Appendix for
RLIPv2: Fast Scaling of Relational Language-Image Pre-training

Hangjie Yuan[1*] Shiwei Zhang[2] Xiang Wang[3*] Samuel Albanie[4] Yining Pan[5*]
Tao Feng[2] Jianwen Jiang[2] Dong Ni[1†] Yingya Zhang[2] Deli Zhao[2]

[1]Zhejiang University [2]Alibaba Group [3]Huazhong University of Science and Technology
[4]CAML Lab, University of Cambridge [5]Singapore University of Technology and Design

{hj.yuan, dni}@zju.edu.cn wxiang@hust.edu.cn {pyn.sigrid, fengtao.hi, zhaodeli}@gmail.com
sma71@cam.ac.uk {zhangjin.zsw, jianwen.jjw, yingya.zyy}@alibaba-inc.com

A. Appendix

In this Appendix, we first elaborate on the societal impact (Appendix B), limitations (Appendix C) and the use of datasets (Appendix D) in RLIPv2. Next, we provide technical details of the box embedding (Appendix E) described in Sec. 4.2.2 of the main paper and cross attention modules (Appendix F) described in Eq. (2) of the main paper. Finally, we present additional experiments (Appendix G) to further validate the effectiveness of our approach.

B. Societal Impact

Our work RLIPv2 can potentially offer societal and commercial benefits. From the data perspective, RLIPv2 proposes a relational pseudo-labelling pipeline that avoids time- and cost-intensive work and obtains reasonable relation labels. From the pre-training perspective, RLIPv2 could perform more efficient pre-training and show prominent data efficiency that can potentially save computational cost. Nonetheless, we acknowledge that HOI detection and SGG are dual-use, meaning that it can be used for beneficial and malicious purposes. For example, the improved technologies can be applied to facilitate surveillance activities. Moreover, due to the bias of the pre-training and fine-tuning datasets, our algorithms can not guarantee equal performance for all demographic groups. Therefore, we emphasize that our pre-training and fine-tuning method is more of a proof-of-concept and requires rigorous evaluation and oversight when deploying for application.

C. Limitations

As mentioned in the main paper, our framework requires an external captioner to generate captions for relation parsing. One limitation of our method is that the performance depends on the quality of the captions. For instance, web-scale datasets for BLIP pre-training are usually noisy and lack diverse relation descriptions, *e.g.*, some words might be excessively used but convey only ambiguous information like "with" and "near". This is also confirmed by Tab. 6 of the main paper, showing that fine-tuning on curated style dataset like COCO Caption [4] is expected.

D. Datasets Used in This Work

Licenses. We uses three datasets for pre-training and three datasets for downstream transfer in RLIPv2. The following datasets are used, each governed by their license: the Objects365 [25], COCO [23], Visual Genome [17] and Open Images v6 [18] datasets are licensed under a Creative Commons Attribution 4.0 License; the HICO-DET [2, 1] dataset is licensed under a CC0: Public Domain license; the V-COCO [10] dataset is licensed under an MIT license.

Release of personally identifiable information/offensive content/consent. We affirm that no data will be disclosed as part of our research. Our research relies on public domian datasets: Objects365 [25], COCO [23], Visual Genome [17], Open Images v6 [18], HICO-DET [2, 1] and V-COCO [10], which we deem to pose a minimal risk of exposing personal information or offensive content. Regarding consent, we have not undertaken an independent inquiry beyond the scope of the original dataset releases.

E. Details about the Box Embedding

As mentioned in Sec. 4.2.2 of the main paper, we use box embeddings to encode labels and positions of the boxes $\hat{\boldsymbol{B}}_s, \hat{\boldsymbol{B}}_o$ into queries for decoding. Specifically, regarding the label embeddings, we adopt the gradient-detached language features after ALIF. To align the dimension of language features (*i.e.*, 768) with DDETR features (*i.e.*, 256), we apply linear projections to the language features.

*Work conducted during their research internships at DAMO Academy.
†Corresponding author.

Threshold	Overlap	Rare	Non-Rare	Full
0.7	✗	13.54	12.36	12.63
	✓	14.13	**15.01**	14.81
0.8	✗	12.66	12.76	12.74
	✓	**14.63**	14.94	**14.87**
0.9	✗	12.00	12.38	12.49
	✓	13.28	14.35	14.10
0.95	✗	11.66	12.08	11.98
	✓	13.69	14.37	14.21

Table 1: **Parameter sensitivity analysis of the threshold for the CLIP tagging method.** "Overlap" denotes the "overlap" prior for SO pairs introduced in Sec. 4.2.2. We report zero-shot (NF) results pre-trained on VG and COCO.

