HICO-DET-SG and V-COCO-SG: New Data Splits to Evaluate Systematic Generalization in Human-Object Interaction Detection

Kentaro Takemoto
Artificial Intelligence Laboratory
Fujitsu Limited
Kanagawa, Japan
k.takemoto@fujitsu.com

Moyuru Yamada
Artificial Intelligence Laboratory
Fujitsu Limited
Kanagawa, Japan
yamada.moyuru@fujitsu.com

Tomotake Sasaki
Artificial Intelligence Laboratory
Fujitsu Limited
Kanagawa, Japan
tomotake.sasaki@fujitsu.com

Hisanao Akima
Artificial Intelligence Laboratory
Fujitsu Limited
Kanagawa, Japan
akima.hisanao@fujitsu.com

Abstract

Human-Object Interaction (HOI) detection is a task to predict interactions between humans and objects in an image. In real-world scenarios, HOI detection models are required systematic generalization, i.e., generalization to novel combinations of objects and interactions, because it is highly probable that the train data only cover a limited portion of all possible combinations. However, to our knowledge, no open benchmark or existing work evaluates the systematic generalization in HOI detection. To address this issue, we created two new sets of HOI detection data splits named HICO-DET-SG and V-COCO-SG based on HICO-DET and V-COCO datasets. We evaluated representative HOI detection models on the new data splits and observed large degradation in the test performances compared to those on the original datasets. This result shows that systematic generalization is a challenging goal in HOI detection. We hope our new data splits encourage more research toward this goal.

1 Introduction

Human-Object Interaction (HOI) detection has been attracting large interest in computer vision, as it is useful for various applications such as self-driving cars, anomaly detection, analysis of surveillance video, and so on. The task is to detect humans and objects as well as interactions between them and the output is typically represented as <human, **interaction**, object> triplet. After some datasets [1–6] were published, a large number of studies has tackled this problem [7–20].

HOI detection is an advanced computer vision task as it requires a model not only to detect humans and objects but also to predict interactions between them. Moreover, humans can have different interactions with the same object (e.g., **wash** a horse and **walk** a horse) and the same interaction occurs for different objects (e.g., wash a horse and wash a car). In real-world scenarios, it is highly probable that the train data only cover a limited portion of all possible combinations of objects and interactions. Thus, it is important for HOI detection models to generalize to novel combinations of known objects and interactions.

Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

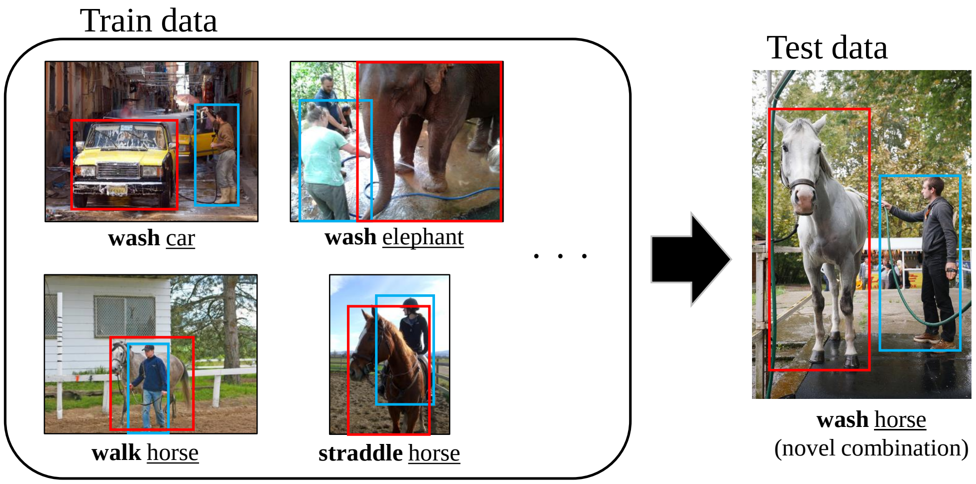


Figure 1: An example of a data split to evaluate systematic generalization in Human-Object Interaction (HOI) detection. The images and annotations are all selected from HICO-DET-SG split3. The train data consists of combinations such as <human, **wash**, car>, <human, **wash**, elephant>, <human, **walk**, horse> and <human, **straddle**, horse>. After trained on such data, the HOI detection model is required to generalize to novel combinations in the test data such as <human, **wash**, horse>. It is important for the model to learn the visual cues of objects (horse) and interactions (**wash**), not depending on the specific paired interaction/object classes in the train data.

This type of generalization to novel combinations of known concepts is called systematic generalization and attracts increasing interest in recent years as it is a highly desirable property for AI systems. Systematic generalization performance is evaluated in various tasks such as sequence-to-sequence parsing [21], language understanding [22, 23], visual properties extraction [24] and visual question answering [25–29]. However, to our knowledge, there is no open benchmark or existing work to evaluate systematic generalization in the HOI detection.

