SCoRD: Subject-Conditional Relation Detection with Text-Augmented Data

Ziyan Yang[1]*, Kushal Kafle[2], Zhe Lin[2], Scott Cohen[2], Zhihong Ding[2], Vicente Ordonez[1]
[1]Rice University, [2]Adobe Research
{zy47,vicenteor}@rice.edu, {kkafle,zlin,scohen,zhding}@adobe.com

Abstract

We propose Subject-Conditional Relation Detection (SCoRD), where conditioned on an input subject, the goal is to predict all its relations to other objects in a scene along with their locations. Based on the Open Images dataset, we propose a challenging OIv6-SCoRD benchmark such that the training and testing splits have a distribution shift in terms of the occurrence statistics of ⟨subject, relation, object⟩ triplets. To solve this problem, we propose an auto-regressive model that given a subject, it predicts its relations, objects, and object locations by casting this output as a sequence of tokens. First, we show that previous scene-graph prediction methods fail to produce as exhaustive an enumeration of relation-object pairs when conditioned on a subject on this benchmark. Particularly, we obtain a recall@3 of 83.8% for our relation-object predictions compared to the 49.75% obtained by a recent scene graph detector. Then, we show improved generalization on both relation-object and object-box predictions by leveraging during training relation-object pairs obtained automatically from textual captions and for which no object-box annotations are available. Particularly, for ⟨subject, relation, object⟩ triplets for which no object locations are available during training, we are able to obtain a recall@3 of 33.80% for relation-object pairs and 26.75% for their box locations.

1. Introduction

Scene Graph Generation (SGG) in computer vision commonly refers to the task of predicting the class and locations for all possible objects along with the relations between them. A variety of methods have been introduced in the past years that attempt this task with increasingly accurate results [1, 5, 8, 9, 22, 30, 35, 45, 47–51]. However, predicting or annotating every possible ⟨subject, relation, object⟩ triplet for every subject in an image, especially under an open-vocabulary setting becomes increasingly impractical. In this work, we instead focus on subject-conditional relation detection (SCoRD) and show that we can generalize to an open-vocabulary setting through extra text annotations.

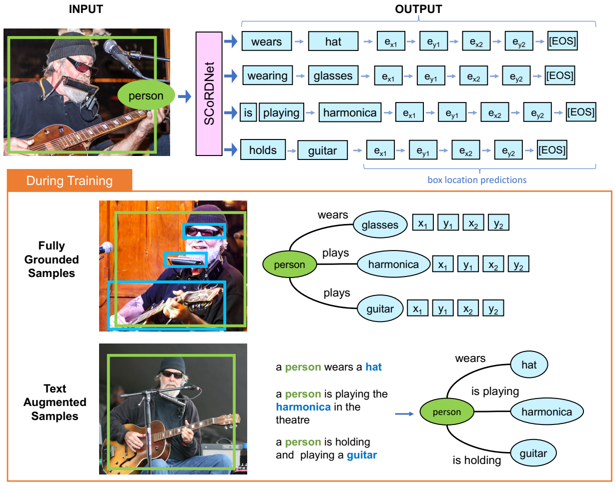

Figure 1. We cast subject-conditional relation detection (SCoRD) as a sequence decoding task where given an input subject in an image, we predict relation-object predicates and their locations. We also demonstrate how to leverage ungrounded training samples extracted from parsing textual captions. These samples are easier to obtain than fully grounded samples and have a potentially wider coverage of relation-object pairs.

Many pairwise relations between objects in a scene might be unnecessary for applications where we want to focus only on a given object of interest –a subject– and its predicates. Consider an image such as the one shown at the top of Figure 1, if we want to crop the person in this picture to create a composite in a different background, we might want to crop the person along with objects that the person interacts with, such as the guitar, the harmonica, the hat, and the glasses. We denote the task of finding objects and predicates that relate to a given subject as Subject-Conditional Relation Detection (SCoRD). We propose a challenging benchmark based on the Open Images dataset such that models trained on our provided training split can not easily take advantage of priors in the highly skewed distribution of ⟨subject, relation, object⟩ triplets. We provide a strong baseline and additionally demonstrate how we can tackle the prediction of relation-object pairs

[*] A portion of this work was done during Ziyan Yang's internship at Adobe Research.

that are unseen during training by leveraging an external dataset containing text annotations.

