This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Abstract

Relational Language-Image Pre-training (RLIP) aims to align vision representations with relational texts, thereby advancing the capability of relational reasoning in computer vision tasks. However, hindered by the slow convergence of RLIPv1[1] architecture and the limited availability of existing scene graph data, scaling RLIPv1 is challenging. In this paper, we propose RLIPv2, a fast converging model that enables the scaling of relational pre-training to large-scale pseudo-labelled scene graph data. To enable fast scaling, RLIPv2 introduces Asymmetric Language-Image Fusion (ALIF), a mechanism that facilitates earlier and deeper gated cross-modal fusion with sparsified language encoding layers. ALIF leads to comparable or better performance than RLIPv1 in a fraction of the time for pre-training and fine-tuning. To obtain scene graph data at scale, we extend object detection datasets with free-form relation labels by introducing a captioner (e.g., BLIP) and a designed Relation Tagger. The Relation Tagger assigns BLIP-generated relation texts to region pairs, thus enabling larger-scale relational pre-training. Through extensive experiments conducted on Human-Object Interaction Detection and Scene Graph Generation, RLIPv2 shows state-of-the-art performance on three benchmarks under fully-finetuning, few-shot and zero-shot settings. Notably, the largest RLIPv2 achieves 23.29mAP on HICO-DET without any fine-tuning, yields 32.22mAP with just 1% data and yields 45.09mAP with 100% data. Code and models are publicly available at https://github.com/JacobYuan7/RLIPv2.

1. Introduction

The pretraining-finetuning paradigm has achieved major breakthroughs in vision and language domains [58, 13, 23, 103, 65, 77]. In this context, a number of particularly notable results have been obtained through *aligned* Vision-Language Pre-training (VLP) [65, 42, 41, 78, 95, 91]. These research efforts have typically employed a robust ***base model*** [14, 25, 12, 74] that is trained on language-image paired ***data*** to produce foundation models.

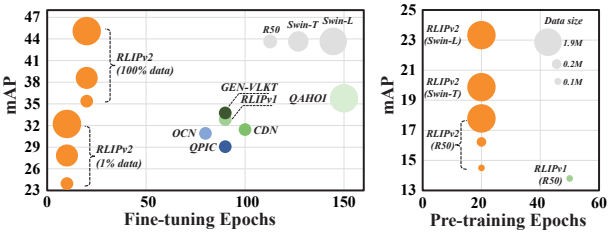

Figure 1: **Left**: Fine-tuning comparison on HICO-DET. **Right**: Pre-training epoch and zero-shot (NF) comparison on HICO-DET. Except where stated, RLIPv2-ParSeDA architecture is adopted.

RLIPv1 [86] presents the first attempt to specifically align vision representations and relational texts using VLP. By pre-training on ***open-vocabulary scene graph data*** like Visual Genome (VG) [38], RLIPv1 demonstrates its usefulness in zero-shot, few-shot and fully-finetuned Human-Object Interaction (HOI) Detection. Although RLIPv1 is proven effective, we find it challenging to scale for the following reasons:

(i) ***Model*** perspective: RLIPv1 converges slowly, as exemplified by DETR [3]-based RLIPv1, which requires 150/90 epochs to converge during pre-training/fine-tuning. Even when building on Deformable DETR (DDETR) [102] and DAB-DDETR [56], 50/60 epochs are still required.

(ii) ***Data*** perspective: as observed by the authors of RLIPv1, data with relation triplet annotations is scarce, impeding RLIPv1's scaling. Annotating triplets in the form of $\langle$*subject, relations, object*$\rangle$ is both time- and cost-intensive.

To resolve the aforementioned challenges, we introduce RLIPv2, a fast converging model that enables relational pre-training on larger-scale pseudo-labelled scene graph data.

*Work conducted during their research internships at DAMO Academy.
†Corresponding author.
[1]RLIPv1 refers to the model presented in [86], and RLIPv2 refers to the model presented in this paper.

From both the model and data perspectives, we summarize the contributions of RLIPv2 as follows.

From the ***model*** perspective, we observe that the slow convergence of DDETR can be attributed to the late language-image fusion strategy: fusion after decoding. Prior works [15, 44] have demonstrated an earlier and deeper fusion mechanism facilitates cross-modal alignment. In light of this, we propose Asymmetric Language-Image Fusion (ALIF) in RLIPv2 that encourages fusion in the encoding stage with sparsified language layers. Without sacrificing inference speed thanks to sparsification, RLIPv2 requires only 20 epochs to pre-train and fine-tune based on DDETR family models [102, 56] as shown in Fig. 1, while performing better than or comparably to RLIPv1.

From the ***data*** perspective, we leverage well-established object detection datasets [53, 69, 39]. Specifically, we extend these datasets with relational annotations by pseudo-labelling. To perform pseudo-labelling, we must tackle two challenges: **(i)** sorting out the relations contained in the image and **(ii)** tagging relation texts to region pairs. Regarding the first challenge, we employ external captioners (*e.g.* BLIP [42]) that generate captions containing relation descriptions. Regarding the second challenge, we reuse RLIPv2 model as a Relation Tagger (R-Tagger) that enables assigning the generated open-vocabulary relation texts to region pairs. Equipped with such a pipeline, we investigate the scaling behavior of both the model and the data for RLIPv2, which demonstrates improved zero-shot, few-shot and fine-tuning performance.

Furthermore, we introduce Scene Graph Generation (SGG), an analogously defined task to HOI detection for evaluating RLIPv2. RLIPv2 achieves state-of-the-art performance on Open Images v6 [39] for SGG, which underscores its robustness and efficacy in tackling relational reasoning tasks.

2. Related Work

Language-image pre-training for detection. Recently, there has been a growing interest in learning visual representations from language supervision [42, 41, 65, 1, 32, 34, 15, 44, 95, 76]. This paradigm of learning from language supervision has also proven effective in improving detection performance. MDETR [34] was the first work to learn region-text correspondences in an end-to-end manner, while X-DETR [2] improved upon MDETR by removing the cross-modal fusion part that improves its training efficiency. GLIP [44] extended this line of research by scaling to web-scale data, leading to significant advances in zero-shot object detection and data efficiency. DetCLIP [84] proposed a paralleled concept formulation and a concept dictionary to enable semantically rich region-text alignment. RLIPv1 [86] was the first work to seek language-image alignment via relations, and our work follows its footsteps to achieve fast scaling of relational pre-training.

End-to-end HOI detection and scene graph generation. Relations can interpret visual content in a fine-grained perspective [10, 31, 33]. Detecting and recognizing relations utilizing HOI detection and scene graph generation have been verified effective in image captioning [85], image retrieval [33, 75], synthesis [83, 18], activity understanding [31, 88, 87] and . The aim of these tasks is to detect relation triplets from a given input image. Before the emergence of DETR [3], the commonly adopted pipeline was to adopt an off-the-shelf object detector [66, 24] and design reasoning modules to infer relations [20, 19, 63, 49, 48, 90, 7, 54, 46, 47, 81]. Initial end-to-end design efforts extended detectors to support relation recognition [50, 99, 82, 35, 55]. Further attempts focus on the adaptation of DETR to the field of HOI detection [89, 6, 36, 104, 71, 93, 94, 79, 37, 57, 101, 62, 64, 97] and SGG [45, 11]. RLIPv1 [86] demonstrated the effectiveness of relational pre-training in improving performance and data efficiency. RLIPv2 builds upon RLIPv1, but seeks to solve its scaling problem. The most related work [100] also seeks scaling with the help of language. However, it remains a two-stage detection pipeline, lacks semantic richness and adopts a naive matching algorithm that hinders its performance.