The existing HOI detection datasets cannot evaluate the systematic generalization for the novel combinations of known objects and interactions since their train and test data provide the same object-interaction combinations. This property of the datasets may lead the model to predict the interactions depending just on the paired object classes or to predict object classes by just interactions rather than by capturing the visual cues such as human posture and positional relationships.

In this paper, we introduce two new sets of HOI detection data splits named HICO-DET-SG and V-COCO-SG, which we created based on HICO-DET [4] and V-COCO [3] datasets to evaluate the systematic generalization (SG) capabilities of HOI detection models. An example of such data split to evaluate systematic generalization in the HOI detection is shown in Figure 1.

In order to make sure that the test performance is not an artifact of a specific selection of combinations in the train and test data, we prepared three distinct train-test splits for both HICO-DET-SG and V-COCO-SG. We evaluated recent representative HOI detection models on our data splits and observed large degradation in the test performances compared to those on the original datasets. In summary, our contributions are the following:

- We created two new sets of HOI detection data splits whose train and test data have no overlapping object-interaction combinations, which serve for studying systematic generalization in HOI detection.

- We evaluated the systematic generalization performance of representative HOI detection models with our new data splits and revealed that there are large decreases in the test performance compared to the original datasets, i.e., systematic generalization is a challenging goal in HOI detection.

Table 1: Statistics of HICO-DET-SG and V-COCO-SG as well as the original HICO-DET and V-COCO. SG splits contain less HOI triplets than the original ones because we eliminated triplets in the test data whose combination of classes is contained in the train data.

	# of images		# of HOI triplets		# of object-interaction classes	
Data splits	train	test	train	test	train	test
Original HICO-DET	38,118	9,061	117,871	33,405	600	600
HICO-DET-SG split1	38,312	8,867	119,331	14,994	540	60
HICO-DET-SG split2	39,213	7,966	122,299	7,966	540	60
HICO-DET-SG split3	40,672	6,597	120,096	13,231	540	60
Original V-COCO	5,400	4,946	14,153	12,649	228	228
V-COCO-SG split1	7,297	3,049	8,214	7,575	160	68
V-COCO-SG split2	7,057	3,289	9,500	8,678	160	68
V-COCO-SG split3	6,210	4,136	10,951	9,940	160	68

HICO-DET-SG, V-COCO-SG, and the source code to create them is publicly available at `https://github.com/FujitsuResearch/hoi_sg`.

2 HICO-DET-SG and V-COCO-SG

2.1 Creation of SG splits

The train and test data of the Systematic Generalization (SG) splits are designed not to have overlapping object-interaction combination classes, thus the HOI detection models are required to generalize to novel combinations. For example, in Figure 1, the train data consists of combinations such as <human, **wash**, car>, <human, **wash**, elephant>, <human, **walk**, horse> and <human, **straddle**, horse>. The HOI detection models are required to generalize to novel combinations in the test data such as <human, **wash**, horse>.

We make sure that, in the train data, every object class is paired with multiple interaction classes, and every interaction class is paired with multiple object classes. This design of the splits enables the model to learn the concept of object/interaction, not depending on the specific paired interaction/object class. To ensure that the test performance is not an artifact of a specific selection of combinations in the train and test data, we prepared three distinct train-test splits.

We eliminate triplets in the test data whose combination of classes is contained in the train data. As a result, the SG splits contain less HOIs in total than the original.

See Appendix B for further details of the creation process of SG splits.

2.2 Statistics of HICO-DET-SG and V-COCO-SG

Based on HICO-DET dataset [4] we created HICO-DET-SG, new data splits to evaluate the systematic generalization performance in the way explained above. The statistics of the original HICO-DET and HICO-DET-SG data splits are shown in Table 1 (upper half). The original HICO-DET dataset contains 80 classes of objects, 117 classes of interactions, and 600 classes of object-interaction combinations. In HICO-DET-SG splits, 540 object-interaction classes out of 600 are contained only in the train data, and the test data consists of 60 remaining classes.

Based on V-COCO dataset [3] we created V-COCO-SG, new data splits to evaluate the systematic generalization performance in the same manner. The statistics of the original V-COCO and V-COCO-SG are shown in Table 1 (lower half). The original V-COCO dataset contains 91 classes of objects, 29 classes of interactions, and 228 classes of object-interaction combinations. In the V-COCO-SG splits, 160 object-interaction classes out of 228 are contained only in the train data, and the test data consists of 68 remaining classes.

Table 2: Comparison of systematic generalization performances on HICO-DET-SG data splits. We show mAPs (%) for the test data, which are higher the better. The mAPs reported in the original papers are written in brackets. The term "Pre-training" means that the initial weights of the model's encoder and decoder are copied from DETR [31] trained on object detection task. Test accuracies significantly decreased on HICO-DET-SG compared to the original HICO-DET for all the models.

