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *ICLR*, 2022.

[32] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv*, 2023.

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

[34] Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In *CVPR*, 2023.

[35] Bo Pang, Kaiwen Zha, Yifan Zhang, and Cewu Lu. Further understanding videos through adverbs: A new video task. In *AAAI*, 2020.

[36] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2024.

[37] 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 *ICML*, 2021.

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[40] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *CVPR*, 2019.

[41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv*, 2020.

[42] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021.

[43] Boris van Breugel, Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. Can you rely on your model evaluation? improving model evaluation with synthetic test data. *NIPS*, 2024.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NIPS*, 2017.

[45] Guangzhi Wang, Yangyang Guo, Yongkang Wong, and Mohan Kankanhalli. Chairs can be stood on: Overcoming object bias in human-object interaction detection. In *ECCV*, 2022.

[46] Yanhui Wang, Ning Xu, An-An Liu, Wenhui Li, and Yongdong Zhang. High-order interaction learning for image captioning. *TCSVT*.

[47] Chi Xie, Fangao Zeng, Yue Hu, Shuang Liang, and Yichen Wei. Category query learning for human-object interaction classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15275–15284, 2023.

[48] Jie Yang, Bingliang Li, Fengyu Yang, Ailing Zeng, Lei Zhang, and Ruimao Zhang. Boosting human-object interaction detection with text-to-image diffusion model. *arXiv*, 2023.

[49] Jie Yang, Bingliang Li, Ailing Zeng, Lei Zhang, and Ruimao Zhang. Open-world human-object interaction detection via multi-modal prompts. In *CVPR*, 2024.

[50] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *ICLR*, 2024.

[51] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014.

[52] H. Yuan, S. Zhang, X. Wang, S. Albanie, Y. Pan, T. Feng, J. Jiang, D. Ni, Y. Zhang, and D. Zhao. Rlipv2: Fast scaling of relational language-image pre-training. In *ICCV*, 2023.

[53] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. *arXiv*, 2021.

[54] Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *CVPR*, 2022.

[55] Frederic Z. Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring predicate visual context in detecting human–object interactions. In *ICCV*, 2023.

[56] Sipeng Zheng, Boshen Xu, and Qin Jin. Open-category human-object interaction pretraining via language modeling framework. In *CVPR*, 2023.

[57] Fangrui Zhu, Yiming Xie, Weidi Xie, and Huaizu Jiang. Diagnosing human-object interaction detectors. *arXiv*, 2023.

[58] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ICLR*, 2020.