Threshold η	Rare	Non-Rare	Full
-	12.12	14.07	13.62
0.1	12.95	14.98	14.49
0.15	13.12	15.14	14.67
0.2	**15.33**	**15.54**	**15.49**
0.25	12.81	14.68	14.25
0.3	11.25	14.52	13.77

Table 2: **Parameter sensitivity analysis of η for R-Tagger.** Zero-shot (NF) is reported after pre-training RLIPv2-ParSeD on VG and pseudo-labelled COCO. "-" denotes pre-training on VG.

Regarding the position embedding, we project the boxes (x, y, w, h) into 256 dimensions where x, y are center coordinates and w, h are width and height of the box. Equipped with these two embeddings as queries, we can perform DDETR decoding [33].

F. Details about Cross-attn$(\boldsymbol{C}^{(0)}, \boldsymbol{L}^{(0)})$

To perform cross attention as mentioned in Eq. (2) of the main paper, we compute the attention scores of one modality with respect to the other modality, and then use scores to aggregate features from the other modality [6, 31, 7]. Specifically, we follow the instantiation of the cross attention module from [20, 31]. The calculation can be formulated as:

$$\boldsymbol{C}^{(0,q)} = \boldsymbol{C}^{(0)}\boldsymbol{W}_1, \boldsymbol{L}^{(0,q)} = \boldsymbol{L}^{(0)}\boldsymbol{W}_2, Att = \frac{\boldsymbol{C}^{(0,q)}(\boldsymbol{L}^{(0,q)})^T}{\sqrt{d}} \tag{1}$$

$$\boldsymbol{L}^{(0,v)} = \boldsymbol{L}^{(0)}\boldsymbol{W}_3, \tilde{\boldsymbol{C}}^{(0)} = \text{softmax}(Att)\boldsymbol{L}^{(0,v)}\boldsymbol{W}_4 \tag{2}$$

$$\boldsymbol{C}^{(0,v)} = \boldsymbol{C}^{(0)}\boldsymbol{W}_5, \tilde{\boldsymbol{L}}^{(0)} = \text{softmax}(Att^T)\boldsymbol{C}^{(0,v)}\boldsymbol{W}_6 \tag{3}$$

where $\boldsymbol{W}_1, \boldsymbol{W}_2$ are trainable parameters for query embedding; $\boldsymbol{W}_3, \boldsymbol{W}_5$ are trainable parameters for value embedding; $\boldsymbol{W}_4, \boldsymbol{W}_6$ are trainable parameters for output embedding; Att denotes attention scores; d is embedding dimension, which is set to 256 following [31, 20]; the feature dimension of $\boldsymbol{C}^{(0)}$ and $\boldsymbol{L}^{(0)}$ are 256 and 768, respectively. Note that to reduce computation, we only perform cross attention on the flattened vision features from the last layer with the smallest scale.

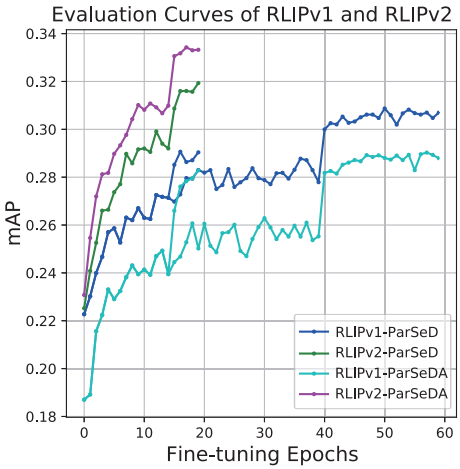

Figure 1: **Evaluation curve comparisons of RLIPv1 and RLIPv2 on HICO-DET.** RLIPv1 and RLIPv2 are both pre-trained on VG. By default, RLIPv1 is fine-tuned for 60 epochs following [27] and RLIPv2 is fine-tuned for 20 epochs. We also fine-tune RLIPv1 for 20 epochs for a fair comparison.

Model	Datasets	Epochs	Zero-shot (NF)
RLIP-ParSeDA	VG	50	11.34 / 14.56 / 13.82
RLIP-ParSeDA	VG+COCO	50	15.05 / 15.55 / 15.44
RLIPv2-ParSeDA	VG	20	13.03 / 14.98 / 14.53
RLIPv2-ParSeDA	VG+COCO	20	15.00 / 16.60 / 16.23

Table 3: **Comparisons of scaling with different models.** We use ResNet-50 by default. Results are evaluated on HICO-DET under zero-shot (NF) setting.