We propose SCoRDNet, a model that is trained to predict given an input image and a subject, a relation-object pair along with the object location auto-regressively. At test time, we use a Two-step decoding approach to generate a diverse set of relation-object pairs for a given test input image-subject pair. Figure 2 shows an overview of this model. SCoRDNet can leverage during training images annotated only with textual captions to improve its capabilites on rare or unseen ⟨subject, relation, object⟩ triplets. For example, even when our training data contains no annotations for *holding umbrella*, we can augment the training data with image-text data to provide an un-grounded example for *holding umbrella*; Due to our unified decoding, our model has the ability to not only predict but also provide accurate grounding in the form of a bounding box location for *holding umbrella* during inference. Figure 1 shows an overview of the inputs and outputs of our model during test time, as well as the two types of annotated samples that it can leverage during training: Fully grounded samples where relation-object pairs as well as object locations are provided, and text-augmented samples where the images are also annotated with image captions, which potentially mention a larger set of object categories than datasets that are fully annotated with ⟨subject, relation, object⟩ triplets.

Our contributions can be summarized as follows:

- We propose a new setting for relationship prediction, called Subject-Conditional Relation Detection (SCoRD), which consists of enumerating relations, objects, and boxes conditioned on an input subject in an image.
- We propose OIv6-SCoRD, a challenging benchmark that contains a training split and a set of testing splits, such that models trained on this dataset are not able to easily take advantage of the statistical distribution of ⟨subject, relation, object⟩ triplets.
- We propose SCoRDNet and conduct extensive experiments under different levels of supervision, demonstrating that our model can ground objects for new relation-object pairs with only additional text supervision. Moreover, we compare to prior general scene-graph generation (SSG) models by conditioning them on a subject and demonstrate the advantages of SCoRDNet.

2. Related Work

Our work is related to prior work on general scene-graph detection, and human-object interaction, and in terms of technical contributions to a few works on general object detection by sequence prediction and vision-language pre-training with grounded inputs.

Visual Relation Detection. In this task, the model is given an image and expected to produce outputs consisting of all the subject-predicate-object triplets along with their bounding boxes. There have been many works tackling this challenging problem [1, 5, 9, 30, 45, 47–51]. Many existing frameworks rely on two steps: First, using object detectors to extract candidate objects with bounding boxes; and then, proposing specific methods to predict relations between objects. A recent method, RelTR [8] uses a single stage by integrating an end-to-end object detector. Our work focuses instead on selective but exhaustive prediction of relation-object pairs conditioned on a subject. In order to compare with the previous literature, we use publicly available implementations of Neural Motif [48] and RelTR [8] and evaluate them in a subject conditioned setting. We compare against these representative one-stage and two-stage scene graph prediction methods by re-training our model under the same training conditions. The Human-Object Interaction Detection task (HOI-det) is a special case of Visual Relation Detection when the subjects are people [13–15, 18–21, 35, 41, 42]. Our SCoRD task also aims to detect related objects given an input subject but our subjects are not limited to instances of people.

Object Locations as Tokens. Our proposed SCoRDNet model leverages auto-regressive sequence prediction and predicts relation-object pairs and box locations by considering the box locations as additional discrete output tokens in a sequence. Some recent works have shown the effectiveness of encoding object box coordinates as discrete tokens [7, 39, 43, 44]. Tan et al [39] predicts box locations to compositionally generate a scene by discretizing the output space. Chen et al [7] considers object detection as a language modeling task by generating bounding box coordinates as text tokens conditioned on input images and previously predicted tokens. UniTAB [43] solves multiple tasks by using both text and box tokens as supervision. In the context, of vision-language pre-training, Yao et al [44] recently proposed PEVL which incorporates grounded input text in its training image-text pairs by encoding object box locations as discrete tokens. We build upon this idea to incorporate a decoding transformer that also predicts discrete object-location tokens.

Visual Grounding. These methods aim to localize an arbitrary input text given an input image and are often framed as referring expression comprehension [2, 16, 24, 31, 33, 36, 46]. Recent methods such as GLIP [28] and MDETR [23] have impressive general capabilities in their ability to localize arbitrary textual input and can be used as open vocabulary object detectors. Our work also aims to localize objects, however, our goal is to selectively detect objects that are related to a given subject, and also assign a label to the relation. In one of our experiments, we adopt GLIP to produce noisy annotations for input subjects in images that only have textual caption annotations and for comparing our method against a recent work that predicts relations but does not localize objects [32].