3. Recap of RLIPv1

In this section, we will briefly review RLIPv1 [86], a model that leverages both entity and relation descriptions to perform VLP and that forms the basis for our approach. As triplet detection architecture, RLIPv1 proposes a ParSeD model that allocates decoupled embeddings for subjects, objects and relations. The collective model RLIPv1-ParSeD consists of three stages: *Parallel Entity Detection*, *Sequential Relation Inference* and *Cross-Modal Fusion*, as shown in Fig. 2(a).

For *Parallel Entity Detection*, RLIPv1-ParSeD defines two sets of queries $\boldsymbol{Q}_s, \boldsymbol{Q}_o \in \mathbb{R}^{N_Q \times D}$ with N_Q pairs of subjects and objects to perform subject and object detection. For *Sequential Relation Inference*, the model generates relation queries $\boldsymbol{Q}_r \in \mathbb{R}^{N_Q \times D}$ based on the decoded subject and object queries $\tilde{\boldsymbol{Q}}_s, \tilde{\boldsymbol{Q}}_o \in \mathbb{R}^{N_Q \times D}$, and performs decoding to obtain $\tilde{\boldsymbol{Q}}_r \in \mathbb{R}^{N_Q \times D}$ for relation recognition. The design of RLIPv1-ParSeD obeys the following probabilistic factorization:

$$\begin{aligned} &\mathbb{P}(\boldsymbol{G}|\boldsymbol{Q}_s, \boldsymbol{Q}_o, \boldsymbol{C}_E; \boldsymbol{\theta}_{Par}, \boldsymbol{\theta}_{Se}) = \\ &\mathbb{P}(\boldsymbol{B}_s, \boldsymbol{B}_o|\boldsymbol{Q}_s, \boldsymbol{Q}_o, \boldsymbol{C}_E; \boldsymbol{\theta}_{Par}) \cdot \mathbb{P}(\boldsymbol{R}|\boldsymbol{B}_s, \boldsymbol{B}_o, \boldsymbol{C}_E; \boldsymbol{\theta}_{Se}) \end{aligned} \tag{1}$$

where $\boldsymbol{C}_E$ denotes features from the DDETR encoder[2]; $\boldsymbol{\theta}_{Par}, \boldsymbol{\theta}_{Se}$ denote parameters for *Parallel Entity Detection*

[2]Since RLIPv2 targets fast scaling, we only focus on DDETR-based detection architecture.

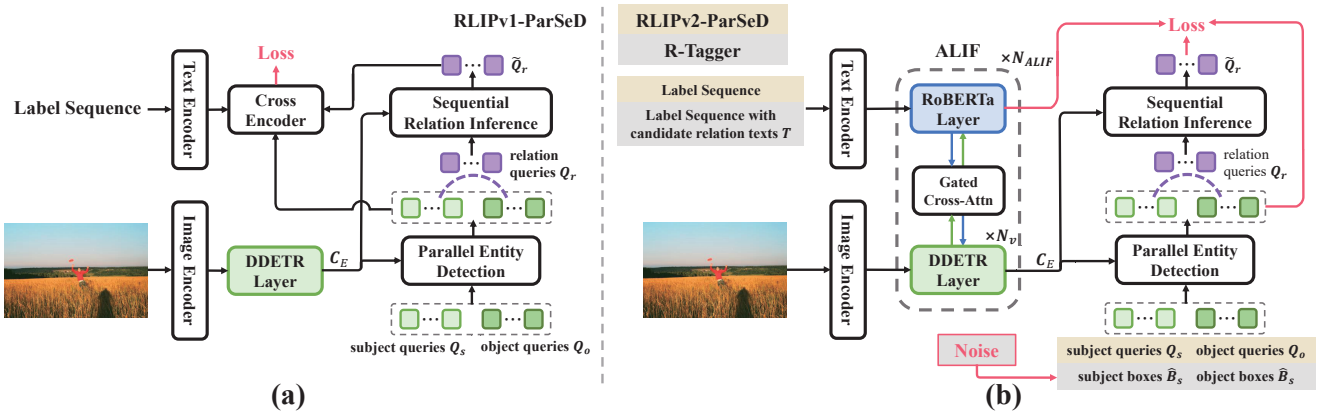

Figure 2: **The overview of (a) RLIPv1-ParSeD and (b) RLIPv2-ParSeD and R-Tagger.** The red part (loss calculation and noise injection) is only valid during training. In (a), *Cross-Modal Fusion* is achieved by the cross encoder. In (b), *Cross-Modal Fusion* is achieved by ALIF. The two architectures have an equivalent number of DDETR layers.

and *Sequential Relation Inference*; $\boldsymbol{B}_s, \boldsymbol{B}_o, \boldsymbol{R}$ are sets of detected subject boxes, object boxes and relations, respectively. They collectively comprise the detection results $\boldsymbol{G}$.

For *Cross-Modal Fusion*, RLIPv1-ParSeD appends additional Transformer encoding layers [74, 34, 43] to perform language-image feature fusion on top of the decoded relation features $\boldsymbol{Q}_r$, entity label features $\boldsymbol{L}_E \in \mathbb{R}^{N_E \times D}$ and relation label features $\boldsymbol{L}_R \in \mathbb{R}^{N_R \times D}$. $\boldsymbol{L}_E, \boldsymbol{L}_R$ are extracted from RoBERTa [58].

4. Methodology

In this section, we will introduce: **(i)** Asymmetric Language-Image Fusion (ALIF) as an efficient *Cross-Modal Fusion* mechanism in RLIPv2, as shown in Fig. 2(b); **(ii)** the overall framework of extending the off-the-shelf object detection datasets [53, 69] to support relational pre-training.

4.1. Asymmetric Language-Image Fusion

The core idea underpinning ALIF is to perform efficient cross-modal fusion in the early stages of RLIPv2 as highlighted by [15, 44]. Unlike RLIPv1 that encourages cross-modal alignment for entities and relations after the decoding phase, ALIF performs this during the detection encoding phase. This is particularly challenging for DDETR's encoder, since it relies on deformable attention that makes it challenging to adopt dedicated encoder layers during the detection encoding phase like [34, 60, 16].