	HOTR		QPIC		QAHOI	STIP	
Pre-training	✗	✓	✗	✓	✗	✗	✓
Original HICO-DET	17.63	26.30	21.70	29.59	35.30	13.21	31.57
(mAP in the original paper)		(25.73)		(29.90)	(35.78)		(32.22)
HICO-DET-SG split1	0.31	2.59	10.57	22.08	4.53	0.00	22.14
HICO-DET-SG split2	0.28	2.73	12.53	19.95	4.76	0.00	23.03
HICO-DET-SG split3	0.52	2.75	13.28	20.50	6.12	0.00	24.39
Average over SG splits	0.37	2.69	12.13	20.84	5.14	0.00	23.19

Table 3: Comparison of systematic generalization performances on V-COCO-SG data splits. We show mAPs (%) for the test data, which are higher the better. The mAPs reported in the original papers are written in brackets. "Pre-training" means that the initial weights of the model's encoder and decoder are copied from DETR [31] trained on object detection task. Test accuracy significantly decreased on V-COCO-SG data splits compared to the original V-COCO for all the models.

	HOTR		QPIC		QAHOI	STIP	
Pre-training	✗	✓	✗	✓	✗	✗	✓
Original V-COCO	24.26	62.54	27.64	63.41	40.74	18.43	70.43
(mAP in the original paper)		(63.8)		(61.0)			(70.65)
V-COCO-SG split1	0.24	2.10	0.77	4.21	2.73	0.12	6.25
V-COCO-SG split2	0.27	2.60	0.66	4.16	1.80	0.00	6.22
V-COCO-SG split3	0.15	1.68	0.32	2.89	1.53	0.00	6.37
Average over SG splits	0.22	2.13	0.58	3.75	2.02	0.04	6.28

3 Evaluation results

In this section, we report systematic generalization performance of four representative HOI detection models, HOTR [14], QPIC [16], QAHOI [15] and STIP [20], on HICO-DET-SG and V-COCO-SG. We trained and tested each model once on each split. See Appendix C for the details of these HOI detection models and other experimental setup. Further results and discussions are available in Appendix D.

3.1 Degradation in the systematic generalization performance

We show the results on the HICO-DET-SG in Table 2 and the results on the V-COCO-SG in Table 3. As the evaluation metrics, we use mean average precision (mAP), which is higher the better. In order to evaluate all the models on the equal condition in terms of pre-training, we report the results of HOTR, QPIC and STIP for both pre-trained (with object detection task on MS COCO dataset [30]) and not pre-trained encoder and decoder.

The mAPs of all models significantly decrease on all systematic generalization splits compared to the ones on the original splits. This shows the difficulty of systematic generalization in HOI detection task, i.e., recognizing novel combinations of known objects and interactions. Note that this degradation occurs at any selection of object-interaction combinations in the train and test data: the differences in the test mAPs are less than 3% among three systematic generalization splits for all the models and datasets.

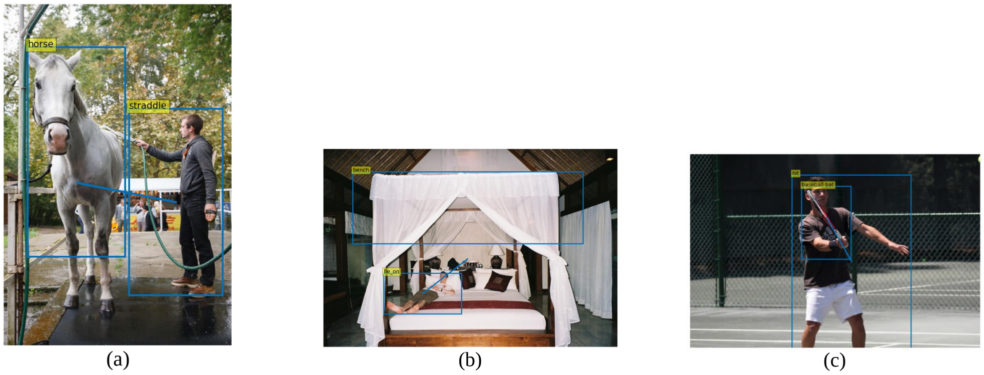
3.2 Qualitative analysis

In order to further reveal the difficulty of systematic generalization in HOI detection, we analyze the failure cases here. Figure 2 (a), (b) and (c) show the outputs of pre-trained STIP trained and tested on HICO-DET-SG split3, which has the highest mAP among all models and all SG splits.

Figure 2 (a) shows an example of predicting wrong interaction class, which is the most frequently observed type of errors. In this example the model predicts the interaction as **straddle**, even though the correct class is **wash**. The <human, **straddle**, horse> triplet appears in the train data but the <human, **wash**, horse> triplet appears only in the test data (**wash** interaction appears with other objects in the train data). The model predicts the interaction depending just on the object class (horse), i.e., the model cannot generalize to the novel combination <human, **wash**, horse>.