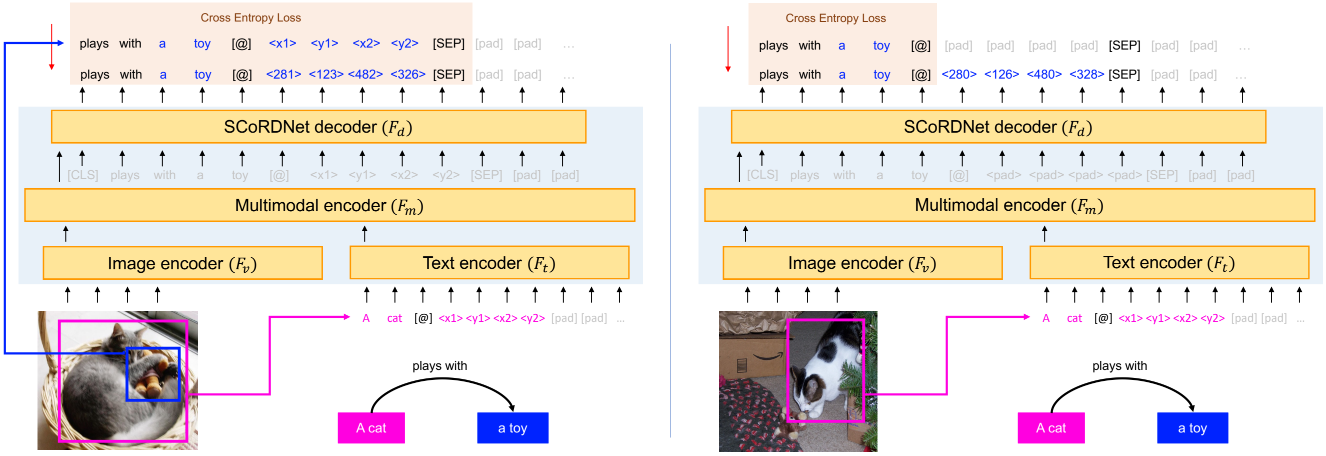
Cross Entropy Loss
plays with a toy [@] <x1> <y1> <x2> <y2> [SEP] [pad] [pad] …
plays with a toy [@] <281> <123> <482> <326> [SEP] [pad] [pad] …
SCoRDNet decoder (F_d)
[CLS] plays with a toy [@] <x1> <y1> <x2> <y2> [SEP] [pad] [pad]
Multimodal encoder (F_m)
Image encoder (F_v)
Text encoder (F_t)
A cat [@] <x1> <y1> <x2> <y2> [pad] [pad] ...
plays with
A cat
a toy
Cross Entropy Loss
plays with a toy [@] [pad] [pad] [pad] [pad] [SEP] [pad] [pad] …
plays with a toy [@] <280> <126> <480> <328> [SEP] [pad] [pad] …
SCoRDNet decoder (F_d)
[CLS] plays with a toy [@] <pad> <pad> <pad> <pad> [SEP] [pad] [pad]
Multimodal encoder (F_m)
Image encoder (F_v)
Text encoder (F_t)
A cat [@] <x1> <y1> <x2> <y2> [pad] [pad] ...
plays with
A cat
a toy

Figure 2. Here we show an overview of SCoRDNet. On the left we show how we handle fully grounded training samples consisting of a ⟨subject, relation, object⟩ triplet for which a box location is available for both the subject and the object. On the right, we show how we handle a training sample for which we only have a ⟨subject, relation, object⟩ triplet but no object location is available. In the second case, we do not backpropagate gradients through the prediction heads corresponding to object location coordinates although the model often makes reasonable predictions based on parameter sharing with other similar samples that are fully annotated.

3. Problem Definition

Let the inputs to our task be image I and target subject s along with its corresponding box location b_s. Our goal in this work is to predict a set of all possible relations, object, and object locations:

$$\{\langle r, o, b_o\rangle \text{ s.t. } \exists\langle s, r, o\rangle \in \Pi_I\}, \tag{1}$$

where Π_I is an enumeration of all valid relation triplets in input image I, and b_o is the box location corresponding to object o. In this work, we propose a model Φ such that we can sample output relation-object predictions from a model trained to predict a distribution over all possible values as follows:

$$\langle r, o, b_o\rangle \sim \Phi(I, s, b_s). \tag{2}$$

By sampling multiple times from this model, we can obtain an arbitrary set of predictions $\{\langle r_j, o_j, b_{o_j}\rangle\}$. For reference, scene graph generation methods aim to generate a full list of subject, relation, objects along with their locations, i.e. $\langle s, b_s, r, o, b_o\rangle$. However, the general scene graph generation (SGG) setup often limits the number of object categories for s, r, and o. In our setup, s, r and o are open-vocabulary and modeled as text, so they can potentially support an arbitrary set of objects and relations.