To address this, we propose ALIF, a mechanism that leverages DDETR encoding for the vision branch, RoBERTa encoding for the language branch and gated cross-attention for fusion. In contrast to previous work that encodes image and language with an equivalent number of layers [34, 16, 44, 15], we experimentally find that excessive RoBERTa layers do not improve its generalization capability due to its potential for overfitting to pre-trained data. Moreover, such a paradigm results in training difficulty due to the increased model complexity. As a result, we perform DDETR encoding densely while performing RoBERTa encoding sparsely. We denote the vision features from the backbone as $\boldsymbol{C}^{(0)}$ and language features from RoBERTa as $\boldsymbol{L}^{(0)}$ (the concatenation of $\boldsymbol{L}_E$ and $\boldsymbol{L}_R$). The first ALIF module can be formulated as:

$$\tilde{\boldsymbol{C}}^{(0)}, \tilde{\boldsymbol{L}}^{(0)} = \text{Cross-attn}(\boldsymbol{C}^{(0)}, \boldsymbol{L}^{(0)}) \tag{2}$$

$$\boldsymbol{C}^{(N_v)} = \text{DDETR}^{N_v}(\boldsymbol{C}^{(0)} + \text{G}(\tilde{\boldsymbol{C}}^{(0)})) \tag{3}$$

$$\boldsymbol{L}^{(1)} = \text{RoBERTa}^{1}(\boldsymbol{L}^{(0)} + \text{G}(\tilde{\boldsymbol{L}}^{(0)})) \tag{4}$$

where $\tilde{\boldsymbol{C}}^{(0)}$ denotes language features aggregated by cross attention, and $\tilde{\boldsymbol{L}}^{(0)}$ is analogously defined; N_v is the number of DDETR layers in one ALIF; $\text{G}(\boldsymbol{x})$ defines a gating function that gates the aggregated cross-modal features. Note that $\boldsymbol{C}_E$ is $\boldsymbol{C}^{(N_v)}$ after stacking N_{ALIF} ALIF layers for encoding. For the instantiation of $\text{G}(\boldsymbol{x})$, we experiment with three options: **(i)** $\text{G}(\boldsymbol{x}) = \alpha\boldsymbol{x}$ where α is a learnable scalar; **(ii)** $\text{G}(\boldsymbol{x}) = \boldsymbol{a}\boldsymbol{x}$ where $\boldsymbol{a}$ is a learnable vector; **(iii)** $\text{G}(\boldsymbol{x}) = \text{SE}(\boldsymbol{x})$ where SE denotes a Squeeze-and-Excitation block [30] with a reduction ratio of 4. We also experiment augmenting $\text{G}(\boldsymbol{x})$ with $\tanh()$ introduced in [1] (*e.g.* $\tanh(\alpha\boldsymbol{x})$), while only observing side effects. This indicates the magnitude of language and vision features are disparate, and a proper gating method is desired.

4.2. Relational Pseudo-labelling

RLIPv2 reuses object detection datasets to benefit pre-training by pseudo-labelling, as shown in Fig. 3. We assume that during the pre-training phase, the downstream entity and relation distributions are unavailable to us. Then, to tag

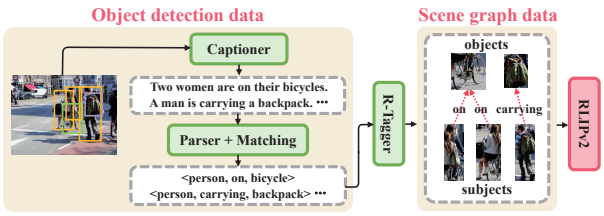

Figure 3: **An overview of relational pseudo-labeling**, which tags object detection data with free-form relation texts. This process enables pre-training to be performed on two kinds of data.

relation texts, we need to sort out the relations contained in the image as detailed in Sec. 4.2.1, and to tag relation texts to region pairs as detailed in Sec. 4.2.2.

4.2.1 Relation Candidate Set Generation

For a given image, relation candidate set generation aims to generate a coarse-grained set of candidate subject-object (SO) region pairs $\boldsymbol{P}$ and their candidate relation texts $\boldsymbol{T}$. First, we adopt BLIP to generate N_{Cap} captions for each image. Note that when $N_{Cap} = 1$, we generate the caption via beam search, a deterministic generation method; when $N_{Cap} > 1$, we generate captions via nucleus sampling [26], a stochastic generation methods with cumulative probability threshold set to 0.9. Nucleus sampling adds semantic diversity to the generated captions, which contributes to diversity in the relations.

Second, we adopt a scene graph parser [68] to parse the obtained captions into relation triplets. To filter out invalid parsed triplets, we perform string matching and keep those whose subjects and objects can be matched with any entities' names or entities' synonyms within the image [86]. This operation encloses a small set of possible SO region pairs $\boldsymbol{P}$ and possible relation texts $\boldsymbol{T}$ for the pairs (which are **inputs to R-Tagger**), without needing to traverse all possible pairs.

4.2.2 Relation Tagger via RLIPv2 Architecture

The R-Tagger aims to assign candidate relation texts $\boldsymbol{T}$ to candidate SO region pairs $\boldsymbol{P}$ for a given image. While one alternative is executing the pre-trained RLIPv2 model to perform SGG on object detection datasets to obtain pseudo-triplets, the informative object annotations fail to be utilized directly during the execution of RLIPv2. As a result, the quality of pseudo-labels is degraded. A prior work [100] adopts a coarse rule-based pseudo-labelling method that employed a greedy matching algorithm to randomly assign one relation text to an SO region pair as long as the relation texts' corresponding SO texts match with the SO region pair. An "overlap" prior was applied to further filter the triplets (*i.e.*, a triplet is deemed valid only if the subject and object are overlapped). This method, however, introduces excessive false positives, causing degraded pre-training. Thus, we propose to **reuse the RLIPv2 architecture to perform relation prediction given ground-truth SO region pairs** as shown in Figure Fig. 2(b).

As outlined in Sec. 3, RLIPv2 also utilizes subject queries $\boldsymbol{Q}_s$ and object queries $\boldsymbol{Q}_o$ as input. To allow for the utilization of object annotations as the input, we propose to replace the detection queries $\boldsymbol{Q}_s$ and $\boldsymbol{Q}_o$ with object embeddings [40], which aims to encourage *Parallel Entity Detection* to reconstruct object representations and *Sequential Relation Inference* to recognize relations. Compared with Eq. (1), the probabilistic factorisation of R-Tagger during inference can be reformulated as:

$$\begin{aligned}&\mathbb{P}(\boldsymbol{R}|\hat{\boldsymbol{B}}_s, \hat{\boldsymbol{B}}_o, \boldsymbol{C}_E; \boldsymbol{\theta}_{Par}, \boldsymbol{\theta}_{Se}) = \\ &\mathbb{P}(\tilde{\boldsymbol{B}}_s, \tilde{\boldsymbol{B}}_o|\hat{\boldsymbol{B}}_s, \hat{\boldsymbol{B}}_o, \boldsymbol{C}_E; \boldsymbol{\theta}_{Par}) \cdot \mathbb{P}(\boldsymbol{R}|\tilde{\boldsymbol{B}}_s, \tilde{\boldsymbol{B}}_o, \boldsymbol{C}_E; \boldsymbol{\theta}_{Se})\end{aligned} \tag{5}$$

where $\hat{\boldsymbol{B}}_s, \hat{\boldsymbol{B}}_o$ denote the ground-truth SO boxes from the region pair set $\boldsymbol{P}$ that include their positions and labels); $\tilde{\boldsymbol{B}}_s, \tilde{\boldsymbol{B}}_o$ denote contextualized SO representations. To allow for decoding, we embed $\hat{\boldsymbol{B}}_s, \hat{\boldsymbol{B}}_o$ to have an equivalent dimension with queries. Specifically, we use MLPs to project positions and label text embeddings, and concatenate them along the channel dimension to obtain query-like input.

