# RLIP: Relational Language-Image Pre-training for Human-Object Interaction Detection

**Hangjie Yuan**[1*] **Jianwen Jiang**[2*] **Samuel Albanie**[3]
**Tao Feng**[2] **Ziyuan Huang**[4] **Dong Ni**[1†] **Mingqian Tang**[2]

[1]Zhejiang University [2]Alibaba Group [3]University of Cambridge
[4]National University of Singapore
{hj.yuan, dni}@zju.edu.cn sma71@cam.ac.uk ziyuan.huang@u.nus.edu
{jianwen.jjw, shisi.ft, mingqian.tmq}@alibaba-inc.com

## Abstract

The task of Human-Object Interaction (HOI) detection targets fine-grained visual parsing of humans interacting with their environment, enabling a broad range of applications. Prior work has demonstrated the benefits of effective architecture design and integration of relevant cues for more accurate HOI detection. However, the design of an appropriate pre-training strategy for this task remains underexplored by existing approaches. To address this gap, we propose *Relational Language-Image Pre-training* (RLIP), a strategy for contrastive pre-training that leverages both entity and relation descriptions. To make effective use of such pre-training, we make three technical contributions: (1) a new **Par**allel entity detection and **Se**quential relation inference (ParSe) architecture that enables the use of both entity and relation descriptions during holistically optimized pre-training; (2) a synthetic data generation framework, Label Sequence Extension, that expands the scale of language data available within each minibatch; (3) mechanisms to account for ambiguity, Relation Quality Labels and Relation Pseudo-Labels, to mitigate the influence of ambiguous/noisy samples in the pre-training data. Through extensive experiments, we demonstrate the benefits of these contributions, collectively termed RLIP-ParSe, for improved zero-shot, few-shot and fine-tuning HOI detection performance as well as increased robustness to learning from noisy annotations. Code will be available at `https://github.com/JacobYuan7/RLIP`.

## 1 Introduction

Driven by improvements in storage, sensors and networking technology, humanity is amassing vast archives of image and video data. A significant fraction of this media is *human-centric*—it is content focused on humans and their actions. The task of *human-object interaction (HOI) detection* [4] aims to provide a step towards fine-grained parsing of such content by detecting all possible triplets of the form <*human, relation, object*> present in visual data. Robust HOI detection has myriad uses for image/video data analysis and represents essential functionality for visual and language applications such as image/video captioning [66, 51], image retrieval [29], image synthesis [28] and video action understanding [26, 69].

Given that sustained progress in object detection has yielded increasingly robust systems for detecting people and objects [56, 17, 10], a key remaining challenge for HOI detection is to develop methods capable of generalising to the many possible pairs of interactions between these entities when provided with non-exhaustive training data. To tackle this challenge, we draw inspiration from recent

*Equal contribution. This work was done when Hangjie Yuan was an intern at DAMO Academy, Alibaba Group, supported by Alibaba Research Intern Program.

†Corresponding author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

developments demonstrating that contrastive language-image pre-training can induce remarkable generalisation for zero-shot classification tasks [55, 27]. These methods perform classification by casting it as a *retrieval problem*, ensuring that the downstream task *aligns closely* with the pre-training objective. Recent work by Alayrac et al. [2] hypothesises that it is this close alignment between the downstream and pre-training objectives that explains why contrastive methods have proven so effective for zero-shot classification. In light of this hypothesis, in this work, we explore whether it is possible to achieve a similarly close alignment between the HOI detection task and its pre-training strategy.

While HOI detection has been widely studied [16, 52, 65, 42, 31, 33, 73, 75, 32, 61, 70, 9], the topic of designing pre-training to reflect the final task objective remains under-explored. Indeed, a widely adopted strategy [70, 72, 61, 7, 75] has been to employ object detection pre-training to initialise the parameters of the model responsible for both entity detection and relation inference. However, while suitable for entity detection, such pre-training may be suboptimal for the detection of *relations between entities* which often requires the model to take account of groups of entities with greater spatial context, rather than individual entities in isolation.

To address this shortcoming of HOI detection, we propose **Relational Language-Image Pre-training** (RLIP) which tasks the model with establishing correspondences from both entities and relations to free-form text descriptions. By doing so, RLIP endows the model with the ability to perform zero-shot HOI detection[3]. Moreover, in contrast to previous pre-training schemes that are limited to predefined finite category lists, RLIP benefits from the rich descriptive nature of natural language supervision.

We encountered three barriers to a naive implementation of RLIP for existing methods: (1) Recent end-to-end HOI detection architectures [61, 72, 75, 70] typically employ joint representations of (some subset of) *subject, object* and *relation* triplets. As a consequence, it is difficult to leverage text descriptions for separate humans, objects and relations provided by existing datasets such as VG [34]. (2) Contrastive pre-training requires negative samples to train effectively, but it is unclear *a priori* how such negatives should be constructed. (3) Free-form text descriptions exhibit label noise and semantic ambiguity (since there can be many ways to describe the same concept in the absence of a canonical list of categories), rendering optimisation challenging.

To overcome these barriers, we make several technical contributions in addition to the RLIP framework. First, to allow end-to-end contrastive pre-training with distinct descriptions of subsets, objects and relations, we propose the **Par**allel entity detection and **Se**quential relation inference (ParSe) architecture. ParSe employs a DETR [3]-like design that allocates separate learnable query groups for subject and object representations, together with an additional set of conditional queries that encode relations. While ParSe enables (and works best with) RLIP, we also find that it yields gains for traditional object detection pre-training schemes. To address the second barrier, we synthesise label sequences by extending in-batch labels with out-of-batch sampling to ensure a plentiful supply of negatives—we term this Label Sequence Extension (LSE). For the third barrier, we exploit cross-modal cues to resolve label noise and relation ambiguity. In particular, to mitigate label noise we use the quality of the visual entity detection phase [39] to assign quality scores to relation-text correspondences, an approach we term Relational Quality Labels (RQL). To mitigate relation ambiguity, we leverage similarities between labels to propagate relations via a pseudo-labeling scheme, which we term Relational Pseudo-Labels (RPL).

We demonstrate through experiments that relational pre-training outperforms traditional object detection pre-training schemes on comparable data. We further find that the a zero-shot application of our combined approach, RLIP-ParSe, surpasses several existing fine-tuned methods.

## 2 Related Work

**Human-object interaction detection.** There is a rich body of work on HOI detection. One theme has focused on the development of effective architectures for this task [42, 31, 61, 9]. A second theme has sought to leverage informative cues ranging from interaction points [42, 73], interaction boxes [31], contextualized embeddings [61, 32, 75, 32], poses [16, 41] and statistical priors [33, 70] to external knowledge in the form of language embeddings [65, 52, 70]. However, relational pre-training and open-vocabulary recognition remains underexplored. The inter-pair transformations of

[3] *Zero-shot* in this context refers to HOI detection without fine-tuning (following the terminology of [55]), a formulation that assesses the generalization of a pre-training model to unseen distributions. This offers a practical alternative to the scenario (unseen combinations) considered in several prior HOI detection papers [21, 23].

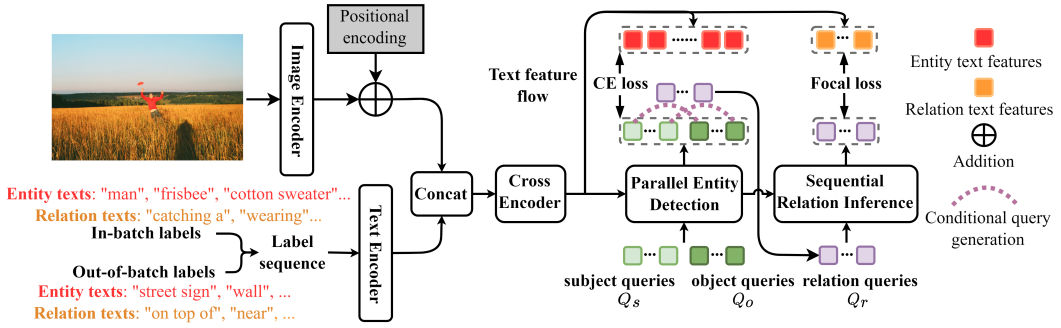


Figure 1: An overview of our pre-training framework, RLIP-ParSe. The *Parallel Entity Detection* and *Sequential Relation Inference* blocks represent independent DETR-style decoders responsible for entity and relation detection, respectively. We omit the localisation loss for clarity. See Sec. 3 for further details.

IDN [40] and affordance transfer learning of ATL [22] can be interpreted as entity augmentations to train a stronger verb classifier, but these methods do not directly optimise for all the components of HOI detection. Similar to ParSe, CDN [72] explores disentangled embeddings. However, it still couples the embedding of subjects and objects, rendering it suboptimal for RLIP. The concurrent work, GEN-VLKT [43] aims to derive knowledge from image-level language-image pre-training [55], while we aim to achieve aligned entity- and relation-level pre-training for HOI detection.

**Leveraging free-form text for visual pre-training.** A series of recent papers have illustrated the significant value of employing free-form language to provide supervision for vision systems. CLIP [55] and ALIGN [27] demonstrated striking improvements in zero-shot image classification ability through contrastive training of image-level representations. Further work has sought to additionally leverage correspondences between objects/regions and text to learn flexible grounding models [30, 38, 71]. We similarly seek to benefit from natural language supervision. However, to the best of our knowledge, we are the first to leverage correspondences between descriptions of relations and explicit pairings of *subjects and objects* (rather than descriptions of images, objects or regions) as a pre-training signal.

## 3 Methodology

In this section, we first present our triplet detection architecture, ParSe. Second, we describe how ParSe is used to perform relational language-image pre-training (RLIP). Finally, we introduce techniques to synthesise contrastive negatives and mitigate noise and ambiguities among labels. The overall RLIP-ParSe framework is illustrated in Fig. 1.

### 3.1 ParSe for Triplet Detection

**Structure overview.** The core idea underpinning the ParSe architecture is to allocate distinct representations of subjects, objects and relations in a holistically optimized model (rather than representing their combination, as commonly pursued in prior work [61]). The motivation for doing so is two-fold: (i) distinct representations enable the direct use of contrastive RLIP, since these representations can be put in correspondence with separate entity and relation annotations; (ii) the separation of responsibilities allows for a more fine-grained control over the context available for each decision (a theme that has proven important for detection tasks [59]). In particular, note that when detecting subjects and objects, local context is typically most useful. However, when it comes to relations, detection will benefit not only from informative local cues, but also neighbouring context [68] (for instance, it is useful to be aware of *water* and *hoses* when inferring the relation in the triplet <*human, wash, car*>). To instantiate this idea we follow [72] and implement triplet detection in a two-stage end-to-end manner. Our probabilistic model factorises as follows:

$$\mathbb{P}(\boldsymbol{G}|\boldsymbol{Q}_s, \boldsymbol{Q}_o, \boldsymbol{C}; \boldsymbol{\theta}_{Par}, \boldsymbol{\theta}_{Se}) = \mathbb{P}(\boldsymbol{B}_s, \boldsymbol{B}_o|\boldsymbol{Q}_s, \boldsymbol{Q}_o, \boldsymbol{C}; \boldsymbol{\theta}_{Par}) \cdot \mathbb{P}(\boldsymbol{R}|\boldsymbol{B}_s, \boldsymbol{B}_o, \boldsymbol{C}; \boldsymbol{\theta}_{Se}) \quad (1)$$

where $\boldsymbol{Q}_s, \boldsymbol{Q}_o \in \mathbb{R}^{N_Q \times D}$ define two sets of independent queries for $N_Q$ subjects and $N_Q$ objects; $\boldsymbol{C}$ denotes features from the detection encoder; $\boldsymbol{B}_s, \boldsymbol{B}_o, \boldsymbol{R}$ denote sets of detected subject boxes, object boxes and relations, respectively (these collectively comprise the detection results $\boldsymbol{G}$); $\boldsymbol{\theta}_{Par}$ and $\boldsymbol{\theta}_{Se}$ represent learnable parameters from the entity detection decoder and the relation inferring decoder, respectively. To construct ParSe, we design two components, *Parallel Entity Detection* and *Sequential Relation Inference*, to implement the second and third terms in Eq. (1), respectively. These are described next.