4. Benchmark

We construct a test benchmark for subject-conditional relation detection based on the Open Images v6 (OIv6) dataset [26] which contains fully grounded ⟨subject, relation, object⟩ triplets for a diverse array of object categories. We select training and testing splits so that models trained on this benchmark can not rely on the highly skewed statistics of naturally occurring ⟨subject, relation, object⟩ triplets.

4.1. Data

We use two types of supervision: First, fully-grounded data that contains ⟨subject, relation, object⟩ triplets with corresponding box locations for both subjects and objects, and a second type of samples that contain noisy ⟨subject, relation, object⟩ triplets extracted automatically from textual image captions but for which no object locations are available (ungrounded samples).

Fully-grounded Data: We use a combination of Visual Genome [25], Flickr30k [33] and Open Images V6 (OIv6) [26]. Visual Genome has 100k images annotated with 2.3M relations between objects with annotated box locations. Flickr30k contains 30k images with bounding boxes annotated with corresponding referring expressions. We use a language dependency parser [17] to extract ⟨subject, relation, object⟩ triplets from these expressions, and map them to phrase bounding boxes. Open Images has 526k images with annotations for subjects, objects, relations and boxes. However, many of the relation-object pairs are relations of the type "A is B" which is not a true relation between two different objects. We filter out images with such relations and end up with 120k images. In the following sections, *fully grounded data (FG)* is used to refer to any sample from the three sets described in this section.

Text-Augmented Data: While *fully grounded data* is scarce and hard to annotate, it is considerably easier to obtain images annotated with textual captions. We aim to show how automatically extracted ⟨subject, relation, object⟩ triplets from captions can help on this task even when no box annotations are provided. We use two data sources: COCO [29] and Conceptual Captions (CC3M [38] + CC12M [4] referred together as CC). COCO includes

120k images, 5 captions for each image, and bounding box annotations for 80 categories. Despite having box annotations, the COCO dataset does not contain a mapping between object annotations and references to objects in text descriptions. Moreover, object boxes are only annotated for 80 object categories, which does not cover the full range of objects mentioned in captions. Hence, we use COCO as an *ungrounded* data source. Additionally, we augment our use of COCO with Localized Narrative descriptions [34] to get more varied subjects, objects and relations. This combination of ungrounded image-text data from COCO and CC is meant to provide additional supervision from an external source. Using image-text pairs, we can extract ⟨subject, relation, object⟩ triplets with a dependency parser [17]. We also generate noisy box locations for the input subjects in these triplets using GLIP [28] but leave target objects ungrounded without any corresponding box locations.

4.2. Test-Train Splits

Our test splits are all based on the Open Images v6 validation and testing sets which contain 7,000 images with 20,000 ⟨subject, relation, object⟩ triplets. Evaluating visual relation prediction tasks is challenging as some relations such as ⟨sitting on, chair⟩ are overly represented in the dataset to an extent that a model could be relying entirely on dataset statistics. To avoid models from taking advantage of the highly skewed distribution of ⟨subject, relation, object⟩ triplets and in order to test whether models can generalize to under-represented ⟨subject, relation, object⟩ triplets, we devise two test splits designed to under-sample the triplets included in the training splits. First, we select all relation-object pairs from the parsed captions obtained from COCO and CC that occur between 100 and 20,000 times. Among those, 104 relation-object pairs also appear in the test split of Open Images V6. We divide these unique 104 unique relation-object pairs into two non-overlapping sets which we denote as *Rel-Obj Set A*, and *Rel-Obj Set B*, each with 52 unique relation-object pairs. Table 1 shows a detailed count of how many samples exist for each of these relation-object pairs across all our data sources. Next, we describe our testing and training splits:

Test A: This test split consists of ∼4k samples belonging to any of the ⟨subject, relation, object⟩ triplets covered by the 52 unique relation-object pairs in the *Rel-Obj Set A* set.

Test B: This test split includes ∼4k samples belonging to any of the ⟨subject, relation, object⟩ triplets covered by the 52 unique relation-object pairs in the *Rel-Obj Set B* set.