Denoising training of R-Tagger. The training losses of R-Tagger are identical to RLIPv1:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{l1} + \lambda_2 \mathcal{L}_{GIoU} + \lambda_3 (\mathcal{L}_s + \mathcal{L}_o) + \lambda_4 \mathcal{L}_r \tag{6}$$

where $\mathcal{L}$ is comprised of the ℓ_1 loss for box regression $\mathcal{L}_{l1}$, GIoU loss [67] $\mathcal{L}_{GIoU}$, Cross-Entropy (CE) loss for subject and object classes $\mathcal{L}_s, \mathcal{L}_o$, and Focal loss [52] for relations $\mathcal{L}_r$. $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are set to $2.5, 1, 1, 1$ as fixed weights to balance multi-task training [86, 89, 71].

However, due to the identical input and training objectives of *Parallel Entity Detection*, the model will attempt to find a shortcut, *i.e.*, identity mapping, to achieve the minimum loss value. To avoid this, we draw inspiration from [40] which adds noise to $\hat{\boldsymbol{B}}_s, \hat{\boldsymbol{B}}_o$ during training. Specifically, we follow [40] to add center shifting and box scaling noise to box positions and add label flipping noise to box labels. The noise scale of center shifting and box scaling is set to 0.4, and the noise scale of label flipping is set to 0.2 following [40]. Furthermore, to prevent the information leakage between the same region with different noise, we apply attention masks to block the information flow between the same regions in *Parallel Entity Detection*.

Inference of R-Tagger. After the training of R-Tagger, we can use it to infer relations without additional noise based on Eq. (5). For each inference, the maximum number of region pairs is N_Q. If candidate SO region pairs exceed N_Q, we infer multiple times and merge the results. To quantify the confidence of a relation, we calculate the product of the top-1 score from the $\mathrm{softmax}$ distribution over the subject, the object and the original relation $\mathrm{sigmoid}$ score. To

N_v	N_{ALIF}	Rare	Non-Rare	Full	#Params	FPS
1	6	10.92	13.99	13.28	246.8M	18.93
2	3	**12.12**	**14.07**	**13.62**	206.6M	21.49
3	2	10.57	14.05	13.25	193.2M	22.15
6	1	11.26	13.90	13.30	179.8M	23.17

Table 1: **The effect of the sparsification of language encoding layers in *Cross-Modal Fusion***. Results are evaluated on HICO-DET under zero-shot (NF) setting. FPS (frames per second) is tested on a single NVIDIA A100 with a minibatch size 1.

Gating	tanh()	Rare	Non-Rare	Full
α	✓	10.47	13.83	13.05
$\boldsymbol{a}$	✓	9.91	13.54	12.70
SE	✓	11.07	13.89	13.24
α	✗	**12.12**	**14.07**	**13.62**
$\boldsymbol{a}$	✗	10.98	**14.07**	13.36
SE	✗	11.07	14.00	13.32

Table 2: **Comparisons of different gating functions** introduced in Sec. 4.1. We report zero-shot (NF) results on HICO-DET.

select pseudo-labels, we choose those whose relation confidence exceeds a threshold η. η is set to 0.2 by default (details can be found in the Appendix).

4.3. Pre-training, Fine-tuning and Inference

Regarding pre-training and fine-tuning of RLIPv2, we first merge results from *Parallel Entity Detection* and *Sequential Relation Inference* to obtain N_Q triplets. Next, we employ the bipartite matching algorithm originally proposed in [71] to match the predicted and ground-truth triplet annotations. The overall loss is identical to Eq. (6). The techniques introduced in RLIPv1, *i.e.* Label Sequence Extension, Relation Quality Labels and Relation Pseudo-Labels, are employed by default for a fair comparison. Regarding the inference of RLIPv2, we sort the relation confidence (defined in Sec. 4.2.2) of the correctly localised triplets (IoU $>$ 0.5) and select the Top-K triplets. K is set to 100 by default following [86, 71, 45, 11].

5. Experiments

Pre-training datasets. To pre-train RLIPv2, we utilize VG [38], COCO [53] and Objects365 [69]. VG has 108k images annotated with free-form relation and object annotations. COCO has 117k images with only object annotations in 80 classes, and Objects365 has 1,742k images with only object annotations in 365 classes. Thus, we use relational pseudo-labelling to tag relation labels for COCO and Objects365, enabling them to support relational pre-training.

Downstream datasets. For **HOI detection**, we follow [71, 86, 89, 93] to evaluate on HICO-DET [4] and V-COCO [22]. For HICO-DET that contains 117 verbs and 80 objects, we evaluate under the **Default** setting on *Full*, *Rare* and *Non-Rare* sets. For V-COCO that contains 24 interactions and 80 objects, we evaluate following the official evaluation code [22] under two scenarios: $\text{AP}^{\#1}_{role}$ and $\text{AP}^{\#2}_{role}$. For **SGG**, we assess RLIPv2 on the widely-used Open Images v6 [39] dataset, which is annotated with 288 objects and 30 relations. We evaluate RLIPv2 using standard evaluation metrics [39, 96]: Recall@50 (R@50), weighted mean Average Precision for relation detection (wmAP_{rel}) and phrase detection (wmAP_{phr}). The final score is calculated as: $\text{score}_{wtd} = 0.2 * \text{R@50} + 0.4 * \text{wmAP}_{rel} + 0.4 * \text{wmAP}_{phr}$.

Implementation details. To assess the effectiveness of RLIPv2, we choose to adopt DDETR [102] to compose **RLIPv2-ParSeD** and adopt DAB-DDETR [56] to compose **RLIPv2-ParSeDA**. To peform a fair comparison with previous works, we ensure that RLIPv2 has $N_v * N_{ALIF} = 6$ DDETR/DAB-DDETR encoding layers. For *Parallel Entity Detection* and *Sequential Relation Inference*, 3 layers are adopted following [86, 89, 71, 93]. N_Q is set to 100 during pre-training and fine-tuning, except when we fine-tune on HICO-DET where N_Q is set to 64 following [93]. Regarding model initialization, we use COCO detection parameters as initialisation when using VG or VG and COCO for pre-training; when using VG, COCO and Objects365 for pre-training, we use COCO and Object365 detection parameters as initialisation. Regarding the configuration of minibatch sizes and learning rate (LR), we set the minibatch size to 64, LR for the text encoder to 1.41e-5 and LR for other modules to 1.41e-4 when using ResNet-50 [25] and Swin-T [59]; We set the minibatch size to 32, LR for the text encoder to 1e-5 and LR for other modules to 1e-4 when using Swin-L [59]. RLIPv2 and R-Tagger are pre-trained and fine-tuned for 20 epochs unless otherwise stated, with LR dropping by a factor of 10 at the 15th epoch. For the BLIP captioner, we adopt the ViT-L/16 [14] model fine-tuned on COCO Caption [8].