**Parallel Entity Detection.** Following the DETR family of architectures [3, 74, 45], we first extract visual features using an image encoder, add positional encodings and then pass the result through a customized Transformer encoder according to the detector we adopt (we explore both DETR [3] and DDETR [74] variants) to obtain detection features $\boldsymbol{C}$. Then, two sets of queries $\boldsymbol{Q}_s$ and $\boldsymbol{Q}_o$ are fed into the entity decoder to perform self-attention [62], cross-attention and feed-forward network (FFN) inference, obtaining $\tilde{\boldsymbol{Q}}_s, \tilde{\boldsymbol{Q}}_o \in \mathbb{R}^{N_Q \times D}$ which are used to predict box locations and classes.
**Sequential Relation Inference.** To encode relations, we perform *Sequential Relation Inference* as a sequential step after entity detection (similarly to [72]). In the first stage, subjects and objects are detected via *Parallel Entity Detection*. In the second stage, we adopt a simple parameter-free matching scheme between subjects and objects to generate relation queries: matching by their indices. Using this pairing scheme, we obtain relation queries via a conditional query generation function:

$$\boldsymbol{Q}_r = F_{so}(\tilde{\boldsymbol{Q}}_s, \tilde{\boldsymbol{Q}}_o) \tag{2}$$

where for simplicity, we adopt addition as the query generation function. Since we match by indices, $\boldsymbol{Q}_r \in \mathbb{R}^{N_Q \times D}$ contains $N_Q$ relation queries. $\boldsymbol{Q}_r$ is then fed into the second decoder to perform *Sequential Relation Inference* via self-attention, cross-attention and FFN inference to obtain the corresponding relation features $\tilde{\boldsymbol{Q}}_r \in \mathbb{R}^{N_Q \times D}$ which are then used for relation classification.

### 3.2 RLIP-ParSe for Relational Language-Image Pre-training

For each iteration of pre-training, we construct a minibatch of images and their annotated relation triplets comprising all entities' locations, $N_E$ unique entity text labels as well as $N_R$ unique relation text labels. We describe how these are used for contrastive pre-training next.
**Formation of target label sequences.** We construct targets from in-batch labels (free-form text descriptions forming subject, object, relation triplets). In more detail, we first aggregate all entity labels within the batch and append to this sequence a *no objects* label. Next, we similarly aggregate all in-batch relation labels. Then, all entity and relation labels are respectively fed into a text encoder (RoBERTa [46] in our implementation) to extract label features denoted as $\boldsymbol{L}_E$ and $\boldsymbol{L}_R$, respectively. Note that a free-form text label can have multiple tokens after tokenization—we use only the feature derived from the [CLS] token to represent the label. We concatenate the label feature sequence with features from the image encoder as shown in Fig. 1. To fuse the concatenated features, we adopt a simple approach: applying a Transformer encoder [36, 30, 1] to obtain fused label features $\tilde{\boldsymbol{L}}_E \in \mathbb{R}^{N_E \times D}$ and $\tilde{\boldsymbol{L}}_R \in \mathbb{R}^{N_R \times D}$.
**Cross-modal alignment through classification.** To implement RLIP, we task the model with establishing correspondences between entities/relations and their text descriptions using a classification objective, following [38]. In particular, we align the $i$th relation $\tilde{\boldsymbol{Q}}_r(i) \in \mathbb{R}^D$ with relation its text via a Focal loss [44]:

$$\boldsymbol{P}_r(i) = \tilde{\boldsymbol{Q}}_r(i)\tilde{\boldsymbol{L}}_R^T + \tilde{\boldsymbol{Q}}_r(i)\boldsymbol{W}_b^T + \boldsymbol{W}_c; \quad \mathcal{L}_r(j) = \text{Focal}(\text{sigmoid}(\boldsymbol{P}_r(i,j))) \tag{3}$$

where $\tilde{\boldsymbol{Q}}_r(i)\boldsymbol{W}_b^T + \boldsymbol{W}_c$ is the learnable bias term introduced in [44]; $\boldsymbol{W}_b \in \mathbb{R}^{N_R \times D}$ is a learnable linear projection and $\boldsymbol{W}_c \in \mathbb{R}^{N_R}$ is a constant vector filled with $-\log((1-\pi)/\pi)$ with $\pi = .01$. The Focal loss is defined via $\text{Focal}(p) = -(1-p)^\gamma \log(p)$ where $\gamma$ is set as a hyperparameter. In the argument to this loss in Eq. (3), $j$ indexes along $\boldsymbol{P}_r(i) \in \mathbb{R}^{N_R}$. To encourage matching of subjects and objects with their corresponding entity descriptions, an analogous objective to Eq. (3) is used except that a softmax and a CE loss are applied and $\boldsymbol{W}_c$ is omitted (note that entities are uni-label and relations multi-label, as defined by the downstream task). The central benefit of the RLIP objectives defined above is that they bring the pre-training and downstream HOI detection losses into close alignment since the task of classifying entities and relations in the downstream task reflects the same matching task used in pre-training. As a result, RLIP produces models that can perform HOI detection under *zero-shot with no fine-tuning* (NF) evaluation protocols.
**Label Sequence Extension (LSE).** Within a given batch, the number of negative samples available for matching is limited. However, provision of plentiful negatives has been widely shown to improve contrastive learning [8, 11, 25, 63]. To this end, we propose Label Sequence Extension as a mechanism to leverage out-of-batch text descriptions. Concretely, we sample additional text descriptions with a ratio of two thirds entity labels and one third relation labels. To ensure computational tractability in the presence of the quadratic complexity of Transformer, we limit the label sequence to a predefined length $N_L$. We experiment with two sampling strategies: (i) *Uniform sampling* that draws among candidate labels with equal probability; (ii) *Frequency-based sampling* that samples according to the label frequency in the training set.

### 3.3 Addressing Relational Semantic Ambiguity

Datasets with crowd-sourced language annotations [34, 35] exhibit significant label noise and ambiguity. First, the descriptions themselves may be noisy (inaccurate), particularly when the underlying image is challenging to interpret. A second challenge for traditional training schemes is that similar relations can be described differently, thanks to synonyms. For example, the *stand near* relation may be annotated "stand near", "stand next to", "stand by", *etc.*. These forms of *semantic ambiguity* make supervised cross-modal pre-training (which relies on access to consistent labels) challenging. To mitigate this issue, we focus on two aspects of the pre-training input data: (i) the quality of the relation text labels; (ii) the presence of semantically-similar labels in sampled label sequences.

**Relational Quality Labels (RQL).** To tackle the first challenge, we propose a label smoothing [48] approach. The key idea is that we expect the difficulty of subject and object detection for a particular instance to correlate with the confidence of the annotated relation. We therefore propose to estimate annotation quality from the quality of the entity detection stage. Drawing inspiration from the generalised focal loss [39], we instantiate this idea by assessing the quality of the $i$th subject and object detection after bipartite matching [61] as

$$\boldsymbol{e}(i) = \min(\mathrm{GIoU}_{0\text{–}1}(\boldsymbol{B}_s(i), \hat{\boldsymbol{B}}_s(i)), \mathrm{GIoU}_{0\text{–}1}(\boldsymbol{B}_o(i), \hat{\boldsymbol{B}}_o(i))) \tag{4}$$

where $\mathrm{GIoU}_{0\text{–}1}$ denotes generalized IoU from [58] together with a linear scaling function to scale the GIoU value to the range of 0 to 1 and $\hat{}$ denotes ground-truth annotation. The resulting value $\boldsymbol{e}(i)$ is then employed to calibrate the relation label confidence via multiplication: $\tilde{\boldsymbol{R}}(i) = \boldsymbol{e}(i)\hat{\boldsymbol{R}}(i)$.

**Relational Pseudo-Labels (RPL).** To address the second issue, we propose a pseudo-labelling strategy [67] to account for synonyms in the extended sequence. We exploit the fact that text embeddings with high semantic similarity will lie close together, as measured by an appropriate distance function $M(\cdot,\cdot)$. We define the distance between the $i$th annotated relation label $\hat{\boldsymbol{R}}(i) = \{0,1\}^{N_R}$ and the $j$th relation text feature $\tilde{\boldsymbol{L}}_R(j) \in \mathbb{R}^D$ from the extended sequence as

$$M(\hat{\boldsymbol{R}}(i), \tilde{\boldsymbol{L}}_R(j)) = \sum\nolimits_{k=1}^{N_R} \hat{\boldsymbol{R}}(i,k) \cdot m(\tilde{\boldsymbol{L}}_R(k), \tilde{\boldsymbol{L}}_R(j)) \tag{5}$$

where $m(\cdot,\cdot)$ denotes Euclidean distance. Given the $i$th relation label, we apply a scaling function to $M(i,j)$ via $\bar{M}(i,j) = \frac{\max_k(M(i,k)) - M(i,j)}{\max_k(M(i,k))}$ where we have abbreviated $M(\hat{\boldsymbol{R}}(i), \tilde{\boldsymbol{L}}_R(j))$ as $M(i,j)$ for clarity. Next, we use a global threshold $\eta$ to select label texts with high similarities: we set the $j$th label in the $i$th relation labels as $\bar{M}(i,j)$ if $\bar{M}(i,j) > \eta$. Note that when applying either RQL or RPL, the ground truth labels are continuous (rather than discrete). We therefore employ the Quality Focal Loss [39] (rather than the standard Focal Loss [44]) as our objective function.

### 3.4 Pre-training, Fine-tuning and Inference

By design, our pre-training (RLIP) and fine-tuning phases follow a similar process. For a given batch of images with corresponding annotations, we aggregate the results from *Parallel Entity Detection* and *Sequential Relation Inference* to form $N_Q$ triplets per image. During pre-training and fine-tuning, we employ bipartite matching similarly to prior work [61, 75, 6, 72, 70], following in particular the matching cost proposed in [61]. The overall loss is then constructed as follows:

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

where $\mathcal{L}_{l1}, \mathcal{L}_{GIoU}, \mathcal{L}_s, \mathcal{L}_o, \mathcal{L}_r$ denote the $\ell_1$ loss for box regression, GIoU loss [58], CE loss for subject and object classes, and Focal loss for relations (or Quality Focal loss [39] when applying RQL or RPL), respectively. The $\lambda$ terms are fixed weights to balance multi-task training following [61], with $\lambda_1 = 2.5, \lambda_2 = 1, \lambda_3 = 1, \lambda_4 = 1$. During pre-training, the label sequence is constructed from both in-batch and out-of-batch labels. During fine-tuning, we use all text labels contained in the dataset to form the label sequence (unlike pre-training, these labels fall within a pre-defined category list of limited size). Note that we follow [61, 72] and exclude $\mathcal{L}_s$ during fine-tuning since HOI detection detects only humans as subjects. During inference, the confidence score for an object is simply the top-1 score from the softmax distribution over objects, and the relation score is obtained by multiplying the original score from the $\mathrm{sigmoid}$ function and the object score. We rank relation scores and filter out the top-$K$ within those correctly localised triplets (IoU > 0.5) for evaluation. $K$ is set to 100 by default following [61, 72, 42, 70].