Full Test: This test split includes the full test set samples of OIv6 containing ∼20k fully grounded samples ⟨subject, relation, object⟩ triplets.

Base Training Split: Consists of fully grounded training samples. However, we remove half of the samples belonging to relation-object pairs present in *Rel-Obj Set A*: ∼11.5k from VG, ∼2.5k from Flickr30k and ∼34k from OIv6 (Table 1). Our goal is to skew the distribution between the training and test splits so that the relation-object pairs present in the test split of *Test A* are underrepresented in the training split. We additionally remove all the samples that include triplets containing relation-object pairs present in *Rel-Obj Set B*: ∼37k from VG, ∼5k from Flickr30k and ∼104k from OIv6, (Table 1). This will make *Test B* challenging since the relation-object pairs in this part of the test set will not be present in this training split.

Source	**Obj-Loc**	**Rel-Obj Set A**	**Rel-Obj Set B**	**Full Rel-Obj Set**
VG [25]	✓	23k	37k	2M
Flickr30k [33]	✓	5k	5k	277k
OIv6 [26]	✓	68k	104k	350k
COCO [6, 29]*	-	49k	36k	85k
CC [4, 38]*	-	38k	56k	94k
OIv6 (test) [26]	✓	4k	4k	20k

Table 1. **Rel-Obj Set A** focuses on a set of 52 unique relation-object pairs, and **Rel-Obj Set B** focuses on a non-overlapping set of 52 unique pairs. This table shows how many individual samples exist in each of the data sources we use to construct our training and test splits for each of these relations. (*) For ungrounded text-augmented data sources we only consider samples included in the 104 unique relation-object pairs in Test A and B.

Text-Augmented Training Split: This training split includes all the fully grounded samples in the base training split and additionally includes ungrounded samples that were automatically annotated from textual captions: ∼85k from COCO and ∼94k from CC. These samples were obtained by filtering COCO and CC to target the 104 relation-object pairs targeted by both *Test A* and *Test B*. The goal is to demonstrate compositional generalization from ungrounded samples to grounded but underrepresented (Test A), and unseen ⟨subject, relation, object⟩ triplets (Test B).

5. Method

In this section, we describe our proposed SCoRDNet model. Figure 2 shows an overview of our model, as used in two different training modes of supervision. Our model consists of four transformer models, an image encoder, an input text encoder, a multimodal fusion encoder, and an output auto-regressive decoder. Here we describe these components along with our proposed decoding process.

5.1. Model

We cast the problem of subject-conditional relation detection as a sequence-to-sequence model where the input is an image I coupled with an input token sequence representing the input subject s and its box location b_s, and the output is a sequence of tokens representing the predicted relation-object pair $\langle r, o \rangle$ and the corresponding object location coordinates b_o. The left part of Figure 2 shows an overview of this process for a given input fully grounded training sample. In order to cast the input and output coordinates for the grounded samples as tokens, it is necessary to discretize them. For a given set of box coordinates (x_1, y_1, x_2, y_2) and a corresponding image with height h and width w, the goal is to discretize the coordinates as:

$$(P\frac{x_1}{w}, P\frac{y_1}{h}, P\frac{x_2}{w}, P\frac{y_2}{h}), \tag{3}$$

with a pre-defined number P representing the total number of position tokens. Additionally, we add special separator tokens to indicate the start and the end of box coordinates.

SCoRDNet consists of an image transformer encoder F_v which encodes images into a sequence of image features $\{I_i\}_{i=1}^{N_I}$, and a text transformer encoder F_t which encodes subjects with position tokens as a sequence of text features $\{T_j\}_{j=1}^{N_T}$. Then, a multimodal fusion encoder F_m will fuse image and text features through cross-attention layers and produce a context vector z. Finally, the output context vector z of the multimodal fusion encoder F_m is forwarded to a decoder transformer F_d which is trained auto-regressively to predict a relation, an object, and its corresponding box coordinate locations as a sequence of tokens.