Model	Pseudo-label Type	Rare	Non-Rare	Full
RLIPv2-ParSeDA	R-Tagger	15.00	16.60	16.23
	RLIPv2-ParSeD	11.96	14.87	14.20

Table 4: **Comparisons of different pseudo-label types.** RLIPv2-ParSeDA is pre-trained on VG and pseudo-labelled COCO. Results are evaluated on HICO-DET under zero-shot (NF) setting.

G. Additional Experiments

The choice of the threshold for the CLIP tagging method. The CLIP tagging method requires a pre-defined threshold to tag relations as described in the Experiment section of the main paper (**Comparisons of different relation tagging strategies.**). Therefore, to choose an op-

	VG	VG+COCO	VG+COCO+O365
ResNet-50	12.12 / 14.07 / 13.62	15.08 / 15.10 / 15.09	17.21 / 16.84 / 16.93
Swin-T	12.17 / 15.01 / 14.36	14.89 / 16.70 / 16.28	20.34 / 18.27 / 18.75
Swin-L	15.19 / 17.46 / 16.94	20.03 / 19.75 / 19.81	26.75 / 20.61 / 22.02

Table 5: **Model and dataset scaling experiments using RLIPv2-ParSeD.** Results are evaluated on HICO-DET *Rare*/*Non-Rare*/*Full* sets under zero-shot (NF) setting.

Method	Backbone	UC-RF	UC-NF
VCL [12]	ResNet-50	10.06 / 24.28 / 21.43	16.22 / 18.52 / 18.06
ATL [13]	ResNet-50	9.18 / 24.67 / 21.57	18.25 / 18.78 / 18.67
FCL [14]	ResNet-50	13.16 / 24.23 / 22.01	18.66 / 19.55 / 19.37
GEN-VLKT [22]	ResNet-50	21.36 / 32.91 / 30.56	**25.05** / 23.38 / 23.71
RLIPv1-ParSeD [27]	ResNet-50	16.43 / 30.59 / 27.76	16.99 / 24.71 / 22.93
RLIP-ParSe [27]	ResNet-50	19.19 / 33.35 / 30.52	20.27 / 27.67 / 26.19
RLIPv2-ParSeD	ResNet-50	**19.33** / **34.22** / **31.24**	21.18 / **28.95** / **27.40**
RLIPv2-ParSeD	Swin-T	**23.80** / **38.23** / **35.34**	21.88 / **33.63** / **31.28**
RLIPv2-ParSeD	Swin-L	**30.98** / **43.67** / **41.13**	23.16 / **39.97** / **36.61**
RLIPv2-ParSeDA	ResNet-50	**21.45** / **35.85** / **32.97**	22.81 / **29.52** / **28.18**
RLIPv2-ParSeDA	Swin-T	**26.95** / **39.92** / **37.32**	21.07 / **35.07** / **32.27**
RLIPv2-ParSeDA	Swin-L	**31.23** / **45.01** / **42.26**	22.65 / **40.51** / **36.94**

Table 6: **Comparisons with methods on HICO-DET under UC-RF and UC-NF settings.** Results are reported on *Unseen*/*Seen*/*Full* sets.

timal threshold, we traverse a range of values and evaluate the pre-training performance. The results are presented in Tab. 1. From this table, we observe that utilizing the "overlap" prior can constantly outperform its naive counterpart without the "overlap" prior. Moreover, our analysis indicates that a threshold of 0.8 leads to the best performance for the CLIP tagging method.

The choice of threshold η for R-Tagger. To choose pseudo-labelled triplets for pre-training, we traverse a range of η values for R-Tagger to select triplets with relation confidence higher than this threshold. By default, we adopt oracle captions from **COCO Caption** [4], which provides an average of $N_{Cap} = 5$ captions per image. We tag pseudo-labels on COCO and pre-train RLIPv2-ParSeD on VG and pseudo-labelled COCO. The results are presented in Tab. 2, which indicates that the optimal value for η is 0.2. It is worth noting that all experiments in this table are initialized with COCO object detection parameters. Therefore, the highest performance boost ($13.62 \rightarrow 15.49$) is obtained by the additional tagged relations from COCO, rather than by the inclusion of an additional COCO dataset.