## 4 Experiments

**Datasets.** We use the Visual Genome (VG) [34] dataset for RLIP. This dataset contains 108,077 images annotated with free-form text for a wide array of objects and relations (100,298 object

annotations and 36,515 relation annotations). The dataset is pre-processed prior to use (see supplementary material for details). For downstream tasks, we conduct experiments on HICO-DET [5] and V-COCO [15]. HICO-DET contains 37,536 training images and 9,515 testing images, annotated with 600 HOI triplets derived from combinations of 117 verbs and 80 objects. We evaluate under the **Default** setting. V-COCO comprises 2,533 training images, 2,876 validation images and 4,946 testing images annotated with 24 interactions and 80 objects. Results are assessed under two scenarios denoted as $\text{AP}^{\#1}_{role}$ and $\text{AP}^{\#2}_{role}$ as defined by the official evaluation code [15]. Note that all object classes in HICO-DET and V-COCO are identical to COCO.

**Implementation details.** The basic architecture of the encoder and decoder are based on DETR [3] for **ParSe** and DDETR [74] for an additional variant named **ParSeD**. A detailed architecture description of RLIP-ParSeD (which uses an additional transformer for cross-modal fusion) is provided in the supplementary. For *Parallel Entity Detection* and *Sequential Relation Inference*, 3 decoding layers are used. The number of queries $N_Q$ is set to 100 during pre-training and 64 during fine-tuning (following [72]). $\gamma$ in the Focal loss is set to 2 following [61, 72]. $N_L$ in LSE is set to 500 to ensure computational tractability. $\eta$ in RPL is set to 0.3. For pre-training and fine-tuning, the initial learning rate (LR) of the image and text encoders is set to 1e-5, while all other modules are set to 1e-4. For RLIP-ParSeD and ParSeD (object detection and relation detection pre-training), we pre-train model on VG for 50 epochs and drop LR by a factor of 10 at epoch 40. For ParSeD and RLIP-ParSeD, We fine-tune for 60 epochs and drop LR at epoch 40 by a factor of 10. For ParSe and RLIP-ParSe, We fine-tune for 90 epochs and drop LR at epoch 60 by a factor of 10. The pre-training and fine-tuning strategy follow above descriptions unless stated otherwise. Experiments are conducted on 8 Tesla V100 GPU cards with a minibatch size of 32.

**Experimental protocols.** To assess performance under *fine-tuning* and *zero-shot with no fine-tuning* (NF) scenarios, we evaluate on HICO-DET across three HOI sets under the mAP metric: *Rare* (HOIs with training samples less than 10, of which there are 138), *Non-Rare* (HOIs with samples equal to or more than 10, of which there are 462) and *Full* (all HOIs, of which there are 600).

To evaluate performance under the zero-shot formulation considered by [22, 23, 21], we report results on unseen combinations (UC). In particular, we report results under two settings: *UC with rare-first* (UC-RF) selection and *UC with non-rare first* selection (UC-NF), both of which are assessed across three subsets: Unseen (120 HOIs), Seen (480 HOIs) and Full (600 HOIs).

To evaluate few-shot transfer performance, we follow [30] and sample subsets of training annotations from HICO-DET. In detail, we sample $1\%$ and $10\%$ of the total annotations available among the HICO-DET training data, ensuring that all objects and verbs (but not all combinations) exist in the selected annotations. Similarly to the fine-tuning protocol, we evaluate on the *Rare*, *Non-Rare* and *Full* sets.

To assess the robustness of RLIP and its sensitivity to noise in the relation labels, we follow [37, 64] and artificially inject noise into the relation labels by randomly flipping a fixed ratio of verbs in HOI triplets across the training set.

### 4.1 Results and Analysis

**Comparing object detection and relation detection pre-training with RLIP.** An assessment of the benefits of object detection pre-training using the COCO dataset may offer a somewhat optimistic evaluation of this approach, since COCO shares identical object classes with the downstream HOI detection evaluation datasets HICO-DET and V-COCO (and in the latter case shares training images). If we control for this effect by performing object detection pre-training on VG rather than COCO, we observe a significant drop in performance for our ParSeD baseline (from 29.12 to 23.78 across the *Full* set on HICO-DET and from 61.8 to 41.4 on V-COCO for $\text{AP}^{\#1}_{role}$, as shown in Tab. 1). However, since COCO lacks relation annotations, we investigate the benefits of RLIP on VG. We observe that RLIP outperforms vanilla object detection pre-training by a wide margin (boosting performance from 23.78 to 29.21 across the *Full* set on HICO-DET and from 41.4 to 53.1 on V-COCO for $\text{AP}^{\#1}_{role}$), demonstrating the value of incorporating relations as a pre-training cue. Another way to pre-train with relations is to perform relation detection which is still inferior to RLIP (27.45 < 29.21), demonstrating the importance of relational language-image pre-training.

**Leveraging off-the-shelf object detection data without relations.** As shown in the previous experiment, while VG provides relation annotations, it provides a much weaker basis for object

Table 1: Comparisons with previous fine-tuned methods on HICO-DET and V-COCO (column $AP^{\#1}_{role}$ and $AP^{\#2}_{role}$). PT, PTP, OD, RD and MD abbreviate Pre-Training, Pre-Training Paradigm, Object Detection, Relation Detection and Modulated Detection. FRCNN, R, HG and Swin-T denote Faster R-CNN [57], ResNet [18], Hourglass [50] and Swin-Tiny [47]. * denotes RLIP is performed on VG, initialized with parameters from COCO object detection. † denotes fine-tuning for 150 epochs following QAHOI [6].

| PT Data | PTP | Method | Detector | Backbone | Rare | Non-Rare | Full | $AP^{\#1}_{role}$ | $AP^{\#2}_{role}$ |
|---|---|---|---|---|---|---|---|---|---|
| - | - | QAHOI† [6] | DDETR | Swin-T | 22.44 | 30.27 | 28.47 | - | - |
| | - | ParSeD† | DDETR | Swin-T | **25.76** | **31.84** | **30.44** | - | - |
| COCO | OD | InteractNet [14] | FRCNN | R50-FPN | 7.16 | 10.77 | 9.94 | 40.0 | - |
| | | GPNN [54] | FRCNN | R152-DCN | 9.34 | 14.23 | 13.11 | 44.0 | - |
| | | iCAN [12] | FRCNN | ResNet-50 | 10.45 | 16.15 | 14.84 | 45.3 | 52.4 |
| | | UnionDet [31] | RetinaNet | R50-FPN | 11.72 | 19.33 | 17.58 | 47.5 | 56.2 |
| | | PPDM [42] | CenterNet | HG104 | 13.97 | 24.32 | 21.94 | - | - |
| | | HOTR [32] | DETR | ResNet-50 | 17.34 | 27.42 | 25.10 | 55.2 | 64.4 |
| | | HOITransformer [75] | DETR | ResNet-50 | 16.91 | 25.41 | 23.46 | 52.9 | - |
| | | QPIC [61] | DETR | ResNet-50 | 21.85 | 31.23 | 29.07 | 58.8 | 61.0 |
| | | OCN [70] | DETR | ResNet-50 | 25.56 | 32.51 | 30.91 | 64.2 | 66.3 |
| | | CDN [72] | DETR | ResNet-50 | 27.39 | 32.64 | 31.44 | 61.7 | 63.8 |
| | | QAHOI [6] | DDETR | ResNet-50 | 18.06 | 28.61 | 26.18 | - | - |
| | | ParSeD | DDETR | ResNet-50 | 22.23 | 31.17 | 29.12 | 61.8 | 64.0 |
| | | ParSe | DETR | ResNet-50 | 26.36 | 33.41 | 31.79 | 62.5 | 64.8 |
| | | ParSe | DETR | ResNet-101 | **28.59** | **34.01** | **32.76** | **64.4** | **66.5** |
| VG | OD | ParSeD | DDETR | ResNet-50 | 19.59 | 25.03 | 23.78 | 41.4 | 43.0 |
| | RD | ParSeD | DDETR | ResNet-50 | 21.36 | 29.27 | 27.45 | 51.5 | 53.2 |
| | RLIP | RLIP-ParSeD | DDETR | ResNet-50 | **24.45** | **30.63** | **29.21** | **53.1** | **55.0** |
| COCO+VG | RLIP* | RLIP-ParSeD | DDETR | ResNet-50 | 24.67 | 32.50 | 30.70 | 61.7 | 63.8 |
| | RLIP* | RLIP-ParSe | DETR | ResNet-50 | **26.85** | **34.63** | **32.84** | **61.9** | **64.2** |
| GoldG+ | MD | MDETR-ParSe [30] | DETR | ResNet-101 | 22.91 | 31.07 | 29.19 | 53.6 | 56.0 |

detection pre-training than COCO. More broadly, we may expect that object annotations are likely to be more readily available (and greater in scale) than relation annotations. To mitigate this, a simple solution is to simply load object detection parameters pre-trained from an object-annotated dataset (like COCO), to complement the abilities of RLIP. Tab. 1 indicates that RLIP-ParSeD indeed benefits from this approach, surpassing both object detection pre-training and RLIP (29.12, 29.21 $\rightarrow$ 30.70 on the *Full* set). We pre-trained on DETR for 150 epochs, outperforming an expert object detection pre-training (31.79 $\rightarrow$ 32.84). On V-COCO, there is a degradation of performance (62.5 $\rightarrow$ 61.9) which we believe may be caused by the reduced domain alignment (i.e. common training images) relative to COCO object-detection pre-training [61].

**Comparing cross-modal regional alignment pre-training with RLIP.** We next compare to the use of language-region alignment pre-training introduced by MDETR [30], which employed the GoldG+ dataset (this comprises VG, COCO and Flickr30k [53] together with the corresponding annotations for referring expressions, VG regions, Flickr entities, and GQA [24]). For comparison, we initialise RLIP-ParSe with MDETR's parameters and then fine-tune on HICO-DET. The results are reported in Tab. 1 as MDETR-ParSe. Although MDETR makes use of a heavier backbone (ResNet-101) and additional pre-training data, RLIP-ParSe nevertheless surpasses this baseline with a lighter ResNet-50 backbone, demonstrating the effectiveness of RLIP for this task.

**Zero-shot HOI detection.** To assess performance under the zero-shot NF protocol, we compare with other methods using RLIP-ParSe (initialised with COCO parameters followed by RLIP on VG) and ParSe (initialised with COCO parameters). Note that during RLIP, we match subjects against a diverse collection of categories. However HOI detection only needs to detect *person* as subjects. Consequently, for zero-shot NF inference, we filter out subjects that are not classified as *person*. We report results in Tab. 2, where we observe that RLIP outperforms several fully fine-tuned methods. We find, as expected, that regional language-image pre-training methods like MDETR fail under an NF evaluation, since its pre-training lacks the notion of explicit relations. Under UC-RF and UC-NF protocols (fine-tuning for 40 epochs under UC-NF to avoid over-fitting), RLIP-ParSe outperforms previous methods and ParSe by performing RLIP on VG. **Few-shot transfer on HICO-DET.** To

evaluate few-show transfer, we fine-tune ParSeD for 60 epochs as above, while RLIP-ParSeD is fine-tuned for 10 epochs to avoid over-fitting. The results are shown in Tab. 3. We observe that RLIP significantly benefits few-shot fine-tuning relative to object detection pre-training and relation detection pre-training, especially when data is scarce.