Figure 2 (b) shows an example of detecting wrong object. The model predicts an irrelevant region as a wrong class, bench, even though it should detect a bed under the person in the image. The <human, **lie on**, bench> triplet appears in the train data but the <human, **lie on**, bed> triplet appears only in the test data (bed appears with other interactions in the train data). This result shows not only that the model cannot generalize to the novel combination but also that the interaction decoder predicts the **lie on** interaction depending mainly on the visual cue of the human posture, not on the visual cue of the object nor on the positional relationships.

Figure 2 (c) shows an example of wrong class prediction for both objects and interaction. The model predicts the tennis racket as a baseball bat and predicts **swing** as **hit**. The <human, **hit**, baseball bat> triplet appears in the train data but the <human, **swing**, tennis racket> triplet appears only in the test data (moreover, <human, **swing**, baseball bat> triplet appears in the train data but <human, **hit**, tennis racket> triplet does not appear in the train data). The model detects the object as a baseball bat at the first stage, and then at the second stage it predicts the interaction as **hit** based on the detection results because baseball bat frequently appeared with the **hit** interaction in the train data.



(a)

(b)

(c)

Figure 2: Three failure cases of pre-trained STIP trained and tested on HICO-DET-SG split3. (a) An example of predicting wrong interaction class. The model predicts the interaction as **straddle**, even though the correct class is **wash**. (b) An example of detecting wrong object. The model predicts an irrelevant region as a wrong class, bench, even though it should detect a bed under the person. (c) An example of wrong class prediction for both objects and interaction. The model predicts <human, **hit**, baseball bat> triplet even though the correct answer is <human, **swing**, tennis racket> triplet.

4 Conclusion

We created new splits of two HOI detection datasets, HICO-DET-SG and V-COCO-SG, which are designed not to contain overlapping combinations of object-interaction classes in the train and test data for evaluating systematic generalization. We observed large degradation in the test performances on our SG splits compared to those on the original datasets for all the representative HOI detection models we evaluated. This shows that systematic generalization is a challenging goal in HOI detection. We hope our data splits encourage more research toward this goal.

References

[1] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from RGB-D videos. *International Journal of Robotics Research*, 32(8):951–970, 2013.

[2] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The Pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision (IJCV)*, 111:98–136, 2014.

[3] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. arXiv preprint, arXiv:1505.04474, 2015. The V-COCO dataset is publicly available at `https://github.com/s-gupta/v-coco` (Accessed on September 29th, 2022).

[4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 381–389, 2018. The HICO-DET dataset is publicly available at `http://www-personal.umich.edu/~ywchao/hico/` (Accessed on September 29th, 2022).

[5] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6047–6056, 2018.

[6] Meng-Jiun Chiou, Chun-Yu Liao, Li-Wei Wang, Roger Zimmermann, and Jiashi Feng. ST-HOI: A spatial-temporal baseline for human-object interaction detection in videos. In *Proceedings of the 2021 Workshop on Intelligent Cross-Data Analysis and Retrieval*, page 9–17, 2021.

[7] Chen Gao, Yuliang Zou, and Jia-Bin Huang. iCAN: Instance-centric attention network for human-object interaction detection. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[8] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3580–3589, 2019.

[9] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. DRG: Dual relation graph for human-object interaction detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[10] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. HOI analysis: Integrating and decomposing human-object interaction. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pages 5011–5022, 2020.

[11] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. GEN-VLKT: Simplify association and enhance interaction understanding for HOI detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20123–20132, 2022.

[12] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. PPDM: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 482–490, 2020.

[13] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[14] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. HOTR: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 74–83, 2021. The official source code of HOTR is available at `https://github.com/kakaobrain/HOTR` (Accessed on September 29th, 2022).

[15] Junwen Chen and Keiji Yanai. QAHOI: Query-based anchors for human-object interaction detection. arXiv preprint, arXiv:2112.08647, 2021. The official source code of QAHOI is available at `https://github.com/cjw2021/QAHOI` (Accessed on September 29th, 2022).

[16] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. The official source code of QPIC is available at `https://github.com/hitachi-rd-cv/qpic` (Accessed on September 29th, 2022).

[17] 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. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

[18] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating HOI detection as adaptive set prediction. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9000–9009, 2021.

[19] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, and Jian Sun. End-to-end human object interaction detection with HOI transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11825–11834, 2021.

[20] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19548–19557, June 2022. The official source code of STIP is available at `https://github.com/zyong812/STIP` (Accessed on September 29th, 2022).

[21] Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

[22] L. Ruis, J. Andreas, M Baroni, D. Bouchacourt, and B. M. Lake. A benchmark for systematic generalization in grounded language understanding. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[23] Leon Bergen, Timothy J. O'Donnell, and Dzmitry Bahdanau. Systematic generalization with edge transformers. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 1390–1402, 2021.