5.1. Ablation Study

We perform the ablation study using RLIPv2-ParSeD with ResNet-50 as the backbone, and VG as the pre-training dataset unless otherwise stated. We pre-train for 20 epochs and evaluate performance on HICO-DET under the zero-shot with no fine-tuning (NF) setting.

5.1.1 Asymmetric Language-Image Fusion

Sparsification of the language layers. First of all, we ablate on the number of language layers to investigate the effect of sparsifying language encoding layers in *Cross-Modal Fusion*. As introduced in Sec. 4.1, dense language layers cause difficult training. To avoid training collapse, we compute classification losses ($\mathcal{L}_s$, $\mathcal{L}_o$ and $\mathcal{L}_r$) using all

Model	Pre-training Ep.	Time	Zero-shot (NF)	Fully-finetuning Ep.	Result	FPS
RLIPv1-ParSeD	50	25.9h	11.20 / **14.73** / **13.92**	60	24.67 / 32.50 / 30.70	21.89
RLIPv2-ParSeD	20	10.9h	**12.12** / 14.07 / 13.62	20	**26.47** / **33.51** / **31.89**	21.49
RLIPv1-ParSeDA	50	27.2h	11.34 / 14.56 / 13.82	60	22.85 / 30.87 / 29.03	19.41
RLIPv2-ParSeDA	20	11.3h	**13.03** / **14.98** / **14.53**	20	**27.01** / **35.21** / **33.32**	18.94

Table 3: **Comparisons of RLIPv1 and RLIPv2** under zero-shot (NF) and fully-finetuning settings on HICO-DET pre-trained on VG. Results are reported on *Rare*/*Non-Rare*/*Full* sets. FPS is tested on a single NVIDIA A100 with minibatch size 1. Pre-training time is tested on 8 NVIDIA A100. "Ep." denotes number of epochs.

Initialization	Rare	Non-Rare	Full
COCO (default)	12.12	14.07	13.62
R-Tagger	12.55	13.64	13.39
w/o noise	9.94	12.74	12.09
w/o attn masks	8.55	11.22	10.61

Table 4: **The quality of R-Tagger parameters.** Results are evaluated using RLIPv2-ParSeD with ResNet-50 under zero-shot (NF) setting.

intermediate language features rather than using only the last layer of language features in *Cross-Modal Fusion*. Results are shown in Tab. 1. Our findings suggest that sparsifying the language encoder does not compromise zero-shot performance and improves parameter efficiency. We attribute this to overfitting effects on upstream datasets impairing models' generalization. We choose $N_v = 2$ and $N_{ALIF} = 3$ as the default hyper-parameters for the following experiments.

The choice of the gating function $\mathrm{G}(\boldsymbol{x})$. To investigate the effect of the gating function on cross-modal fusion, we experiment with variants of the gating function as described in Sec. 4.1. The results are shown in Tab. 2. We note that $\tanh$ consistently degrades performance for the three gating methods. We conjecture that the output range of $\tanh$ can constrain the magnitude of the features to be fused, thus limiting the performance. Based on the results, we choose the simplest one: the learnable scalar gating method parameterised by α.

Comparisons of RLIPv2 with RLIPv1. To compare the architectural benefit of RLIPv2 with RLIPv1, we adopt DDETR [102] and DAB-DDETR [56] to follow two designs, and evaluate performance under zero-shot (NF) and fully-finetuning setting. Results are shown in Tab. 3. From this table, we can observe that by modifying the architecture without compromising much inference speed, RLIPv2 generally outperforms its RLIPv1 counterpart. Specifically, RLIPv2 obtains comparable zero-shot results to RLIPv1, while costing about 0.4× pre-training time. In terms of fine-tuning results, RLIPv2 surpasses RLIPv1 with 0.33× fine-tuning time due to earlier and deeper fusion.

5.1.2 Relational Pseudo-labelling

By default, we use RLIPv2-ParSeD with ResNet-50 backbone as the basic structure of R-Tagger.

The necessity of denoising pre-training and attention masks for R-Tagger. If noise is not added during R-Tagger's pre-training, the loss will fluctuate instead of steadily decreasing as the optimization proceeds. If attention masks are not utilized to prevent information leakage, the model will tend to learn an identity mapping, thus degrading its ability to infer relations. R-Tagger is pre-trained for 20 epochs. To assess the quality of R-Tagger's parameters, thanks to R-Tagger's identical structure to RLIPv2-ParSeD, we initialize RLIPv2-ParSeD with R-Tagger's parameters and pre-train for 10 epochs. We evaluate the zero-shot (NF) performance on HICO-DET as shown in Tab. 4. We observe that removing additional noise or attention masks both impair performance, highlighting their importance.

Tagging Method	Overlap	Rare	Non-Rare	Full
Greedy [100]	✗	11.15	11.65	11.55
	✓	13.16	14.70	14.35
CLIP (ViT-L/14) [65]	✗	12.66	12.76	12.74
	✓	14.63	14.94	14.87
R-Tagger (ResNet-50)	✗	15.33	**15.54**	**15.49**
	✓	**15.36**	15.37	15.36

Table 5: **Comparisons of relation tagging methods.** "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. We use oracle captions from **COCO Caption** [8] ($N_{Cap} = 5$).

Caption type	N_{Cap}	Rare	Non-Rare	Full
Oracle	5	15.33	15.54	15.49
BLIP (beam)	1	9.86	12.02	11.52
BLIP (nucleus)	5	14.67	14.76	14.74
BLIP (nucleus)	10	**15.08**	**15.10**	**15.09**
BLIP (nucleus)	20	14.24	14.91	14.75
BLIP (nucleus)*	5	12.31	14.37	13.89

Table 6: **Comparisons of different caption types.** "beam" and "nucleus" denote beam search and nucleus sampling. "Oracle" denotes captions from COCO Caption. By default, we adopt COCO Caption fine-tuned BLIP model. * denotes that we adopt the pre-trained BLIP model.

Comparisons of different relation tagging strategies. We compare R-Tagger with other two methods: **(i)** the greedy matching algorithm [100]; **(ii)** the CLIP [65] tagging method. Specifically, to tag relations for a given SO region pair with a candidate relation text, the CLIP tagging

Dataset	Relation candidate	Rare	Non-Rare	Full
VG	-	12.12	14.07	13.62
VG+COCO	BLIP	**15.08**	**15.10**	**15.09**
VG+COCO	Selection from VG	10.34	11.33	11.11

Table 7: **Comparisons of methods to generate relation candidate sets.** We report zero-shot (NF) results on HICO-DET.