Table 2: Results under zero-shot settings on HICO-DET. NR denotes Non-Rare.

| Zero-shot | Method | Rare | NR | Full |
|---|---|---|---|---|
| NF | MDETR-ParSe [30] | 0.00 | 0.00 | 0.00 |
| | RLIP-ParSe | **15.08** | **15.50** | **15.40** |
| **Zero-shot** | **Method** | **Unseen** | **Seen** | **Full** |
| UC-RF | VCL [21] | 10.06 | 24.28 | 21.43 |
| | ATL [22] | 9.18 | 24.67 | 21.57 |
| | FCL [23] | 13.16 | 24.23 | 22.01 |
| | ParSe | 18.53 | 32.21 | 29.06 |
| | RLIP-ParSe | **19.19** | **33.35** | **30.52** |
| UC-NF | VCL [21] | 16.22 | 18.52 | 18.06 |
| | ATL [22] | 18.25 | 18.78 | 18.67 |
| | FCL [23] | 18.66 | 19.55 | 19.37 |
| | ParSe | 19.65 | 24.50 | 23.38 |
| | RLIP-ParSe | **20.27** | **27.67** | **26.19** |

Table 3: Few-shot transfer on HICO-DET. OD, RD denote object detection pre-training and relation detection pre-training.

| Method | Data | Epochs | Rare | Non-Rare | Full |
|---|---|---|---|---|---|
| ParSeD (VG, OD) | 1% | 60 | 0.18 | 2.05 | 1.62 |
| | 10% | 60 | 6.46 | 12.19 | 10.87 |
| ParSeD (VG, RD) | 1% | 60 | 3.74 | 8.62 | 7.50 |
| | 10% | 60 | 12.29 | 17.98 | 16.67 |
| ParSeD (COCO, OD) | 1% | 60 | 5.86 | 10.16 | 9.17 |
| | 10% | 60 | 12.20 | 20.39 | 18.51 |
| RLIP-ParSeD (VG) | 0% | - | 12.30 | 12.81 | 12.69 |
| | 1% | 10 | 16.24 | 16.05 | 16.09 |
| | 10% | 10 | 15.43 | 20.34 | 19.21 |
| RLIP-ParSeD (COCO + VG) | 0% | - | 11.20 | 14.73 | 13.92 |
| | 1% | 10 | 16.22 | 18.92 | 18.30 |
| | 10% | 10 | 15.89 | 23.94 | 22.09 |

**The influence of relation label noise.** To assess sensitivity to noise, we report fine-tuning results on HICO-DET with increasing ratios of relation label noise in Fig. 2. We observe that as label noise increases, the COCO object detection pre-training adopted by prior work exhibits a greater degradation in performance (29.12→24.52, -4.60) than RLIP (29.21→25.68, -3.53) We also observe that when initialising RLIP-ParSeD with COCO pre-trained parameters, RLIP again helps to ameliorate noise, with a more limited loss of performance (30.70→26.87, -3.83) than COCO pre-training and a similar degradation to RLIP-ParSeD with random initialization. Consequently, we deduce that RLIP offers a route to mitigating label corruption and improving model robustness [19].

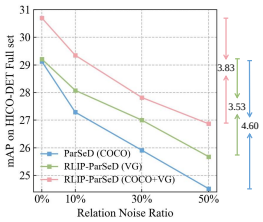


Figure 2: Relation Label Noise.

Table 4: Ablation study of different sampling strategies for label sequence extension using RLIP-ParSeD on HICO-DET.

| Sampling Type | Fine-tuning | | | Zero-shot (NF) | | |
|---|---|---|---|---|---|---|
| | Rare | Non-Rare | Full | Rare | Non-Rare | Full |
| - | 22.58 | 28.98 | 27.51 | 9.77 | 9.97 | 9.92 |
| Uniform | 23.33 | 29.55 | 28.12 | 9.46 | 9.67 | 9.63 |
| Frequency-based | 23.02 | 29.77 | 28.22 | 10.45 | 11.26 | 11.07 |

### 4.2 Ablation studies and analysis

**Ablation study of ParSe on the influence of decoupled representations.** We report an ablation study of the ParSe architecture in Tab. 5 to highlight the importance of decoupling the representation of subjects, objects and relations. The first row of Tab. 5 represents the use of coupled representations for subjects, objects and relations [61]. The second row of Tab. 5 represents the use of coupled representations for subjects and objects that are disentangled from relations [72]. The final row (ParSe) uses fully-disentangled representations. We observe a clear gain resulting from ParSe over methods using a joint representation of (some subset of) subject, object and relation triplets.

**Ablation study of sampling strategy.** In Tab. 4, we present an ablation study to assess the efficacy of out-of-batch sampling strategies. Intuitively, *uniform sampling* will up-weight descriptions from the tail of the distribution while *frequency-based sampling* will preserve the distribution. Both bring improvements to fine-tuning by providing additional negative samples. However, by over-sampling descriptions from the tail, uniform sampling performs less well with common texts in the downstream task, and thus fares less well overall.

**Visualisation of ParSe attention weights.** We visualise attention weights from ParSe for several example images in Fig. 3. We observe that ParSe attends distinct regions for subjects, objects and relations. This aligns with our motivation that entity detection is best supported by local context,

Table 5: Fine-tuning results with ParSe (COCO, OD) on HICO-DET.

| ParSe Architecture | Coupling | Rare | Non-Rare | Full |
|---|---|---|---|---|
| - | coupled subject, objects and relations | 23.18 | 31.45 | 29.55 |
| w/ Se | coupled subject and objects | 25.58 | 32.50 | 30.91 |
| w/ ParSe | fully decoupled | **26.36** | **33.41** | **31.79** |

Table 6: Ablation study of RLIP techniques using RLIP-ParSeD on HICO-DET.

| RLIP Technique | | | Fine-tuning | | | Zero-shot (NF) | | | Relation | |
|---|---|---|---|---|---|---|---|---|---|---|
| LSE | RQL | RPL | Rare | Non-Rare | Full | Rare | Non-Rare | Full | Uniformity↓ | Alignment↓ |
| | | | 22.58 | 28.98 | 27.51 | 9.77 | 9.97 | 9.92 | -0.8233 | 0.3650 |
| ✓ | | | 23.02 | 29.77 | 28.22 | 10.45 | 11.26 | 11.07 | -1.0556 | 0.4542 |
| ✓ | ✓ | | 24.32 | 30.32 | 28.94 | 11.49 | 12.60 | 12.34 | -1.3986 | 0.6072 |
| ✓ | ✓ | ✓ | 24.45 | 30.63 | 29.21 | 12.30 | 12.81 | 12.69 | -1.3265 | 0.5799 |

while relation inference draws on additional spatial context, conditioning on subjects and objects like hands and string, as well as the wet ground and sky where appropriate.

**Ablation study of RLIP techniques.** In Tab. 6, we present an ablation study of the three proposed technical contributions. We observe that each benefits both fine-tuning and zero-shot NF under all metrics. LSE attains a greater boost for the *Non-Rare* set (by sampling according to the training set distribution). On the other hand, RPL enhances results for the *Rare* set, likely due to the propensity of RPL to label rare descriptions as positive (due to the long-tailed distribution of text labels).

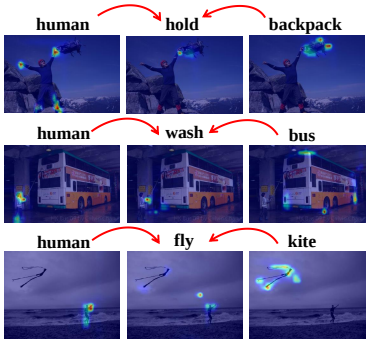


Figure 3: Attention weight analysis for the top-1 scored verb. Weights are extracted from *Parallel Entity Detection* (human and object) and *Sequential Relation Inference* (verb).

**Understanding LSE, RQL and RPL with Uniformity and Alignment.** To gain insight into representation quality, Wang et al. [63] proposed two metrics, *uniformity* and *alignment*, which we employ here to better understand the influence of our contributions. To this end, we perform a zero-shot (NF) evaluation on HICO-DET, using bipartite matching to assign predicted triplets to ground-truth labels. We then calculate uniformity and alignment metrics for relation features via $\mathcal{L}_u(f;t) = \log(\mathbb{E}_{(x,y)\overset{\text{i.i.d}}{\sim}\mathcal{P}_{data}}[e^{-t\|f(x)-f(y)\|_2^2}])$ and $\mathcal{L}_a(f;\alpha) = \mathbb{E}_{(x,y)\sim\mathcal{P}_{pos}}[\|f(x)-f(y)\|_2^\alpha]$, where $\alpha, t = 2$. We present the results in Tab. 6, where we observe that LSE and RQL both reduce uniformity, the former through additional negative pairs and the latter by reducing the loss assigned to over-confident text labels. We also observe that RPL yields better alignment through discovering additional positive labels. The results indicate, however, a trade-off between the metrics—a useful direction for future work would be to determine how to find an appropriate balance between them.

**Verb-wise mAP Analysis for zero-shot (NF) evaluation.** We provide analysis to give a sense of the verb overlap of HICO with VG. We use "relationship aliases" from the official VG website to obtain as many HOI verb annotations from VG as possible by string matching. The result is shown in Tab. 19 in the supplementary material. We observe that in VG there are only 2,203 HOI verb annotations even when considering relationship aliases—approximately 1.47% of the number of relationship annotations in HICO-DET. 30 HOI verbs do not have an annotation and 45 HOI verbs have five or fewer annotations. In RLIP-ParSe (COCO+VG), we observe that mAP for the 30 verbs is 5.56 while mAP for the remaining 87 verbs is 18.12. If we use uni-modal relation detection pre-training, the result for the 30 verbs degrades to zero. In light of this, we conjecture that existing relations can transfer their knowledge to the inference of non-existing relations in HOI detection. To provide a more detailed analysis, we show the verb-wise mAP on HICO verbs in VG (Fig. 6) and not in VG (Fig. 7) with zero-shot (NF) evaluation (figures are provided in the supplementary material), where we observe solid performance for some verbs.

**Probing into reasons for the verb zero-shot performance.** We aim to ***qualitatively understand where the zero-shot ability stems from***. In the above analysis, *pay* has the highest performance among verbs not seen by VG (Fig. 7 in the supplementary material). In the methodology section, we

Table 7: VG verb ranking given similar subject-object triplets from HICO-DET. Verbs are in ascending order of Euclidean distance. (The Cosine distance can also output similar rankings.)

| **"pay"** ("parking meter") | **putting money in** | **collecting money at** | **puts change into** | **repairing** | **checking** | **next to** | ... |
|---|---|---|---|---|---|---|---|
| Count | 1 | 1 | 1 | 1 | 1 | 1 | ... |
| Euclidean | 11.56 | 11.70 | 13.34 | 14.21 | 15.16 | 16.12 | ... |
| Cosine | 0.4560 | 0.4576 | 0.3108 | 0.2554 | 0.1583 | 0.0709 | ... |

present the conditional query generation that constrains the verb inference to be related to subjects and objects, providing verb inference with a conditional context. Thus, to analyze how this ability of verb zero-shot inference emerges, we need to consider the subject and object context as they are essential to predict the verb in ParSe. For the verb *pay* in HICO-DET, there is only one possible triplet annotated, "person pay parking meter". Then, we want to answer, "**Is there any triplet annotated with similar or identical subjects and objects that transfer the inference ability to *pay*?**" To answer this question, we search for triplets annotated with similar subjects and objects to HICO-DET from VG (For details, please refer to the analysis of Tab. 20 in the supplementary material.). We report the verb distribution of the limited number of triplets that are found, ranking the verbs in ascending order of Euclidean distance to the target verb (Tab. 7). From this table, we can see that the verbs quantitatively closer (in Euclidean distance or Cosine distance) to *pay* have similar meanings to *pay*, shown by their lexical variants or grammatical variants (e.g., *putting money in* has a similar meaning to *pay*). Thus, in the VG dataset, there is *human putting money in parking meter*, which may transfer to the zero-shot recognition of *person pay parking meter* in HICO-DET. More examples can be found in Tab. 20. In short conclusion, we conjecture that the zero-shot inference ability of RLIP is not from the scale of annotations (by comparing relation detection pre-training and RLIP using VG), but the ability to transfer the verb inference knowledge from semantically similar annotations. This analysis also accords with previous works [55, 38] that semantic diversity is important as it introduces large-scale potential annotations, ensuring a model transfers well to different data distributions.