[24] Shimon Ullman, Liav Assif, Alona Strugatski, Ben-Zion Vatashsky, Hila Levi, Aviv Netanyahu, and Adam Uri Yaari. Image interpretation by iterative bottom-up top- down processing. Technical Report CBMM Memo No. 120, Center for Brains, Minds and Machines, 2021.

[25] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910, 2017.

[26] Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. Systematic generalization: What is required and can it be learned? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[27] Dzmitry Bahdanau, Harm de Vries, Timothy J. O'Donnell, Shikhar Murty, Philippe Beaudoin, Yoshua Bengio, and Aaron C. Courville. CLOSURE: assessing systematic generalization of CLEVR models. arXiv preprint, arXiv:1912.05783v2, 2020.

[28] Vanessa D' Amario, Tomotake Sasaki, and Xavier Boix. How modular should neural module networks be for systematic generalization? In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 23374–23385, 2021.

[29] Moyuru Yamada, Vanessa D'Amario, Kentaro Takemoto, Xavier Boix, and Tomotake Sasaki. Transformer module networks for systematic generalization in visual question answering. Technical Report CBMM Memo No. 121, Center for Brains, Minds and Machines, 2022.

[30] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, 2014.

[31] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[32] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. HICO: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

[34] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *Association for Computing Machinery (ACM) Computing Survey*, 2021.

[35] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[36] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[37] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 852–869, 2016.

[38] Xuan Kan, Hejie Cui, and Carl Yang. Zero-shot scene graph relation prediction through commonsense knowledge integration. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases(ECML-PKDD)*, page 466–482, Berlin, Heidelberg, 2021. Springer-Verlag.

[39] Federico Baldassarre, Kevin Smith, Josephine Sullivan, and Hossein Azizpour. Explanation-based weakly-supervised learning of visual relations with graph networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 612–630, 2020.

[40] Zhong Ji, Xiyao Liu, Yanwei Pang, Wangli Ouyang, and Xuelong Li. Few-shot human-object interaction recognition with semantic-guided attentive prototypes network. *IEEE Transactions on Image Processing*, 30:1648–1661, 2021.

[41] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[42] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

Appendix (supplementary materials)

A Related works

In Section A.1, we introduce some types of HOI detection datasets and models. In Section A.2, we explain systematic generalization.

A.1 Human-Object Interaction (HOI) detection

There are two popular datasets for HOI detection: HICO-DET and V-COCO datasets. HICO dataset was originally created for classifying objects and interactions in the image (no bounding-box) [32]. Later HICO-DET dataset was created based on HICO dataset by putting bounding-boxes of humans and objects in the image [4]. Also, at this point, one image in the dataset has become to contain multiple humans, objects and interactions between them. V-COCO dataset [3] was created based on Microsoft COCO (MS COCO) dataset [30] by adding annotations of interactions (verbs). Statistics of HICO-DET and V-COCO dataset are shown in Table 1.

HOI detection consists of two tasks. The first task is to localize human and object instances in the given image. The second task is to identify the interactions between them. Due to this property, there are two types of model architectures to solve HOI detection: two-stage model and one-stage model. Two-stage models [7–11, 20] detect instances at the first stage and classify the interaction for all of their combinations in the second stage. In order to improve both instance and interaction detectors via multi-task learning and to reduce inference time, one-stage models [12, 13] have been proposed recently and becoming more popular these days. In response to the great success of Transformer [33] in both natural language processing and computer vision [34], some of the recent one-stage works [14–19] are based on Transformer as they are designed to capture wide-range information from an image.

A.2 Systematic generalization in HOI detection

Systematic generalization [21–23, 26–29] (also called as compositional generalization [25] and combinatorial generalization [24]) can be regarded as one of the Out-of Distribution (OoD) problem settings, i.e., the model needs to generalize to different data distributions from the train data.

Scene Graph Generation (SGG) [35] is a task closely related to HOI detection. In SGG, there are some works to evaluate [36] and improve [37, 38] systematic generalization for new combinations of subject, relation and object classes under the name of zero-shot generalization. They all showed that there are large performance degradation in the systematic generalization compared to the original settings (in-distribution generalization) unless some techniques are intentionally used. There are two main differences between SGG and HOI detection. First, the subjects in SGG can be of any type (humans, cars, etc.), while in HOI detection they are fixed as humans, which results in more various subject-object combinations in SGG. Second, the predicates in SGG can be both positional relations (e.g., next to) and semantic actions (e.g., play with), while in HOI detection they only consist of the latter. Considering these differences, HOI detection can be regarded as a subset of SGG with less various subject-object combinations. On the other hand, HOI detection can be regarded as a task focusing on measuring a model's capability for complex scene understanding, because recognition of semantic actions needs more information in addition to the locations of humans and objects. Thus we believe that it is essential to work on systematic generalization in HOI detection as an early step towards the development of better models also in SGG and other visual understanding tasks.