	VG	VG+COCO	VG+COCO+O365
ResNet-50	13.03 / 14.98 / 14.53	15.00 / 16.60 / 16.23	19.64 / 17.24 / 17.79
Swin-T	13.01 / 16.06 / 15.35	17.13 / 18.74 / 18.37	21.24 / 19.47 / 19.87
Swin-L	19.93 / 18.74 / 19.02	22.59 / 21.09 / 21.44	27.97 / 21.90 / 23.29

Table 8: **Model and dataset scaling experiments using RLIPv2-ParSeDA.** Results are evaluated on HICO-DET *Rare*/*Non-Rare*/*Full* sets under zero-shot (NF) setting.

Model	Backbone	Extra	mR@50	R@50	wmAP rel	wmAP phr	score_{wtd}
VCTree [73]	X101-F	-	33.91	74.08	34.16	33.11	40.21
G-RCNN [82]	X101-F	-	34.04	74.51	33.51	34.21	41.84
Motifs [90]	X101-F	-	32.68	71.63	29.91	31.59	38.93
Unbiased [72]	X101-F	-	35.47	69.30	30.74	32.80	39.27
GPS-Net [54]	X101-F	-	35.26	74.81	32.85	33.98	41.69
RelDN [96]	R101	-	36.80	72.75	29.87	30.42	38.67
BGNN [46]	R101	-	39.41	74.93	31.15	31.37	40.00
SGTR [45]	R101	-	42.61	59.91	36.98	38.73	42.28
RelTR [11]	R50	-	-	64.47	34.17	37.44	41.54
RLIPv2-ParSeD	R50*	-	44.58	58.04	43.30	43.12	46.18
	R50*	-	44.88	60.20	44.73	43.36	47.28
	R50†	-	45.59	61.15	45.71	43.73	48.01
RLIPv2-ParSeDA	R50	**(i)**	50.42	63.35	47.65	45.23	49.82
	R50	**(ii)**	**52.07**	64.53	49.14	**46.14**	51.01
	R50	**(iii)**	51.31	**65.99**	**49.54**	45.71	**51.30**
	Swin-T	**(iii)**	**59.61**	**68.81**	**52.70**	**48.01**	**54.05**
	Swin-L	**(iii)**	**64.72**	**72.49**	**56.38**	**50.70**	**57.34**

Table 9: **Comparisons with previous methods on Open Images v6 SGG benchmark.** X101-F denote ResNeXt-101 FPN [80]. * and † denote ImageNet pretrained and COCO object detection pretrained. "**Extra**" denotes extra relations adopted from **(i)** VG, **(ii)** VG+COCO and **(iii)** VG+COCO+O365.

method clips the minimum bounding box of the SO region pair, creates two prompts ("a photo of {subject} {relation} {object}" and "a photo of {subject} not interacting with {object}" are adopted as positive and negative prompts.), and performs zero-shot prediction. If the softmax probability of the positive prompt is greater than a pre-defined threshold, we tag this relation text to this SO region pair. (We traverse the threshold and find the optimal one for the CLIP tagging method, as detailed in the Appendix.) We also ablate on the "overlap" prior [100] to observe whether a given tagging method relies on strong prior knowledge to filter out false positive relations. The results are shown in Tab. 5. From this table, we conclude that greedy matching and CLIP tagging method generate a significant number of low-quality non-overlapped triplets. Thus, the "overlap" prior is essential for them. R-Tagger, however, suffers a slight performance drop when using the "overlap" prior, indicating that R-Tagger generates more reliable non-overlapped triplets. Besides, although CLIP is pre-trained on a massive quantity of language-image pairs, it struggles in recognizing relations, which R-Tagger is expert in.

Method	Backbone	UC-RF	UC-NF
VCL [27]	ResNet-50	10.06 / 24.28 / 21.43	16.22 / 18.52 / 18.06
ATL [28]	ResNet-50	9.18 / 24.67 / 21.57	18.25 / 18.78 / 18.67
FCL [29]	ResNet-50	13.16 / 24.23 / 22.01	18.66 / 19.55 / 19.37
GEN-VLKT [51]	ResNet-50	21,36 / 32.91 / 30.56	**25.05** / 23.38 / 23.71
RLIPv1-ParSeD [86]	ResNet-50	16.43 / 30.59 / 27.76	16.99 / 24.71 / 22.93
RLIPv1-ParSe [86]	ResNet-50	19.19 / 33.35 / 30.52	20.27 / 27.67 / 26.19
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 10: **Comparisons with methods on HICO-DET under UC-RF and UC-NF settings.** We adopt ResNet-50 as the backbone. Results are reported on *Unseen*/*Seen*/*Full* sets.

Method	Backbone	1% Data	10% Data
RLIPv1-ParSeD [86]	ResNet-50	16.22 / 18.92 / 18.30	15.89 / 23.94 / 22.09
RLIPv1-ParSe [86]	ResNet-50	17.47 / 18.76 / 18.46	20.16 / 23.32 / 22.59
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 11: **Comparisons on HICO-DET under few-shot settings.** Results are reported on *Rare*/*Non-Rare*/*Full* sets.

Comparisons of different caption types. Previous experiments demonstrate that oracle COCO captions can benefit pre-training. However, most datasets lack such high-quality captions. Thus, BLIP provides an alternative for caption generation. We conduct experiments on various caption types, as shown in Tab. 6. From this table, we can observe that: **(i)** BLIP-generated captions achieve a decent performance compared to oracle captions, indicating the practicality of adopting generated captions for pseudo-labelling. **(ii)** If we compare BLIP model with different parameters (COCO Caption fine-tuned or only pre-trained), we can see that the fine-tuned model generates better captions with more boost on the *Rare* set (12.31 $\rightarrow$ 14.67). We conjecture that the caption quality of the curated style dataset COCO Caption is better than pre-training captions harvested from the web. **(iii)** By adopting generated captions, we can increase the number of captions per image, thus diversifying the relations contained in captions and boost RLIPv2 (11.52 $\rightarrow$ 15.09). Besides, although datasets like Conceptual Caption [70] (CC3M) also provide captions, the quantity of captions (average one caption per image) is not enough to describe many relations in an image. In comparison, our pipeline can work without caption annotation while performing better.

The necessity of using captioners. By default, we use BLIP to generate relation texts. Another option is to query all possible SO pairs and select all possible relation texts from VG as it contains an enormous quantity of relations

Model	Backbone	Extra Relations	HICO-DET		V-COCO	
			Zero-shot (NF)	Fully-finetuning	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
InteractNet [21]	R50-FPN	-	-	7.16 / 10.77 / 9.94	40.0	-
UnionDet [35]	R50-FPN	-	-	11.72 / 19.33 / 17.58	47.5	56.2
PPDM [50]	HG104	-	-	13.97 / 24.32 / 21.94	-	-
HOTR [36]	R50	-	-	17.34 / 27.42 / 25.10	55.2	64.4
QPIC [71]	R50	-	-	21.85 / 31.23 / 29.07	58.8	61.0
OCN [89]	R50	-	-	25.56 / 32.51 / 30.91	64.2	66.3
CDN [93]	R50	-	-	27.39 / 32.64 / 31.44	61.7	63.8
GEN-VLKT [51]	R50	-	-	29.25 / 35.10 / 33.75	62.4	64.5
QAHOI [5]	Swin-L*	-	-	29.80 / 37.56 / 35.78	-	-
UniVRD [98]	ViT-H/14†	-	-	31.65 / 39.99 / 38.07	65.8	66.9
RLIPv1-ParSeD [86]	R50	VG	11.20 / 14.73 / 13.92	24.67 / 32.50 / 30.70	61.7	63.8
RLIPv1-ParSe [86]	R50	VG	15.08 / 15.50 / 15.40	26.85 / 34.63 / 32.84	61.9	64.2
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 12: **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 [25] and Hourglass [61]. * denotes the backbone is pre-trained with 384×384 resolution, while others use 224×224. † indicates the backbone is pre-trained using LiT [92], then fine-tuned on Objects365, COCO and HICO with the objective of object detection.