Second, we aim to ***demonstrate quantitatively how RLIP pre-trains the model to perform zero-shot detection from the perspective of representation learning***. We employ the Uniformity metric introduced in [63]. Uniformity is a metric to assess a model's generalization in contrastive learning. In this case, since label textual embeddings serve as a classifier in RLIP, we calculate the Uniformity of the seen verbs, unseen verbs and all verbs, aiming to observe how the generalization changes before and after RLIP, and how the generalization varies between seen verbs and unseen verbs. The results are shown in Tab. 8. As can be seen from the table, Uniformity values are high before RLIP, suggesting that the representations before RLIP are distributed compactly, leading to a poor classifier. However, after RLIP is performed, the 87 seen verbs have a substantially lower Uniformity value, corresponding with decent zero-shot performance. Similarly, the 30 unseen verbs and the combination of 117 verbs also have excellent Uniformity values, contributing to unseen zero-shot performance. Through this quantitative observation, we think that from the perspective of representations, RLIP contributes to improved zero-shot performance.

Table 8: Uniformity analysis of the seen verbs, unseen verbs and all verbs before and after RLIP. Lower uniformity value is better.

| **Verb Set** | **Seen (87)** | **Unseen (30)** | **All (117)** |
|---|---|---|---|
| Before RLIP | -0.00367 | -0.00436 | -0.00388 |
| After RLIP | -3.73780 | -3.59457 | -3.71330 |

From all the above analysis, we conjecture that the zero-shot performance is not caused by the increased dataset size or annotations, but rather from the generalization in representations obtained by pre-training with language supervision.

## 5 Conclusion

In this paper, we propose RLIP as a pre-training strategy for HOI detection. We show that RLIP, together with our additional technical contributions, boosts HOI detection performance under fine-tuning, zero-shot and few-shot evaluations, and improves robustness against noisy annotations.

**Acknowledgements:** We would like to appreciate anonymous reviewers for their valuable feedback and members from DAMO Academy's Vision Lab for sharing computational resources. This work was supported in part by National Natural Science Foundation of China Grant No. 62173298 and by Alibaba Group through Alibaba Research Intern Program.

## References


[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34, 2021.

[2] 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.

[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.

[4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 381–389. IEEE, 2018.

[5] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1017–1025, 2015.

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

[7] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9004–9013, 2021.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

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

[10] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7373–7382, 2021.

[11] Yutong Feng, Jianwen Jiang, Mingqian Tang, Rong Jin, and Yue Gao. Rethinking supervised pre-training for better downstream transferring. In *International Conference on Learning Representations*, 2022.

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

[13] Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*, 2019.

[14] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018.

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

[16] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9677–9685, 2019.

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[19] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721. PMLR, 2019.

[20] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420, 2017.

[21] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *European Conference on Computer Vision*, pages 584–600. Springer, 2020.

[22] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 495–504, 2021.

[23] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14646–14655, 2021.

[24] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019.

[25] Ashraful Islam, Chun-Fu Chen, Rameswar Panda, Leonid Karlinsky, Richard J. Radke, and Rogério Schmidt Feris. A broad study on the transferability of visual representations with contrastive learning. *2021 IEEE/CVF International Conference on Computer Vision*, pages 8825–8835, 2021.

[26] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020.

[27] 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.

[28] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1219–1228, 2018.

[29] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3668–3678, 2015.

[30] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

[31] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *European Conference on Computer Vision*, pages 498–514. Springer, 2020.

[32] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 74–83, 2021.

[33] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *European Conference on Computer Vision*, pages 718–736. Springer, 2020.

[34] 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.

[35] 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.

[36] 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. *Advances in Neural Information Processing Systems*, 34, 2021.

[37] Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

[38] 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.

[39] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33:21002–21012, 2020.

[40] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. Hoi analysis: Integrating and decomposing human-object interaction. *Advances in Neural Information Processing Systems*, 33, 2020.

[41] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2019.

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

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

[44] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[45] 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.

[46] 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.

[47] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[48] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, pages 6448–6458. PMLR, 2020.

[49] 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.

[50] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[51] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10870–10879, 2020.

[52] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1981–1990, 2019.

[53] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2641–2649, 2015.

[54] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[55] 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.

[56] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[57] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[58] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.

[59] Guanglu Song, Yu Liu, and Xiaogang Wang. Revisiting the sibling head in object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11563–11572, 2020.

[60] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[61] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021.

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[63] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[64] Zhi-Fan Wu, Tong Wei, Jianwen Jiang, Chaojie Mao, Mingqian Tang, and Yu-Feng Li. Ngc: A unified framework for learning with open-world noisy data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 62–71, October 2021.

[65] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[66] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 684–699, 2018.

[67] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196, 1995.

[68] Hangjie Yuan and Dong Ni. Learning visual context for group activity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3261–3269, 2021.

[69] Hangjie Yuan, Dong Ni, and Mang Wang. Spatio-temporal dynamic inference network for group activity recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7476–7485, 2021.

[70] Hangjie Yuan, Mang Wang, Dong Ni, and Liangpeng Xu. Detecting human-object interactions with object-guided cross-modal calibrated semantics. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):3206–3214, Jun. 2022.

[71] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021.

[72] 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. *Advances in Neural Information Processing Systems*, 34, 2021.

[73] Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13234–13243, 2021.

[74] 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.

[75] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11825–11834, 2021.

## Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] It's mentioned in paragraph "**Leveraging off-the-shelf object detection data without relations**", and discussed in more detail in the supplementary material.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] It's included in the supplementary material.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Data is publicly available and instructions are included in the experiment section.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] They are included in Methodology and Experiments section.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes]

   (b) Did you mention the license of the assets? [Yes] It's included in the supplementary material.

   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

# A Appendix

In this supplementary material, we first discuss the potential societal impact (Appendix A.1) and limitations (Appendix A.2) of our approach. Next, we provide further details on the architecture of RLIP-ParSeD (Appendix A.3), dataset pre-processing (Appendix A.4), phased pre-training (Appendix A.5), attention analysis (Appendix A.6) and subject-object query pairing (Appendix A.7). Finally, we provide additional experiments and analysis (Appendix A.8) and discuss our use of datasets (Appendix A.9). Codes will be publicly available upon publication.

## A.1 Potential Societal Impact

By targeting improved HOI detection, our work has the potential to bring societal and commercial benefits in medical, retail, security and sports analysis applications. However, it is inherently a dual-use technology, providing functionality with scope for abuse. For example, better HOI detection may make it easier to conduct unlawful surveillance. Moreover, due to biases present among training datasets, it is also likely that our system does not perform equally across all demographics. Therefore, we caution that our approach represents a research proof-of-concept and is not suitable for real-world usage without a rigorous evaluation of the deployment context and appropriate oversight.

## A.2 Limitations and Potential Future Works

As noted in the experiment section of the main paper, one limitation of our method is its dependence on a particular form of annotations (*i.e.* the pre-training data must be annotated with relation triplets), which are not always available, or exist at a diminished scale relative to object detection annotations. In our work, we investigated one potential solution by bootstrapping RLIP from object detection parameters to mitigate annotation scarcity. Although existing relation annotations are limited, we do not anticipate that this will remain the case. Indeed, we hope that our work will inspire future work to focus on this problem and dataset contributions will follow. Besides, we provide ways to scale up datasets as future works. For example, we could reuse a grounding dataset with entities annotated. Then, a language processing tool like spaCy [20] can be adopted as a tool to obtain their relations from captions. Even if we do not have subjects and objects but only image-caption pairs, we can combine the use of spaCy and methods like GLIP [38] to create abundant triplet annotations. Based on the analysis, we think our method is still promising and inspiring, paving a path for further research.

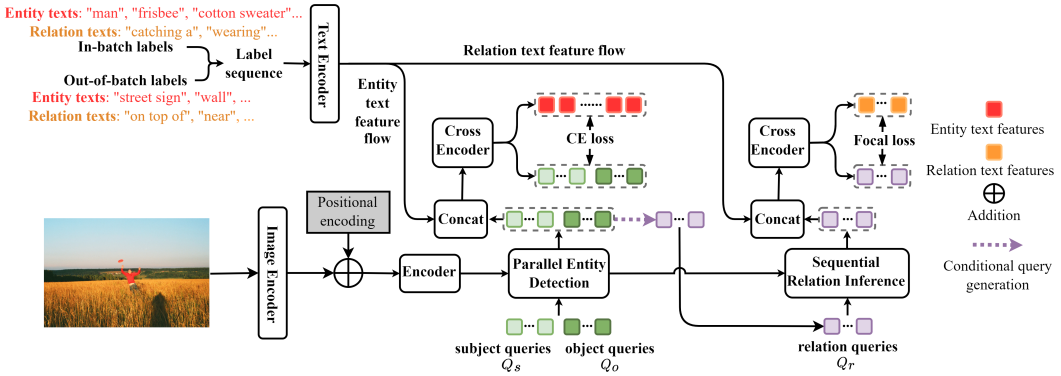


Figure 4: An overview of our pre-training framework, RLIP-ParSeD. The encoder represents DDETR-style encoders. The *Parallel Entity Detection* and *Sequential Relation Inference* blocks represent independent DDETR-style decoders responsible for entity and relation detection, respectively. The cross encoder represents DETR-style encoders. We omit the localisation loss for clarity.

A second limitation of our approach is its requirement for long optimisation schedules during pre-training—a characteristic that is inherited from DETR [3]. Although DDETR [74] ameliorates this issue to some extent, it achieves weaker performance. Thus, one potential for future work is to combine the benefits of DETR's high performance and DDETR's fast convergence speed.

### A.3 Overall Architecture of RLIP-ParSeD

We present the detailed architecture of RLIP-ParSeD as shown in Fig. 4. The major difference between RLIP-ParSe (shown in the main paper) and RLIP-ParSeD is the cross-modal fusion module. For the former, we use the detection Transformer encoder to directly fuse language-image features as previous works have done [36, 30, 1]. For the latter, we use an additional Transformer encoder to fuse the decoded queries and language features following [49] since deformable attention [74] from DDETR relies on spatial coordinates, which language features do not have. Note that the decoded subject and object queries are fused with entity text features and the decoded relation queries are fused with relation text features. The two cross encoders do not share parameters. The localisation loss is applied on decoded queries (after *Parallel Entity Detection* and before the cross encoder). Other architecture/RLIP details follow the descriptions provided in the main paper.