In the HOI detection task, HICO-DET dataset [4] provides rare-triplets evaluation to measure the few-shot generalization ability of the models (rare triplets are defined to appear less than 10 times in the train data). Generalization to the rare-triplets is a type of OoD generalization and some works [39, 40] aim at improving this performance. However, to our knowledge, there have been no benchmarks or previous works to tackle systematic generalization, i.e., zero-shot generalization in HOI detection, and our work is the first one providing the data splits to evaluate the systematic generalization performance and benchmarking the representative HOI models.

Algorithm 1 Creation of SG splits. “Dataset” represents the set of images and annotations of the original dataset (combined the train and test data) and “scene” represents a set of one image and HOI triplets in the image. “COUNT” function returns the number of components in the first argument which is equal to the second argument. “test_combinations” is a list of object-interaction classes to be contained only in the test data.

```
train_data = []
test_data = []
for scene in Dataset do
    sum = 0
    test_hois = []
    for hoi in scene.hois do
        match = COUNT(test_combinations, [hoi.object_class, hoi.interaction_class])
        sum = sum + match
        if match > 0 then
            test_hois.append(hoi)
        end if
    end for
    if sum == length(scene.hois) then
        test_data.append(scene)
    else if sum == 0 then
        train_data.append(scene)
    else
        test_data.append([scene.image, test_hois])
    end if
end for
```

B Further details of creating systematic generalization splits

In this section, we explain how we created systematic generalization (SG) splits for HICO-DET and V-COCO.

First, we set the number of object-interaction combination classes to be contained in the train and test data. We tried to match the ratio of them to that of the number of HOI triplets in the original train and test data at the beginning, but eventually more combination classes (540 in HICO-DET-SG and 160 in V-COCO-SG) are contained in the train data in order to ensure that every object class is paired with multiple interaction classes and every interaction class is paired with multiple object classes. This enables the model to learn the concept of object/interaction itself, not depending on the specific paired interaction/object.

Then, we created SG splits as described in Algorithm 1. We eliminated triplets in the test data whose object-interaction combinations are contained in the train data. Because of that, SG splits contain less HOI triplets in total than the original.

Finally, we verified that all object and interaction classes are actually contained in the train data, even though some combinations are only contained in the test data and unseen in the train data for evaluating systematic generalization.

To ensure that the test performance is not an artifact of a specific selection of combinations in the train and test data, we prepared three distinct train-test splits.

The actual source code to create SG splits can be seen at the following repository: `https://github.com/FujitsuResearch/hoi_sg`.

C Experimental setup for evaluating representative HOI detection models

In this section, we describe the experiments to evaluate the systematic generalization ability of representative HOI detection models. In Section C.1, we first explain the evaluated models and the reasons for selecting them. Then we explain experimental setup in Section C.2.

Table 4: Comparison of four HOI detection models, HOTR, QPIC, QAHOI and STIP. We show reported mAPs (%) on the original HICO-DET and V-COCO (higher is better). "mAP on HICO-DET (rare)" shows the performance on rare triplets, i.e., few-shot generalization. The QAHOI paper does not report mAP on V-COCO dataset.

	HOTR [14]	QPIC [16]	QAHOI [15]	STIP [20]
Architecture type	One-stage parallel	One-stage end-to-end	One-stage end-to-end	Two-stage
Feature extractor	CNN	CNN	Multi-scale Transformer	CNN
Base model	DETR [31]	DETR [31]	deformable DETR [41]	DETR [31]
mAP on HICO-DET (full)	25.73	29.90	35.78	32.22
mAP on HICO-DET (rare)	17.34	23.92	29.80	28.15
mAP on V-COCO	63.8	61.0	-	70.65

C.1 HOI detection models

We tested systematic generalization performance for four HOI detection models, HOTR [14], QPIC [16], QAHOI [15] and STIP [20]. The comparison of the four models is shown in Table 4. The selected models except STIP adopt one-stage architecture as it is popular recently. The backbone (feature extraction) network of each model is all pre-trained with object detection task on MS COCO dataset [30]. We used the source code of these models taken from the official publicly-available repositories that we show URLs in the reference section.

HOTR. "Human-Object interaction detection TRansformer" (HOTR) [14] is one of the first Transformer-based models for the HOI detection. The model is one-stage parallel architecture and it has a backbone, shared encoder, instance (human + object) decoder, and interaction decoder. The backbone network is CNN-based to extract features from images. In order to get the matching between instance and interaction decoder outputs, three independent feed-forward networks, which are so-called HO-pointers, are trained to predict correct <human, **interaction**, object> combination matching. Most of the network except HO-pointers is based on "DEtection TRansformer" (DETR) [31], a Transformer-based object detector. Therefore we can pre-train backbone, shared encoder, instance decoder, and interaction decoder with the DETR's weights trained on the MS COCO dataset.