Evaluation curve comparisons. As detailed in Tab. 3 of the main paper, we show the comparisons of RLIPv1 and RLIPv2 concerning pre-training and fine-tuning. To further validate the effectiveness of ALIF, we compare their fine-tuning evaluation curves on HICO-DET in Fig. 1. As can be observed from the figure, RLIPv2 can converge much faster than its RLIPv1 counterparts.

Dataset scaling using RLIP. This paper introduces RLIPv2 as a novel method that facilitates scaling due to its convergence speed. Another alternative is to adopt RLIPv1 to scale up relational pre-training. In Tab. 3, we aim to compare the effect of scaling using RLIPv1 and RLIPv2. We can observe that by performing earlier and deeper gated fusion, ALIF gains slightly better performance boost by costing $0.4\times$ pre-training time. Therefore, RLIPv2 is a more efficient and scalable approach for scaling up relational pre-training.

Comparisons of pseudo-label types. To validate the effectiveness of R-Tagger that utilizes groundtruth object annotations as input, we compare R-Tagger with another baseline that generates pseudo-labels by applying the pre-trained RLIPv2 model to perform SGG on object detection datasets as mentioned in Sec. 4.2.2. To leverage the annotated object boxes for the new baseline, we match the generated boxes with the annotated ones. If a generated box is matched with one given annotated box (IoU > 0.5), we substitute the generated box with the annotated box to ensure the accuracy of the box position. To seek a fair comparison, we use the pre-trained RLIPv2-ParSeD (ResNet-50) to generated pseudo-triplets, which is identical to the base model of R-Tagger. To select triplets, we use an identical selection threshold to R-Tagger, *i.e.*, 0.2. Then, we pre-train on VG and pseudo-labelled COCO datasets to compare the quality

Method	Backbone	1% Data	10% Data
RLIP-ParSeD [27]	ResNet-50	16.22 / 18.92 / 18.30	15.89 / 23.94 / 22.09
RLIP-ParSe [27]	ResNet-50	17.47 / 18.76 / 18.46	20.16 / 23.32 / 22.59
RLIPv2-ParSeD	ResNet-50	**19.87 / 24.04 / 23.08**	**21.51 / 27.84 / 26.38**
RLIPv2-ParSeD	Swin-T	**26.37 / 27.29 / 27.08**	**27.85 / 31.41 / 30.59**
RLIPv2-ParSeD	Swin-L	**30.49 / 30.80 / 30.73**	**33.90 / 35.93 / 35.46**
RLIPv2-ParSeDA	ResNet-50	**22.13 / 24.51 / 23.96**	**23.28 / 30.02 / 28.46**
RLIPv2-ParSeDA	Swin-T	**24.26 / 28.92 / 27.85**	**28.31 / 32.93 / 31.87**
RLIPv2-ParSeDA	Swin-L	**31.89 / 32.32 / 32.22**	**34.75 / 38.27 / 37.46**

Table 7: **Comparisons with methods on HICO-DET under few-shot settings.** Results are reported on *Rare*/*Non-Rare*/*Full* sets.