### A.4 Pre-Processing Steps for Visual Genome

Due to the crowd-sourcing process used to construct Visual Genome [34], there are many redundant annotations. Thus, we conduct basic cleaning steps to filter out such annotations and all of our experiments are conducted on the dataset after pre-processing. The steps are listed as follows:

- We keep the first object text description for every object because a very small proportion of objects have multiple text descriptions.
- We filter out redundant triplets by **i)** keeping only one if there are multiple identical triplets and **ii)** keeping only one if there are multiple triplets with identical subject descriptions, object descriptions, relation descriptions and similar box locations (with both the subject box's and the object box's IoU>0.5)
- We filter out redundant triplets if the number of triplets in one image is greater than the number of queries $N_Q = 100$.

### A.5 Additional Details for Phased Pre-training

In the experiments section of the main paper, we describe **i)** how object detection parameters (obtained via pre-training on COCO) can be used to initialise RLIP-ParSe and RLIP-ParSeD and **ii)** how MDETR [30] parameters (obtained via pre-training on GoldG+) can be used to initialise MDETR-ParSe. Here, we provide further details on how these are implemented.

For the main blocks of RLIP-ParSe pre-trained on COCO, we initialise the image encoder, cross encoder, *Parallel Entity Detection* block and *Sequential Relation Inference* block parameters with

parameters from the image encoder, detection Transformer encoder, detection Transformer decoder (first 3 layers) and detection Transformer decoder (first 3 layers) of a COCO pre-trained DETR model following [72]. We initialise the FFN layers of RLIP-ParSe pre-trained on COCO using the localization FFN layer parameters and entity classification FFN layer parameters (since object categories in HICO-DET and V-COCO are identical to COCO). For RLIP-ParSeD pre-trained on COCO, parameter initialisation follows RLIP-ParSe. Other parameters are randomly initialised.

For the main blocks of MDETR-ParSe [30], we initialise the parameters of the text encoder, image encoder, cross encoder, *Parallel Entity Detection* block and *Sequential Relation Inference* block with parameters from the text encoder, image encoder, cross encoder, detection Transformer decoder (first 3 layers) and detection Transformer decoder (first 3 layers) of a GoldG+ pre-trained MDETR model. For the FFN layers of MDETR-ParSe, we initialise from localization FFN layers. Other parameters are randomly initialised.

### A.6 Details of Attention Weight Analysis

In Figure 3 of the main paper, we provide an analysis of the attention weights produced by ParSe (extracted from ParSe in RLIP-ParSe). Here, we provide additional details of how this analysis is conducted. During image inference, we employ a Transformer [62] architecture to decode queries. In this decoding process, a $\mathrm{softmax}$ function is used to normalize attention weights calculated by a scaled dot-product attention. The logits after the $\mathrm{softmax}$ function indicate the importance of regions (since queries aggregate values according to the logits). Thus, we extract the logits after the $\mathrm{softmax}$ function in the last Transformer layer of both the *Parallel Entity Detection* and the *Sequential Relation Inference* block for the top-1 scored verb. To visualize the attention weights, we linearly scale the range of logits to 0–255 (and cast to integers to produce an image).

### A.7 Computational Overhead of the Subject and Object Query Pairing

The pairing of humans and objects are performed by index-matching as is stated in the main paper. Thus, we pair humans and objects with identical indices (e.g., the first decoded feature from the subject queries and the first decoded feature from the object queries are paired.). Due to the simplicity of this matching strategy, the cost is trivial ($\mathcal{O}(1)$ cost) compared to the overall overhead during model inference.

### A.8 Additional Experiments and Analysis

**Ablation study of ParSe on the influence of decoupled representations.** We report a further ablation study of the ParSe architecture in Tab. 9 to highlight the importance of decoupling the representation of subjects, objects and relations. The first row of Tab. 9 represents the use of coupled representations for subjects, objects and relations [61]. The second row of Tab. 9 represents the use of coupled representations for subjects and objects that are disentangled from relations [72]. The final row (ParSe) uses fully-disentangled representations. We observe a clear gain resulting from ParSe over methods using a joint representation of (some subset of) subject, object and relation triplets.

Table 9: Fine-tuning results with ParSe (COCO) on HICO-DET.

| ParSe Architecture | Coupling | Rare | Non-Rare | Full |
|---|---|---|---|---|
| - | coupled subject, objects and relations | 23.18 | 31.45 | 29.55 |
| w/ Se | coupled subject and objects | 25.58 | 32.50 | 30.91 |
| w/ ParSe | fully decoupled | **26.36** | **33.41** | **31.79** |

**Robustness towards different backbones.** Compare more thoroughly with CDN [72] and QAHOI [6], we perform extensive experiments to demonstrate the effectiveness of the uni-modal detection pipeline ParSe as shwon in Tab. 10. As the table indicates, ParSe outperforms CDN-L with half the number of decoding layers and a single-stage fine-tuning with a clear gain (+0.69 mAP on Full set). When compared to QAHOI, ParSe improves by 1.18 mAP on Full set with only two fifths number of fine-tuning epochs. If using the same number of epochs, ParSe can surpass it by 1.97 mAP on Full set and more improvement on the Rare set (+3.32mAP).

Table 10: Fully-finetuned results on HICO-DET with different backbones. PTP, DL and PT denote Pre-training paradigm, decoding layers and pre-training.

| Method | Backbone | DL | PTP | PT data | #Tuning Epochs | Rare | Non-Rare | Full |
|---|---|---|---|---|---|---|---|---|
| CDN-L [72] | ResNet-101 | 12 | OD | COCO | 90+10 | 27.19 | 33.53 | 32.07 |
| ParSe | ResNet-101 | 6 | OD | COCO | 90 | 28.59 | 34.01 | 32.76 |
| QAHOI [6] | Swin-T | 6 | - | - | 150 | 22.44 | 30.27 | 28.47 |
| ParSe | Swin-T | 6 | - | - | 60 | 23.77 | 31.40 | 29.65 |
| ParSe | Swin-T | 6 | - | - | 150 | 25.76 | 31.84 | 30.44 |

**Superiority over other models with VG and COCO data** To have a fairer comparison with previous methods using VG and COCO data, we adopt CDN [72] as a base method and then add the VG dataset to its pre-training stage. Since RLIP also adopts the relation annotations in VG, we also try to include these annotations in the uni-modal pre-training. Thus, we resort to relation detection on VG. To be more specific, we perform uni-modal relation detection pre-training by using linear classifiers for verbs and entities rather than matching with texts. The results are shown in Tab. 11. We can see from the table that by using uni-modal relation detection pre-training, CDN still trails RLIP-ParSe with the same number of epochs of pre-training and fine-tuning, which shows the effectiveness of RLIP. Even if comparing it with ParSe using relation detection pre-training, we can still observe an improvement of ParSe over CDN, demonstrating the usefulness of decoupling triplet representations.

Table 11: Fully-finetuned results on HICO-DET with VG and COCO dataset. RD and PT denote relation detection and pre-training.

| Method | Detector | Data | PT Paradigm | PT #Epochs | Rare | Non-Rare | Full |
|---|---|---|---|---|---|---|---|
| CDN | DETR | COCO+VG | Relation Detction | 150 | 25.65 | 32.75 | 31.12 |
| ParSe | DETR | COCO+VG | Relation Detction | 150 | 26.00 | 33.40 | 31.70 |
| RLIP-ParSe | DETR | COCO+VG | RLIP | 150 | 26.85 | 34.63 | 32.84 |

**Few-shot transfer with RLIP-ParSe.** To evaluate few-show transfer with ParSe and RLIP-ParSe, we fine-tune ParSe for 90 epochs as above, while RLIP-ParSe is fine-tuned for 10 epochs to avoid over-fitting. The detailed results are shown in Tab. 12 and evaluation curves are shown in Fig. 5. We observe that RLIP with object detection initialisation significantly benefits few-shot fine-tuning relative to object detection pre-training in terms of final results and convergence speed, especially when data is scarce. RLIP-ParSe with 1% data achieves similar performance with ParSe with 10% data.

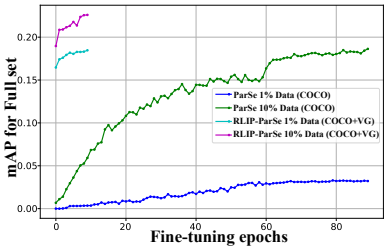


Figure 5: Evaluation curves for few-shot transfer on HICO-DET with ParSe and RLIP-ParSe.

Table 12: Few-shot transfer on HICO-DET with ParSe and RLIP-ParSe.

| Method | Data | Epochs | Rare | Non-Rare | Full |
|---|---|---|---|---|---|
| ParSe (COCO) | 1% | 90 | 1.69 | 3.67 | 3.21 |
| | 10% | 90 | 14.61 | 19.56 | 18.42 |
| RLIP-ParSe (COCO+VG) | 0% | - | 15.08 | 15.50 | 15.40 |
| | 1% | 10 | 17.47 | 18.76 | 18.46 |
| | 10% | 10 | 20.16 | 23.32 | 22.59 |

**Detailed results of relation label noise with RLIP-ParSe and RLIP-ParSeD** We present detailed results concerning the influence of relation label noise on RLIP-ParSeD and ParSeD in Tab. 13 (which corresponds to Figure 2 in the main paper) and the influence of relation label noise on RLIP-ParSe and ParSe in Tab. 14. The RLIP-ParSe and ParSe results support our claim that RLIP helps to ameliorate noise since ParSe suffers a greater degradation of performance (31.79→25.19, -6.6) than RLIP-ParSe (32.84→27.75, -5.09).

Table 13: Relation label noise on HICO-DET with ParSeD and RLIP-ParSeD.

| Method | Noise | Rare | Non-Rare | Full |
|---|---|---|---|---|
| ParSeD (COCO) | 0% | 22.23 | 31.17 | 29.12 |
| | 10% | 19.63 | 29.58 | 27.29 |
| | 30% | 17.14 | 28.52 | 25.91 |
| | 50% | 15.82 | 27.12 | 24.52 |
| RLIP-ParSeD (VG) | 0% | 24.45 | 30.63 | 29.21 |
| | 10% | 21.59 | 30.02 | 28.08 |
| | 30% | 19.60 | 29.21 | 27.00 |
| | 50% | 17.11 | 28.24 | 25.68 |
| RLIP-ParSeD (COCO+VG) | 0% | 24.67 | 32.50 | 30.70 |
| | 10% | 19.86 | 32.20 | 29.35 |
| | 30% | 18.45 | 30.62 | 27.82 |
| | 50% | 17.81 | 29.58 | 26.87 |

Table 14: Relation label noise on HICO-DET with ParSe and RLIP-ParSe.

| Method | Noise | Rare | Non-Rare | Full |
|---|---|---|---|---|
| ParSe (COCO) | 0% | 26.36 | 33.41 | 31.79 |
| | 10% | 21.59 | 30.80 | 28.68 |
| | 30% | 20.52 | 29.99 | 27.81 |
| | 50% | 15.01 | 28.23 | 25.19 |
| RLIP-ParSe (COCO+VG) | 0% | 26.85 | 34.63 | 32.84 |
| | 10% | 24.62 | 33.54 | 31.49 |
| | 30% | 23.12 | 31.75 | 29.77 |
| | 50% | 20.09 | 30.04 | 27.75 |

**Sensitivity analysis of hyper-parameter $\eta$ in RPL.** In Tab. 15, we present a sensitivity analysis for the $\eta$ hyperparamter used in RPL. As $\eta$ decreases, RPL selects more descriptions with high similarities as positive, boosting performance. However, if $\eta$ is too low, there is an increased risk of false positives arising in the pseudolabeling process. We choose $\eta = 0.3$ in the main paper according to this experiment.