QPIC. "Query-based Pairwise human-object interaction detection with Image-wide contextual information" (QPIC) [16] is another Transformer-based HOI detection model that is proposed around the same time as HOTR. It is also based on DETR for most of the network. The main difference from HOTR is that QPIC is a one-stage end-to-end architecture that has only one decoder without HO-pointers. We can use pre-trained weights of DETR on MS COCO datasets for most of the network, including backbone, encoder, and decoder.

QAHOI. "Query-based Anchors for Human-Object Interaction detection" (QAHOI) [15] exhibits the best performance on HICO-DET dataset as of the submission time on Papers with Code[1]. The model is based on deformable DETR [41] (a modified DETR), which computes self-attention from a limited range of feature maps for computational efficiency. This enables to extract multi-scale feature maps from the image. There is another technique to reduce computational costs in the encoder and decoder. The encoder is trained to generate query-based anchors which represent the points with high objectness score in the image. The decoder predicts objects and interactions only on those anchors for efficient computation. Although the backbone of the network can be pre-trained because it is based on Swin Transformer [42], it is impossible to pre-train all the encoder and decoder because

[1] `https://paperswithcode.com/sota/human-object-interaction-detection-on-hico`

there are some modifications to deformable DETR (object detection model). This is the reason why we do not report accuracies of pre-trained QAHOI in Tables 2 and 3.

STIP. "Structure-aware Transformer over Interaction Proposals" (STIP) [20] exhibits the best performance on V-COCO dataset as of the submission time on Papers with Code[2]. The model has two-stage architecture to perform HOI set prediction from non-parametric interaction queries detected by the independent instance detector. This enables the model to explore inter-interaction and intra-interaction structure from early epochs of the training by fixing the correspondence between interaction query and each target HOI. We can use pre-trained weights of DETR on MS COCO datasets for the backbone and object detector because the first stage of the network is exactly the same as DETR.

C.2 Pre-training, hyperparameters, and other conditions

In order to evaluate all the models on the equal condition in terms of pre-training, we report the results of HOTR, QPIC and STIP for both pre-trained and not pre-trained encoder and decoder, despite that the original papers encourage to use pre-trained weights for the best performance in the original HOI detection task. We note that for all of our experiments, the backbone of each model, i.e., feature extraction part, is pre-trained with the same weights as the original experiments.

For all of the models, we used almost the same hyperparameters as reported in the original papers and repositories. One change is the batch size of QAHOI: we reduced it from 2 to 1, because we got a memory error while training with the default setting. In addition, we trained QAHOI on V-COCO and V-COCO-SG with the same hyperparameters as HICO-DET and HICO-DET-SG, because the original paper does not report hyperparameters for V-COCO dataset.

We trained and tested seven types of model once on each split, and one training took about 1 to 2 days with 4 NVIDIA V100 GPUs.

D Further results and discussions

D.1 Difference between HICO-DET-SG and V-COCO-SG

Comparing two datasets, there are larger gaps between the original and the systematic generalization splits of V-COCO compared to HICO-DET. We speculate that the cause of this gap is the difference in the number of images and HOIs in the dataset. As shown in Table 1, HICO-DET-SG train data contains about 5.6 times as many images and about 12.6 times as many HOIs as the train data of V-COCO-SG. This means that there are more examples for one object-interaction combination in HICO-DET-SG train data, although the data has only 3.4 times as many variety of object-interaction combination classes. Therefore the models tend to achieve higher systematic generalization on HICO-DET-SG compared to V-COCO-SG.

D.2 Comparison across models

Each model showed different performance on systematic generalization splits. HOTR could not generalize at all as they achieved almost 0% mAP when its encoder and decoder are not pre-trained with object detection task, and achieved less than 3% mAP even with its encoder and decoder pre-trained. QAHOI also achieved low mAPs around 5% to show its inability to systematic generalization. However, QPIC generalizes to novel combinations to some extent especially when using pre-trained DETR weights: they achieved around 20% mAPs on HICO-DET-SG splits. STIP showed the best SG performance among all the models on both HICO-DET-SG and V-COCO-SG with pre-training. This may be because STIP has two-stage architecture. Instance and interaction detector is less affected by each other as they are independent. In other computer vision tasks, this type of modularity is proved to improve systematic generalization ability [29].

[2] `https://paperswithcode.com/sota/human-object-interaction-detection-on-v-coco`

D.3 Importance of pre-trained encoder and decoder

For three models except QAHOI, pre-training its encoder and decoder with the weights of DETR trained on object detection task plays a key role in improving systematic generalization performance. The accuracies on HICO-DET changed from around 0.4% to around 2.7% for HOTR, from around 12.1% to around 20.8% for QPIC and from 0.0% to around 23.2% for STIP. Also the accuracies on V-COCO changed from around 0.2% to around 2.1% for HOTR, from around 0.6% to around 3.8% for QPIC and from around 0.0% to 6.3% for STIP.