(36,515 kinds). Then, we run R-Tagger with selected relation texts as $\boldsymbol{T}$ and all region pairs as $\boldsymbol{P}$. The results are shown in Tab. 7. We can observe that selecting from VG generates low-quality candidates, harming performance.

Model scaling and dataset scaling. Equipped with the labelling pipeline introduced above, we can scale RLIPv2 to larger models and datasets. In Tab. 8, we adopt RLIPv2-ParSeDA as the base architecture and observe the benefit of scaling by zero-shot (NF) performance on HICO-DET. In terms of data, adding COCO and Objects365 can both boost performance, and the benefit of adding data exhibits a log scaling trend [9]. Models pre-trained with Objects365 consistently have better *Rare* result, which we attribute to the distribution misalignment of Objects365 and HICO-DET [17]. In terms of models, switching to stronger backbone models can improve the data efficiency at the cost of larger amounts of computation. Regarding scaling experiments using RLIPv2-ParSeD, we present it in the Appendix.

5.2. Comparisons with State-of-the-Arts

Scene graph generation. We compare RLIPv2 series models with previous methods on Open Images v6 in Tab. 9. We also report mean Recall@50 (mR@50) to better show the usefulness of RLIPv2. Our findings suggest that **(i)** with the assistance of DDETR family models, RLIPv2 can serve as a strong baseline without any pre-training; **(ii)** naive object detection pre-training can boost the performance to some extent ($47.28 \rightarrow 48.01$), while pre-training on VG can further boost the performance especially on mR@50 ($45.59 \rightarrow 50.42$); **(iii)** adding pseudo-labelled relation annotations in pre-training or switching to stronger backbones both contribute to better performance. However, the boost of adding Objects365 is negligible. We attribute this to the distribution discrepancy of Objects365 and Open Images v6.

HOI Detection under UC-NF and UC-RF settings. We report results in Tab. 10 on unseen combinations (UC) under UC-RF and UC-NF settings following [27, 28]. We only fine-tune for 10 epochs under UC-NF setting. Our method outperforms previous methods except on one metric. We attribute this to the strong transferability of CLIP features that GEN-VLKT adopts. Regarding experiments using RLIPv2-ParSeD, we present it in the Appendix.

Few-shot HOI Detection. We follow [86] to only fine-tune 10 epochs on partial data (1% and 10%), results of which are shown in Tab. 11. We can observe significant improvements upon all metrics by scaling up pre-training compared with previous methods. This improves the practicality of RLIPv2 in low-data scenarios. Regarding experiments using RLIPv2-ParSeD, we present it in the Appendix.

HOI detection under fully-finetuning and zero-shot (NF) settings. We compare the performance of RLIPv2 series models with previous methods on HICO-DET and V-COCO in Tab. 12. We can observe from the 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. Regarding experiments using RLIPv2-ParSeD, we present it in the Appendix.

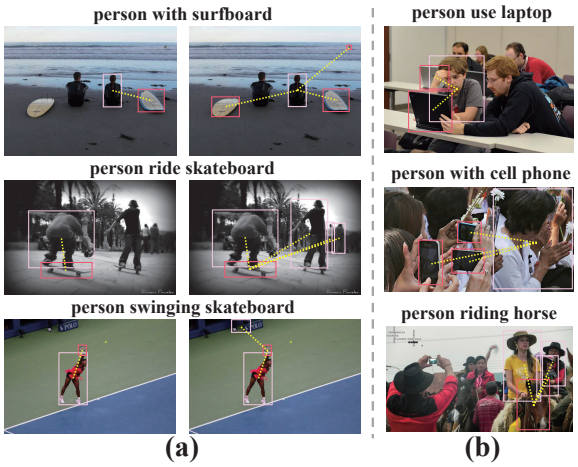

Figure 4: **(a) Visualization of pseudo-labelled relations on COCO**. Left column: R-Tagger; right column: CLIP tagging method. **(b) Visualization of failure cases of R-Tagger.**

5.3. Qualitative Analysis

Comparisons of relation tagging methods. We visualize three examples to compare the quality of pseudo-labelled relations by R-Tagger and CLIP tagging method in Fig. 4(a). Generally, CLIP is more object-centric and position-agnostic, and thus struggles in discriminating relations. It tends to tag relations as long as the subject and object have strong co-occurrence priors. However, R-Tagger tags more reasonable relations.

Failure cases of R-Tagger. Recognizing relations is challenging, especially in complex scenes. In Fig. 4(b), we present three examples of R-Tagger's failure cases. In particular, we observe failure cases when the scene contains multiple similar subjects or objects.

6. Conclusion

In this paper, we propose RLIPv2, a fast converging model that enables the scaling of relational pre-training to larger-scale pseudo-labelled datasets. Comprehensive experiments on HOI detection and scene graph generation under various settings demonstrate its effectiveness compared to previous methods. We anticipate that our work can galvanize further research efforts to focus on relational reasoning, fostering advancements that yield tangible benefits not only to the research community but also to broader society and humanity.

Acknowledgments

We would like to extend our sincere gratitude to the anonymous reviewers for their invaluable feedback. Additionally, we appreciate the Fundamental Vision Intelligence Team of Alibaba DAMO Academy for their generous provision of essential computational resources. This research received support from the National Natural Science Foundation of China under Grant No. 62173298 and was additionally backed by Alibaba Group via the Alibaba Research Intern Program.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 2, 3

[2] Zhaowei Cai, Gukyeong Kwon, Avinash Ravichandran, Erhan Bas, Zhuowen Tu, Rahul Bhotika, and Stefano Soatto. X-detr: A versatile architecture for instance-wise vision-language tasks. In *ECCV*, pages 290–308. Springer, 2022. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020. 1, 2

[4] 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. 5

[5] Junwen Chen and Keiji Yanai. Qahoi: Query-based anchors for human-object interaction detection. *arXiv preprint arXiv:2112.08647*, 2021. 8

[6] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *CVPR*, pages 9004–9013, 2021. 2

[7] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *CVPR*, pages 6163–6171, 2019. 2

[8] 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. 5, 6

[9] 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. 8

[10] Yuren Cong, Wentong Liao, Bodo Rosenhahn, and Michael Ying Yang. Learning similarity between scene graphs and images with transformers. *arXiv preprint arXiv:2304.00590*, 2023. 2