Table 15: Results with varying $\eta$ in RPL with RLIP-ParSeD on HICO-DET. LSE and RQL are used by default.

| | Fine-tuning | | | Zero-shot (NF) | | |
|---|---|---|---|---|---|---|
| $\eta$ | Rare | Non-Rare | Full | Rare | Non-Rare | Full |
| 0.2 | 24.05 | **30.73** | 29.19 | 10.95 | 12.71 | 12.30 |
| 0.3 | **24.45** | 30.63 | **29.21** | **12.30** | **12.81** | **12.69** |
| 0.4 | 23.67 | 29.90 | 28.47 | 11.97 | 12.80 | 12.61 |
| 0.5 | 22.63 | 29.79 | 28.14 | 11.70 | 12.09 | 12.00 |

**Design choice of distance function $m(\cdot,\cdot)$ in RPL.** For the design choice of $m(\cdot,\cdot)$, we also experiment on another widely-adopted distance function Cosine distance. We conduct a sensitivity analysis of the hyper-parameter $\eta$ to compare with the one chosen in the main paper (last row of results). The zero-shot (NF) results of Cosine distance using RLIP-ParSeD is shown in Tab. 16. We observe that Euclidean distance is slightly better (the last row of results is selected in the paper). Since both methods have similar computational overhead, in the paper, we choose the Euclidean distance.

Table 16: Zero-shot (NF) results with varying $\eta$ in RPL with RLIP-ParSeD on HICO-DET. LSE and RQL are used by default.

| Distance Function | $\eta$ | Rare | Non-Rare | Full |
|---|---|---|---|---|
| Cosine | 0.3 | 11.21 | 12.53 | 12.23 |
| Cosine | 0.4 | 11.92 | 12.82 | 12.61 |
| Cosine | 0.5 | 11.76 | 12.71 | 12.49 |
| Cosine | 0.6 | 11.30 | 12.22 | 12.01 |
| **Euclidean** | **0.3** | **12.30** | **12.81** | **12.69** |

**Comparing RQL and RPL with Label Smoothing Regularization.** Since we employ RPL and RQL to smooth the target distributions used during training to account for ambiguity, it is useful to compare their effectiveness to a manually-designed Label Smoothing Regularisation (LSR) method, such as the one introduced in [60]. We experiment with using LSR for relation labels (following RPL and RQL). However, we find that LSR tends to degrade performance, while the proposed RPL and RQL approaches boost all metrics.

Table 17: Results comparing Label Smoothing Regularization (LSR) [60] with RPL+RQL on HICO-DET. We use RLIP-ParSeD with LSE as a base model.

| Ambiguity suppression | Fine-tuning | | | Zero-shot (NF) | | |
|---|---|---|---|---|---|---|
| | Rare | Non-Rare | Full | Rare | Non-Rare | Full |
| - | 23.02 | 29.77 | 28.22 | 10.45 | 11.26 | 11.07 |
| LSR | 23.51 | 29.38 | 28.03 | 10.03 | 10.84 | 10.65 |
| RPL+RQL | **24.45** | **30.63** | **29.21** | **12.30** | **12.81** | **12.69** |

**Similarity analysis between in-batch labels and out-of-batch labels** The in-batch labels are aggregated from images' annotations, and the out-of-batch labels are sampled from the whole dataset, which does not overlap with in-batch labels. Since the contrastive loss optimizes to push away the negative textual labels, we can observe the change of the similarities of the negative and positive labels. To quantitatively analyze it, we simulate the training process by out-of-batch sampling, and observe the change of similarities by calculating the average pairwise distance of the positive labels to the negative labels. We mainly compare the object and relation similarity based on the RoBERTa model before and after RLIP pre-training. The results are shown in Tab. 18. From this table, we can see that the Cosine similarity decreases, and Euclidean distance increases. Note that before RLIP, the discrimination ability of text embeddings are poor, which corresponds with previous work [13]. The results indicate whichever distance function we adopt (Cosine or Euclidean distance) and whichever kind of feature we observe (object or relation), the similarity between in-batch labels and out-of-batch labels decreases after performing RLIP. This enables the language model to adapt well to the visual representations and serve as a good classifier.

Table 18: Similarity analysis between in-batch labels and out-of-batch labels before and after RLIP. Cos and Euc abbreviate Cosine distance and Euclidean distance.

| Model | Object (Cos) | Relation (Cos) | Object (Euc) | Relation (Euc) |
|---|---|---|---|---|
| **RoBERTa (Before RLIP)** | 0.9991 | 0.9986 | 0.2502 | 0.3156 |
| **RoBERTa (After RLIP)** | 0.0084 | 0.0208 | 18.1943 | 16.9177 |

**Verb-wise mAP analysis for zero-shot (NF) evaluation** We provide analysis to give a sense of the verb overlap of HICO with VG. We use "relationship aliases" from the official VG website to obtain as many HOI verb annotations from VG as possible. The result is shown in Tab. 19.

Table 19: Verb overlap of HICO with VG.

| Dataset | #images | HOI verb annos | HOI verb annos' ratio | Imbalance ratio |
|---|---|---|---|---|
| VG | 108K | 2,203 | 1.47% | 304 |

We observe that in VG, there are 2,203 HOI verb annotations even when considering relationship aliases—approximately 1.47% of the number of relationship annotations in HICO-DET. 30 HOI verbs do not have an annotation and 45 HOI verbs have five or fewer annotations. In RLIP-ParSe (COCO+VG), we observe that mAP for the 30 verbs is 5.56 while mAP for the remaining 87 verbs is 18.12. If we use uni-modal relation detection pre-training, the result for the 30 verbs degrades to zero.

To provide a more detailed analysis, we show the verb-wise mAP on HICO verbs in VG (Fig. 6) and not in VG (Fig. 7) with zero-shot (NF) evaluation. We can observe solid performance for some verbs.

**Probing into reasons for the verb zero-shot performance** We aim to ***qualitatively understand where the zero-shot ability stems from***. In the above analysis, *pay* has the highest performance among verbs not seen by VG. In the main paper, we present the conditional query generation that constrains the verb inference to be related to subjects and objects, providing verb inference with a conditional context. Thus, to analyze how this ability of verb zero-shot inference emerges, we need to consider the subject and object context as they are essential to predict the verb in ParSe. For the verb *pay* in HICO-DET, there is only one possible triplet annotated, "person pay parking meter". Then, we want to answer, "**is there any triplet annotated with similar or identical subjects and**

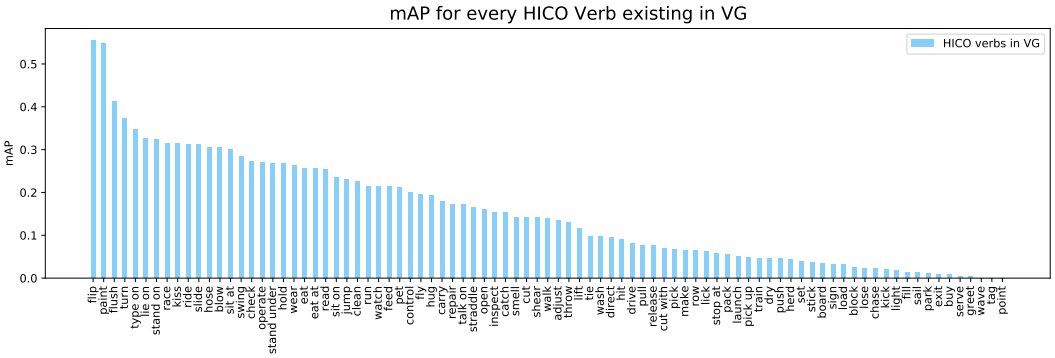


Figure 6: Verb-wise mAP Analysis for Zero-Shot (NF) Evaluation. Presented verbs exist in VG.

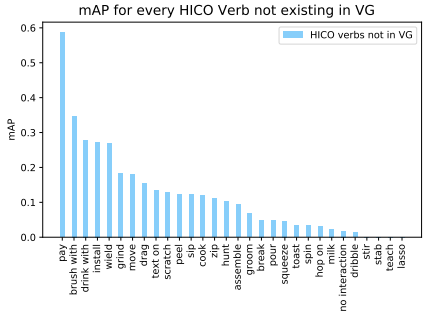


Figure 7: Verb-wise mAP Analysis for Zero-Shot (NF) Evaluation. Presented verbs do not exist in VG.

**objects that transfer the inference ability to *pay*?**" To answer this question, we search for triplets annotated with similar subjects and objects to HICO-DET from VG. For the subjects, we heuristically select ones whose textual descriptions have any one of the following strings: *man, woman, person, friend, guy, dude, human, people, driver, passenger, hand, limb*. For the objects, we heuristically select ones whose textual descriptions have the string of the target object. By this processing, only a limited number of triplets are found. Building on these, we report the verb distribution of the limited number of triplets that are found, ranking the verbs in ascending order of Euclidean distance to the target verb, results of which are shown in Tab. 20. From this table, we can see that the verbs quantitatively closer (in Euclidean distance or Cosine distance) to *pay* have similar meanings to *pay*, shown by their lexical variants or grammatical variants (e.g., *putting money in* has a similar meaning to *pay*). Thus, in the VG dataset, there is *human putting money in parking meter*, which may transfer to the zero-shot recognition of *person pay parking meter* in HICO-DET. Similarly, we could see that in the VG datatset, there is *human putting condiments on hot dog*, which may transfer to the zero-shot recognition of *person cook hot dog* in HICO-DET. Note that there are some grammatical variations that we can not exhaustively set rules to avoid, thus creating a zero-shot setting which is not theoretically strict enough. But we want the model to benefit from this property of natural language in the context of language-image pre-training.

In short conclusion, with the assistance of the sequential inference structure of verbs, we think that the zero-shot inference ability in RLIP is not from the scale of annotations (by comparing relation detection pre-training and RLIP using VG), but the ability to transfer the verb inference knowledge from semantically similar annotations. This analysis also accords with previous papers [55, 38] that semantic diversity is important as it introduces large-scale potential annotations, ensuring a model transfers well to different data distributions.