Generally vision Transformers require a lot of training to achieve high performance when trained from scratch [34]. So it is natural that without pre-training, the Transformer-based HOI detectors cannot solve merely HOI detection task nor systematic generalization. Especially STIP performed quite badly without pre-training because the number of training epochs is much smaller (30 epochs) than the others (100 epochs). For the three models, pre-trained initial weights of encoder and decoder contributes to improving systematic generalization accuracy as well as HOI detection task itself. In this paper we could not evaluate pre-trained QAHOI because the network is designed only for HOI detection task and it is hard to obtain pre-trained weights from other tasks such as object detection. If we modify the final part of QAHOI so that it can be trained with other tasks, we may be able to get better systematic generalization score with the pre-trained weights.

D.4 Evaluation on the train data

In order to make sure that the models are well-trained, we evaluated accuracies on the train data. In Table 5 and Table 6, we show mAP accuracies on the train data as well as the test data for HICO-DET-SG and V-COCO-SG, respectively (the right half of the tables are equivalent to Table 2 and Table 3). For both datasets and for all the models, mAPs on the train data of SG splits are coequal to or slightly better than the original splits. This is probably because the train data of SG splits contain less variety of triplets: in HICO-DET-SG there are only 540 combinations of objects and interactions out of 600 combination contained in the original dataset, and in V-COCO-SG there are only 160 out of 228.

Table 5: Comparison of systematic generalization performances on HICO-DET-SG data splits. We show mAPs (%) for both train and test data, which are higher the better (the right half of this table is equivalent to Table 2). The values reported in the original papers are written in brackets. The term "pre-train" means that the initial weights of the model's encoder and decoder are copied from DETR [31] trained on object detection. Test accuracies significantly decreased on HICO-DET-SG compared to original HICO-DET.

	Evaluation on train data (reference)							**Evaluation on test data (main)**						
	HOTR		QPIC		QAHOI	STIP		HOTR		QPIC		QAHOI	STIP	
pre-train	✗	✓	✗	✓	✗	✗	✓	✗	✓	✗	✓	✗	✗	✓
Original HICO-DET (mAP in the original paper)	30.54	44.80	33.29	46.28	50.23	13.47	32.23	17.63	26.30 (25.73)	21.70	29.59 (29.90)	35.30 (35.78)	13.21	31.57 (32.22)
HICO-DET-SG split1	33.92	46.03	34.05	49.70	52.94	13.69	33.17	0.31	2.59	10.57	22.08	4.53	0.00	22.14
HICO-DET-SG split2	31.48	42.04	30.23	48.28	51.50	15.13	34.44	0.28	2.73	12.53	19.95	4.76	0.00	23.03
HICO-DET-SG split3	32.05	44.91	35.54	47.90	48.27	13.25	30.61	0.52	2.75	13.28	20.50	6.12	0.00	24.39
Average over SG splits	32.48	44.33	33.30	48.63	50.90	14.02	32.74	0.37	2.69	12.13	20.84	5.14	0.00	23.19

Table 6: Comparison of systematic generalization performances on V-COCO-SG data splits. We show mAPs (%) for both train and test data, which are higher the better (the right half of this table is equivalent to Table 3). The values reported in the original papers are written in brackets. "pre-train" means that the initial weights of the model's encoder and decoder are copied from DETR [31] trained on object detection. Test accuracy significantly decreased on V-COCO-SG data splits compared to original V-COCO for all the models.

	Evaluation on train data (reference)							**Evaluation on test data (main)**						
	HOTR		QPIC		QAHOI	STIP		HOTR		QPIC		QAHOI	STIP	
pre-train	✗	✓	✗	✓	✗	✗	✓	✗	✓	✗	✓	✗	✗	✓
Original V-COCO (mAP in the original paper)	28.23	64.72	30.61	65.63	43.81	19.10	72.89	24.26	62.54 (63.8)	27.64	63.41 (61.0)	40.74	18.43	70.43 (70.7)
V-COCO-SG split1	30.57	65.79	31.24	67.25	45.52	23.51	71.91	0.24	2.10	0.77	4.21	2.73	0.12	6.25
V-COCO-SG split2	31.53	67.28	32.53	68.43	43.96	20.04	74.38	0.27	2.60	0.66	4.16	1.80	0.00	6.22
V-COCO-SG split3	28.21	61.07	29.83	60.47	42.88	22.41	73.43	0.15	1.68	0.32	2.89	1.53	0.00	6.37
Average over SG splits	30.10	64.71	31.20	65.38	44.12	21.99	73.24	0.22	2.13	0.58	3.75	2.02	0.04	6.28