Model	Backbone	Extra Relations	HICO-DET		V-COCO	
			Zero-shot (NF)	Fully-finetuning	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
InteractNet [9]	R50-FPN	-	-	7.16 / 10.77 / 9.94	40.0	-
UnionDet [15]	R50-FPN	-	-	11.72 / 19.33 / 17.58	47.5	56.2
PPDM [21]	HG104	-	-	13.97 / 24.32 / 21.94	-	-
HOTR [16]	R50	-	-	17.34 / 27.42 / 25.10	55.2	64.4
QPIC [26]	R50	-	-	21.85 / 31.23 / 29.07	58.8	61.0
OCN [28]	R50	-	-	25.56 / 32.51 / 30.91	64.2	66.3
CDN [30]	R50	-	-	27.39 / 32.64 / 31.44	61.7	63.8
GEN-VLKT [22]	R50	-	-	29.25 / 35.10 / 33.75	62.4	64.5
QAHOI [3]	Swin-L*	-	-	29.80 / 37.56 / 35.78	-	-
UniVRD [32]	ViT-H/14†	-	-	31.65 / 39.99 / 38.07	65.8	66.9
RLIPv1-ParSeD [27]	R50	VG	11.20 / 14.73 / 13.92	24.67 / 32.50 / 30.70	61.7	63.8
RLIPv1-ParSe [27]	R50	VG	15.08 / 15.50 / 15.40	26.85 / 34.63 / 32.84	61.9	64.2
RLIPv2-ParSeD	R50	VG	12.12 / 14.07 / 13.62	26.47 / 33.51 / 31.89	61.9	64.5
RLIPv2-ParSeD	R50	VG+COCO	15.08 / 15.10 / 15.09	26.61 / 33.78 / 32.13	62.9	65.3
RLIPv2-ParSeD	R50	VG+COCO+O365	**17.21 / 16.84 / 16.93**	**27.27 / 35.08 / 33.29**	**63.8**	**66.4**
RLIPv2-ParSeD	Swin-T	VG+COCO+O365	**20.34 / 18.27 / 18.75**	**31.44 / 38.51 / 36.89**	**66.6**	**69.1**
RLIPv2-ParSeD	Swin-L	VG+COCO+O365	**26.75 / 20.61 / 22.02**	**42.76 / 44.67 / 44.23**	**71.0**	**73.2**
RLIPv2-ParSeDA	R50	VG	13.03 / 14.98 / 14.53	27.01 / 35.21 / 33.32	63.0	65.1
RLIPv2-ParSeDA	R50	VG+COCO	15.00 / 16.60 / 16.23	27.89 / 35.27 / 33.57	64.5	66.7
RLIPv2-ParSeDA	R50	VG+COCO+O365	**19.64 / 17.24 / 17.79**	**29.61 / 37.10 / 35.38**	**65.9**	**68.0**
RLIPv2-ParSeDA	Swin-T	VG+COCO+O365	**21.24 / 19.47 / 19.87**	**33.66 / 40.07 / 38.60**	**68.8**	**70.8**
RLIPv2-ParSeDA	Swin-L	VG+COCO+O365	**27.97 / 21.90 / 23.29**	**43.23 / 45.64 / 45.09**	**72.1**	**74.1**

Table 8: **Comparisons with previous methods on HICO-DET and V-COCO.** Results on HICO-DET are reported on *Rare*/*Non-Rare*/*Full* sets. R50 and HG denote ResNet-50 [11] and Hourglass [24]. * denotes the backbone is pre-trained with 384×384 resolution, while others use 224×224. †indicates the backbone is pre-trained using LiT [29], then fine-tuned on Objects365, COCO and HICO with the objective of object detection.

of the pseudo-labels in Tab. 4. As can be observed from the table, by utilizing groundtruth box information when inferring relations, R-Tagger can generate more authentic pseudo-labels, thus benefiting relational pre-training.

Model scaling and dataset scaling using RLIPv2-ParSeD. In addition to scaling experiments using RLIPv2-ParSeDA in the main paper, we also present the model and dataset scaling experiments using RLIPv2-ParSeD in Tab. 5. In terms of data, adding COCO and Objects365 can both boost performance, and the benefit of adding data exhibits a log scaling trend [5]. Models pre-trained with Objects365 consistently have better *Rare* result, which we attribute to the distribution misalignment of Objects365 and HICO-DET [8]. In terms of models, switching to stronger backbone models can improve the data efficiency at the cost of larger amounts of computation.

More results using RLIPv2-ParSeD on HICO-DET under UC-NF and UC-RF settings. We provide more results under UC-NF and UC-RF settings using RLIPv2-ParSeD in addition to RLIPv2-ParSeDA in Tab. 6. It is worth noting that UC-RF denotes 120 unseen combinations (UC) among 600 combinations are selected by a rare-first (RF) order, while UC-NF denotes that 120 unseen combinations among 600 combinations are selected by a non-rare

Caption type	N_{Cap}	N_{Unique}	Rare	Non-Rare	Full
- (baseline: *w/o* captions)	-	-	12.12	14.07	13.62
BLIP (beam)	1	1	9.86	12.02	11.52
BLIP (nucleus)	10	9.97	15.08	15.10	15.09
BLIP-2 (beam)	1	1	9.98	12.23	11.72
BLIP-2 (nucleus)	10	3.26	11.76	12.85	12.60
BLIP + BLIP-2 (nucleus)	20	13.18	14.74	15.52	**15.34**
BLIP (dense captions, beam)	28.63	10.40	14.25	15.14	14.94

Table 9: **Comparisons of different captioners.** "beam" and "nucleus" denote beam search and nucleus sampling. N_{Unique} denotes the number of unique captions after deduplication. By default, we adopt COCO Caption fine-tuned BLIP and BLIP-2 model.

first (NF) order. We can observe that **(i)** under the UC-RF setting, switching to stronger backbones improves the performance of all metrics; **(ii)** under the UC-NF setting, switching to stronger backbones enhances all metrics except the metric on the *Unseen* set. We attribute this to the significant object distribution misalignment between *Seen* and *Unseen* sets.