Secondly, we aim to ***demonstrate quantitatively how RLIP pre-trains the model to perform zero-shot detection from the perspective of representation learning***. We resort to the Uniformity metric introduced in [63]. Uniformity is a metric to assess a model's generalization in contrastive learning. We detail the calculation of this metric in the analysis of the main paper. In this case, since label textual embeddings serve as a classifier in RLIP, we calculate the Uniformity of the seen verbs, unseen verbs and all verbs, aiming to observe how the generalization changes before and after RLIP, and how the generalization varies between seen verbs and unseen verbs. The results are shown in Tab. 21. As can be seen from the table, Uniformity values are all high before RLIP. It means that the

Table 20: VG verb ranking given similar subject-object triplets from HICO-DET. Verbs are in ascending order of Euclidean distance. (The Cosine distance can also output similar rankings.)

| **"pay"** ("parking meter") | **putting money in** | **collecting money at** | **puts change into** | **repairing** | **checking** | **next to** | ... |
|---|---|---|---|---|---|---|---|
| Count | 1 | 1 | 1 | 1 | 1 | 1 | ... |
| Euclidean | 11.56 | 11.70 | 13.34 | 14.21 | 15.16 | 16.12 | ... |
| Cosine | 0.4560 | 0.4576 | 0.3108 | 0.2554 | 0.1583 | 0.0709 | ... |

| **"cook"** ("hot dog") | **putting condiments on** | **prepping** | **displeased with** | **roasts** | **blowing on** | **about to eat** | ... |
|---|---|---|---|---|---|---|---|
| **Count** | 1 | 1 | 1 | 1 | 1 | 1 | ... |
| **Euclidean** | 13.38 | 14.14 | 15.48 | 15.63 | 16.04 | 16.27 | ... |
| **Cosine** | 0.3467 | 0.2565 | 0.0656 | 0.1787 | 0.0471 | 0.0680 | ... |

| **"grind"** ("skateboard") | **race downhill** | **flying off ramp on** | **skating** | **midair on** | **for balancing on** | **competes on** | ... |
|---|---|---|---|---|---|---|---|
| **Count** | 1 | 1 | 2 | | 1 | 1 | ... |
| **Euclidean** | 13.26 | 13.27 | 13.44 | 13.59 | 14.26 | 14.36 | ... |
| **Cosine** | 0.3670 | 0.3401 | 0.3553 | 0.3288 | 0.2510 | 0.2037 | ... |

| **"assemble"** ("kite") | **are preparing their** | **launched** | **managing** | **launch** | **carry** | **directing** | ... |
|---|---|---|---|---|---|---|---|
| **Count** | 1 | 1 | 2 | 1 | 1 | 2 | ... |
| **Euclidean** | 12.48 | 12.72 | 13.13 | 13.20 | 13.31 | 13.71 | ... |
| **Cosine** | 0.3350 | 0.3684 | 0.2974 | 0.2554 | 0.2027 | 0.2565 | ... |

| **"text on"** ("cell phone") | **viewing messages on** | **typing on** | **texting** | **speaking on** | **listening on** | **speaks on** | ... |
|---|---|---|---|---|---|---|---|
| **Count** | 1 | 1 | 2 | 1 | 1 | 2 | ... |
| **Euclidean** | 10.24 | 11.24 | 11.42 | 11.85 | 11.85 | 11.98 | ... |
| **Cosine** | 0.5876 | 0.4437 | 0.3983 | 0.4082 | 0.4281 | 0.3557 | ... |

| **"scratch"** ("dog") | **touching** | **bent over touching** | **touches** | **interacting with** | **holding a hot** | **reaching for** | ... |
|---|---|---|---|---|---|---|---|
| **Count** | 8 | 1 | 2 | 1 | 1 | 2 | ... |
| **Euclidean** | 14.02 | 14.47 | 14.54 | 14.60 | 14.80 | 14.81 | ... |
| **Cosine** | 0.2157 | 0.1566 | 0.2260 | 0.0969 | 0.0512 | 0.2039 | ... |

representations before RLIP are compactly distributed, serving as an awful classifier. However, after RLIP is performed, the seen 87 verbs have a distinctively lower Uniformity value, corresponding with the decent zero-shot performance. Similarly, the 30 unseen verbs and the combination of 117 verbs also have excellent Uniformity values, contributing to the unseen zero-shot performance. Through this quantitative observation, we think that from the perspective of representations, RLIP contributes to the real zero-shotness.

From all the above analysis, we think that the zero-shotness may not be caused by the mounting dataset size or annotations, but stem from the generalization in representations obtained by pre-training with language supervision.

Table 21: Uniformity analysis of the seen verbs, unseen verbs and all verbs before and after RLIP. Lower uniformity value is better.

| **Verb Set** | **Seen (87)** | **Unseen (30)** | **All (117)** |
|---|---|---|---|
| Before RLIP | -0.00367 | -0.00436 | -0.00388 |
| After RLIP | -3.73780 | -3.59457 | -3.71330 |

**The influence of semantic-diverse data (upstream data distributions)** We ablate semantic diversity by significantly altering the distribution of VG annotations and assessing the influence on RLIP's performance. To this end, first note that since VG is human-annotated with free-form text, it is extremely long-tailed. We alter its distribution by dropping tail object classes and verb classes to create a dataset with limited semantic diversity. Concretely, we drop object classes whose instance counts are fewer than 1,000 and relation classes whose instance counts are fewer than 500. We

pre-train RLIP on the resulting dataset and then perform zero-shot (NF) evaluation on HICO-DET. The results are shown in Tab. 22. We observe from this table that despite a very significant change to the training distribution, performance on the Full set drops only moderately. We do, however, witness a relatively larger decline on the Rare set due to the lack of semantic diversity in the modified data. This finding accords with the observations of GLIP [38]. To make full use of language-image pre-training, semantic diversity is important which can ensure a good domain transfer as is indicated by CLIP [55] and GLIP [38].

Table 22: Semantic diversity analysis with zero-shot (NF) evaluation on HICO-DET. Obj, rel and annos denote object, relation and annotations respectively.

| **Method** | **Data** | **Obj classes** | **Obj annos** | **Rel classes** | **Rel annos** | **Rare** | **Non-Rare** | **Full** |
|---|---|---|---|---|---|---|---|---|
| ParSeD | VG | 100,298 | 3.80m | 36,515 | 1.99m | 12.30 | 12.81 | 12.69 |
| ParSeD | VG- | 497 | 1.73m | 151 | 1.27m | 9.45 | 12.13 | 11.51 |

**More successful and failure cases analysis and corresponding potential future work** In this work, we present ParSe as an effective HOI detection structure. In Fig. 8 (a-d), we present several cases where the model successfully predicts by using RLIP-ParSeD (VG), with verb scores greater than 0.3. From Fig. 8 (a) and (b), we can see that although the scene is complex with many possible triplets with multiple labels, the model can detect the right subjects and objects, linking them as triplets. From Fig. 8 (c) and (d), we can observe that although the person is only partially visible, the model still detects him/her and then predicts the right triplets. Both cases show that RLIP-ParSeD can overcome difficult cases during application. However, there are also failure cases, where further works can improve upon. We show the failure cases from top-3 predictions produced by RLIP-ParSeD (VG) in Fig. 8 (e-h). **i)** First of all, the verb inference in ParSe conditions on the detection results since Sequential Relation Inference is fed with queries generated by detection features. Also, VG does not provide a good object detection foundation. Thus, an inferior detection result can lead to false positives. As shown in Fig. 8(f) and (g), the model detects the objects to be *wine glass* and *cell phone* (which should be *fire hydrant* and *handbag*), thus producing wrong predictions. Building on this observation, we think excellent object detectors can be designed and incorporated into HOI detection. **ii)** Secondly, DETR-based models may find it hard to detect fine-grained poses in people, which degrades the performance of inferring some verbs like *wave*. As shown in Fig. 8 (e) and (g), the model fails to predict *wave* although right localizing the subject person. Building on this observation, we think it to be promising to efficiently incorporate pose cues in end-to-end HOI detection models. **iii)** Thirdly, some contextual cues may be hard to be detected. As shown in Fig. 8 (h), the model predicts *ride* rather than *lasso*. If the model captures the rope as a global context, then it can perform well. Building on this observation, we think it to be promising to efficiently incorporate fine-grained contextual cues or introduce external object knowledge to detect the triplet.

### A.9 Datasets used in this work

**Licenses.** The V-COCO [15] dataset is used under an MIT license. The HICO-DET [5, 4] dataset is used under a CC0: Public Domain license. The Visual Genome [34] dataset is used under a Creative Commons Attribution 4.0 International License.

**Release of personally identifiable information/offensive content/consent.** We do not release data as part of this research. We work with standard public domain benchmarks for computer vision: Visual Genome [34], HICO-DET [5, 4], and V-COCO [15]. We therefore assess that the risk of releasing personally identifiable information or offensive content is relatively low. With regards to consent, we did not pursue an independent investigation of consent that goes beyond the considerations of the original dataset releases.

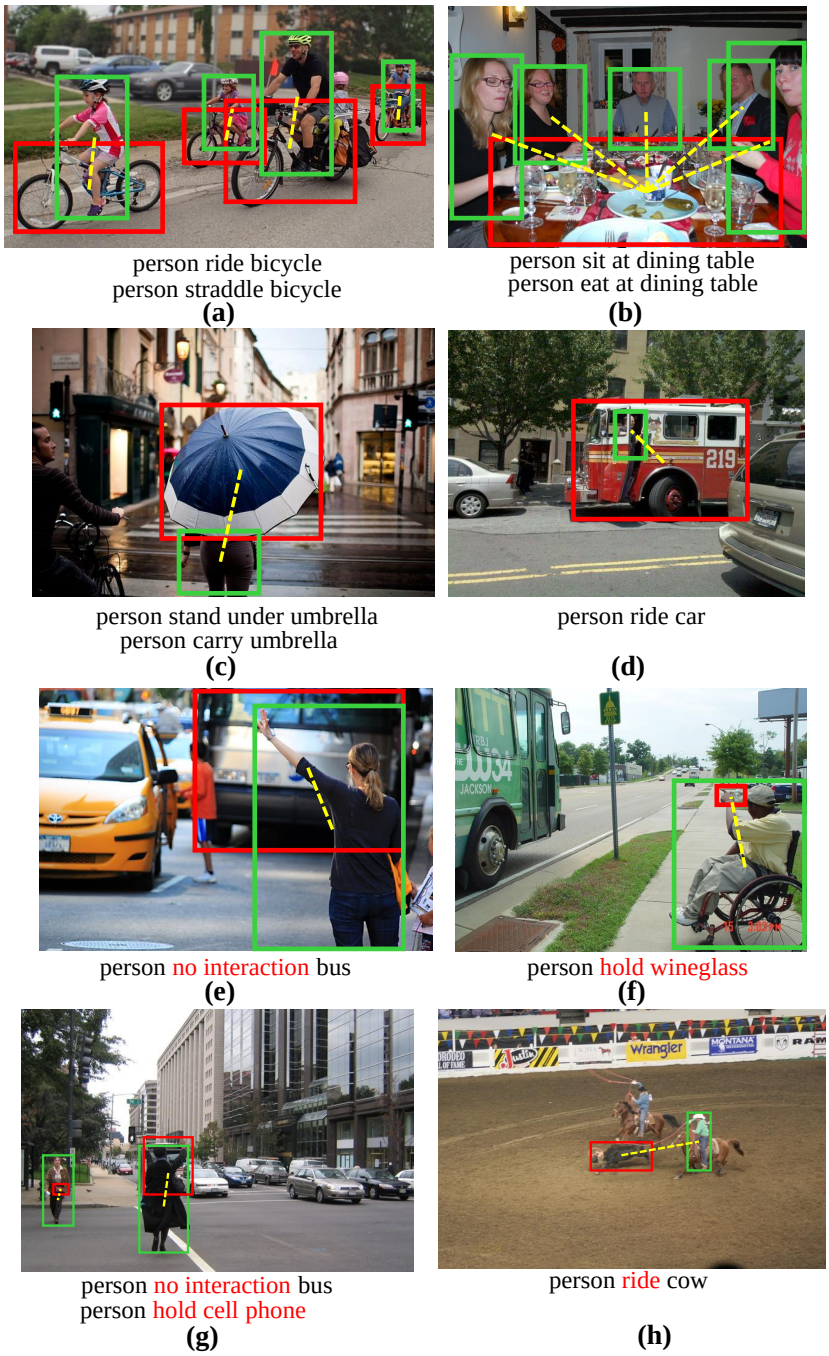


Figure 8: Successful and failure case analysis. We visualize the true positives (a-d) with verb scores higher than 0.3, and the false positives (e-h) from top-3 predictions produced by the RLIP-ParSeD (VG) model. Words in red indicate wrong predictions.