[11] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. Reltr: Relation transformer for scene graph generation. *arXiv preprint arXiv:2201.11460*, 2022. 2, 5, 7

[12] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *CVPR*, pages 7373–7382, 2021. 1

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 5

[15] 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, 3

[16] 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. 3

[17] 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. 8

[18] Azade Farshad, Yousef Yeganeh, Yu Chi, Chengzhi Shen, Björn Ommer, and Nassir Navab. Scenegenie: Scene graph guided diffusion models for image synthesis. *arXiv preprint arXiv:2304.14573*, 2023. 2

[19] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *ECCV*, pages 696–712. Springer, 2020. 2

[20] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018. 2

[21] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, pages 8359–8367, 2018. 8

[22] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 5

[23] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *ICCV*, pages 4918–4927, 2019. 1

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 2

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 5, 8

[26] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019. 4

[27] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, pages 584–600. Springer, 2020. 7, 8

[28] 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. 7, 8

[29] 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. 7

[30] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:2011–2023, 2019. 3

[31] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *CVPR*, pages 10236–10247, 2020. 2

[32] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 2

[33] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, pages 3668–3678, 2015. 2

[34] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, pages 1780–1790, 2021. 2, 3

[35] 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. 2, 8

[36] 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. 2, 8

[37] Bumsoo Kim, Jonghwan Mun, Kyoung-Woon On, Minchul Shin, Junhyun Lee, and Eun-Sol Kim. Mstr: Multi-scale transformer for end-to-end human-object interaction detection. In *CVPR*, pages 19578–19587, 2022. 2

[38] 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, 5

[39] 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. 2, 5

[40] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *CVPR*, pages 13619–13627, 2022. 4

[41] 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. 1, 2

[42] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR, 2022. 1, 2

[43] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi.

Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 34, 2021. 3

[44] 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, 3

[45] Rongjie Li, Songyang Zhang, and Xuming He. Sgtr: End-to-end scene graph generation with transformer. *ArXiv*, abs/2112.12970, 2021. 2, 5, 7

[46] Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *CVPR*, pages 11109–11119, 2021. 2, 7

[47] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: an efficient subgraph-based framework for scene graph generation. In *ECCV*, pages 335–351, 2018. 2

[48] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. Hoi analysis: Integrating and decomposing human-object interaction. *NeurIPS*, 33, 2020. 2

[49] 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 *CVPR*, pages 3585–3594, 2019. 2

[50] 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. 2, 8

[51] 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. 7, 8

[52] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 4

[53] 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. 2, 3, 5

[54] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. *arXiv preprint arXiv:2003.12962*, 2020. 2, 7

[55] Hengyue Liu, Ning Yan, Masood Mortazavi, and Bir Bhanu. Fully convolutional scene graph generation. In *CVPR*, pages 11546–11556, 2021. 2

[56] 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. *arXiv preprint arXiv:2201.12329*, 2022. 1, 2, 5, 6

[57] Xinpeng Liu, Yong-Lu Li, Xiaoqian Wu, Yu-Wing Tai, Cewu Lu, and Chi-Keung Tang. Interactiveness field in human-object interactions. In *CVPR*, pages 20113–20122, 2022. 2

[58] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 1, 3

[59] 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*, pages 10012–10022, 2021. 5

[60] Muhammad Maaz, Hanoona Rasheed, Salman Hameed Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Multi-modal transformers excel at class-agnostic object detection. *ArXiv*, abs/2111.11430, 2021. 3

[61] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016. 8

[62] Jihwan Park, SeungJun Lee, Hwan Heo, Hyeong Kyu Choi, and Hyunwoo J Kim. Consistency learning via decoding path augmentation for transformers in human object interaction detection. In *CVPR*, pages 1019–1028, 2022. 2

[63] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, September 2018. 2

[64] Xian Qu, Changxing Ding, Xingao Li, Xubin Zhong, and Dacheng Tao. Distillation using oracle queries for transformer-based human-object interaction detection. In *CVPR*, pages 19558–19567, 2022. 2

[65] 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. *arXiv preprint arXiv:2103.00020*, 2021. 1, 2, 6

[66] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. 2

[67] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, pages 658–666, 2019. 4

[68] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015. 4

[69] 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. 2, 3, 5

[70] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 7

[71] 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. 2, 4, 5, 8

[72] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, pages 3716–3725, 2020. 7

[73] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, pages 6619–6628, 2019. 7

[74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 1, 3

[75] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *WACV*, pages 1508–1517, 2020. 2

[76] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2

[77] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 1

[78] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 1

[79] Xiaoqian Wu, Yong-Lu Li, Xinpeng Liu, Junyi Zhang, Yuzhe Wu, and Cewu Lu. Mining cross-person cues for body-part interactiveness learning in hoi detection. In *ECCV*, pages 121–136. Springer, 2022. 2

[80] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500, 2017. 7

[81] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, pages 5410–5419, 2017. 2

[82] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *ECCV*, pages 670–685, 2018. 2, 7

[83] Ling Yang, Zhilin Huang, Yang Song, Shenda Hong, Guohao Li, Wentao Zhang, Bin Cui, Bernard Ghanem, and Ming-Hsuan Yang. Diffusion-based scene graph to image generation with masked contrastive pre-training. *arXiv preprint arXiv:2211.11138*, 2022. 2

[84] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *arXiv preprint arXiv:2209.09407*, 2022. 2

[85] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, pages 684–699, 2018. 2

[86] 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. 1, 2, 4, 5, 7, 8

[87] Hangjie Yuan and Dong Ni. Learning visual context for group activity recognition. In *AAAI*, volume 35, pages 3261–3269, 2021. 2

[88] Hangjie Yuan, Dong Ni, and Mang Wang. Spatio-temporal dynamic inference network for group activity recognition. In *ICCV*, pages 7476–7485, 2021. 2

[89] 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. 2, 4, 5, 8

[90] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, pages 5831–5840, 2018. 2, 7

[91] Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. X^2-vlm: All-in-one pre-trained model for vision-language tasks. *arXiv preprint arXiv:2211.12402*, 2022. 1

[92] 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. 8

[93] 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. 2, 5, 8

[94] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *CVPR*, pages 20104–20112, 2022. 2

[95] 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. 1, 2

[96] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *CVPR*, pages 11535–11543, 2019. 5, 7

[97] 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 *CVPR*, pages 19548–19557, 2022. 2

[98] 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. 8

[99] Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *CVPR*, pages 13234–13243, 2021. 2

[100] Yiwu Zhong, Jing Shi, Jianwei Yang, Chenliang Xu, and Yin Li. Learning to generate scene graph from natural language supervision. In *ICCV*, pages 1823–1834, 2021. 2, 4, 6, 7

[101] Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. Human-object interaction detection via disentangled transformer. In *CVPR*, pages 19568–19577, 2022. 2

[102] 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. 1, 2, 5, 6

[103] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *NeurIPS*, 33:3833–3845, 2020. 1

[104] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *CVPR*, pages 11825–11834, 2021. 2