More results on few-shot HOI detection. We provide more results under few-shot settings using RLIPv2-ParSeD in addition to RLIPv2-ParSeDA in Tab. 7. We can observe that RLIPv2 exhibits remarkable data efficiency by scaling up pre-training. Notably, the largest pre-trained model obtains 32.22mAP when fine-tuned on 1% data, which outperforms many methods that fine-tune on 100% data.

More results under fully-finetuning and zero-shot (NF) settings on HOI detection. We present more results using RLIPv2-ParSeD on HICO-DET and V-COCO in Tab. 8. We draw similar conclusions from this table that **(i)** dataset and model scaling can both boost the final performance on two datasets; **(ii)** on HICO-DET, the benefit of pre-training is more prominent on zero-shot than fully-finetuning, especially on the *Rare* set.

The effect of using diverse captioning models. To comprehensively study the effect of using other captioners, we adopt the more advanced BLIP-2 [19] to implement our method. The results are shown in Tab. 9. The results indicate that BLIP-2 (beam) slightly outperforms BLIP (beam), but BLIP-2 (nucleus) trails BLIP (nucleus). We attribute this to the low diversity of BLIP-2 captions as BLIP-2 generates more deterministic captions. To verify this hypothesis, we combine captions from BLIP and BLIP-2, obtaining better generalization performance.

The effect of dense captioning on pairwise union regions. We show the result in the final row of the above table. Dense captioning on pairwise union regions can improve over the baseline (the first row), offering an alternative to improve the caption diversity in addition to utilizing nucleus sampling. Thus, the potential for developing more advanced captioning schemes to bolster the quality of pseudo-labels holds considerable promise. We mark this endeavor as a research focus for one of our future works.

References

[1] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, pages 381–389. IEEE, 2018. 1

[2] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *ICCV*, pages 1017–1025, 2015. 1

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

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1, 3

[5] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143*, 2022. 4

[6] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *arXiv preprint arXiv:2206.07643*, 2022. 2

[7] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *CVPR*, pages 18166–18176, 2022. 2

[8] Rahim Entezari, Mitchell Wortsman, Olga Saukh, M Moein Shariatnia, Hanie Sedghi, and Ludwig Schmidt. The role of pre-training data in transfer learning. *arXiv preprint arXiv:2302.13602*, 2023. 4

[9] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, pages 8359–8367, 2018. 4

[10] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[12] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, pages 584–600. Springer, 2020. 3

[13] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *CVPR*, pages 495–504, 2021. 3

[14] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *CVPR*, pages 14646–14655, 2021. 3

[15] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *ECCV*, pages 498–514. Springer, 2020. 4

[16] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *CVPR*, pages 74–83, 2021. 4

[17] 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. *International Journal of Computer Vision*, 123(1):32–73, 2017. 1

[18] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 1

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 5

[20] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. *arXiv preprint arXiv:2112.03857*, 2021. 2

[21] 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 *CVPR*, pages 482–490, 2020. 4

[22] 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*, pages 20123–20132, June 2022. 3, 4

[23] 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*, pages 740–755. Springer, 2014. 1

[24] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016. 4

[25] 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 *ICCV*, pages 8430–8439, 2019. 1

[26] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, pages 10410–10419, 2021. 4

[27] Hangjie Yuan, Jianwen Jiang, Samuel Albanie, Tao Feng, Ziyuan Huang, Dong Ni, and Mingqian Tang. Rlip: Relational language-image pre-training for human-object interaction detection. In *NeurIPS*, 2022. 2, 3, 4

[28] Hangjie Yuan, Mang Wang, Dong Ni, and Liangpeng Xu. Detecting human-object interactions with object-guided cross-modal calibrated semantics. *AAAI*, 36(3):3206–3214, Jun. 2022. 4

[29] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, pages 18123–18133, 2022. 4

[30] 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. *NeurIPS*, 34, 2021. 4

[31] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. In *NeurIPS*, 2022. 2

[32] Long Zhao, Liangzhe Yuan, Boqing Gong, Yin Cui, Florian Schroff, Ming-Hsuan Yang, Hartwig Adam, and Ting Liu. Unified visual relationship detection with vision and language models. *arXiv preprint arXiv:2303.08998*, 2023. 4

[33] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2