Our model is trained to predict a context vector $z = F_m(F_v(I), F_t(s, b_s))$, which is then used to predict an output relation-object pair along with a bounding box using the auto-regressive transformer decoder $\langle r, o \rangle, b_o = F_d(z)$. During training, the model is optimized to minimize a loss function $\mathcal{L}(\langle r, o \rangle, b_o, F_d(z))$ that aims to produce a sequence from the transformer decoder that matches the true relation-object pair. During inference, predictions are obtained by sampling from the transformer decoder: $\langle r, o \rangle, b_o \sim F_d(z)$. This same model can be trained with ungrounded relation-object pairs for which a set of object box coordinates is not available by simply optimizing $\mathcal{L}(\langle r, o \rangle, F_d(z))$ for samples that do not contain a box corresponding to the object in the relation. However, our model can still be used to sample a full length sequence containing a relation-object pair as well as the object location. We show on the right part of Figure 2 an overview of how this process works. During training, we can simply skip computing the loss terms for input tokens that are not provided in the ground truth output sequences (e.g., boxes).

Loss: We use a standard cross-entropy loss to train our model. Let $\mathrm{H}(\cdot, \cdot)$ be the cross-entropy function, then we compute the cross-entropy between the current target token and generated token conditioned on the input image, subject sequence, and previous tokens as follows:

Algorithm 1 Two-stage decoding

```
Input: $I$: Input image
Input: $\langle s, b_s \rangle$: Input subject, and subject box
Input: $K$: Number of desired output relation-object pairs
Output: $\{\langle r_k, o_k \rangle\}, b_{o_k}$: Output relation-object pairs and boxes
function BEAMSEARCH($F, z, \{t_k\}, K$, [EOS])
    return top K sequences from $F(z, \{t_k\})$ ending in [EOS]
end function
$z \leftarrow F_m(F_v(I), F_t(\langle s, b_s \rangle))$
$\{\langle r_k, o_k \rangle\} \leftarrow$ BEAMSEARCH($F_d, z, \emptyset, K$, [@])
for $\langle r_k, o_k \rangle$ in $\{\langle r_k, o_k \rangle\}$ do
    $b_{o_k} \leftarrow$ BEAMSEARCH($F_d, z, \langle r_k, o_k \rangle, 1$, [SEP])
end for
```

$$z_i = F_m\left(F_v(I_i), F_t(\langle s_i, b_{s_i} \rangle)\right) \tag{4}$$

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{k} \mathrm{H}\left(t_k, F_d\left(t_{0:k-1}, z_i\right)\right), \tag{5}$$

where N is the number of training samples, I_i is the input image, $\langle s_i, b_{s_i} \rangle$ is an input subject and subject-box encoded as a sequence of tokens, and t_k is a token at a given time step k for ground truth annotation $\langle \langle r_i, o_i \rangle, b_{o_i} \rangle$ encoded as a sequence of tokens. For ungrounded samples these tokens t_k are encoding only the ground truth annotation $\langle r_i, o_i \rangle$.

5.2. Two-step Decoding

We could sample output relation-object pairs along with object-locations using beam search directly from our decoder such that $\{\langle r, o \rangle\}, b_o \sim F_d$. However, we find that this leads to a lack of diversity in the predicted relation-object pairs due to the disparity in the vocabulary for relation-object tokens and box location tokens. Hence, we rely on a Two-step decoding process where first a diverse set of K relation object pairs $\{\langle r_k, o_k \rangle\}$ are decoded, and then a corresponding set of object boxes b_{o_k} are decoded conditional on the relation-object pairs. This two-step decoding process is described in Algorithm 1, where we rely on a beam search function that decodes a sequence of K tokens given an input decoder F, conditional state vector z, and a partially generated output sequence t_k. Additionally, this function takes as a parameter a custom end-of-sequence token. First, we perform beam search to decode relation-object pairs until we encounter the end of sequence token [@] which we use to separate relation-object pairs and object-location box coordinates, and then we decode box locations one by one conditioned on the same input but with a partially generated output sequence comprised of the relation-object pair. We decode until finding the end-of-sequence token [SEP].

		Rel-Object		Object-Loc	
Split	**Model**	*R@1*	*R@3*	*R@1*	*R@3*
Test A	Base	39.94	82.19	29.08	57.06
	Text-aug.	43.61	83.91	31.75	58.58
Test B	Base	0.74	4.78	0.67	4.36
	Text-aug.	18.71	33.80	15.14	26.75
Full Test	Base	49.84	69.22	36.54	50.37
	Text-aug.	53.48	75.51	39.36	55.38

Table 2. Results for our three test splits in our benchmark. The base model was trained while removing 50% of training samples with triplets in *Test A* and by removing all samples for triplets present in *Test B*. These results highlight how our Text-augmented model is able to take advantage of ungrounded samples to generalize to both unseen and under-sampled relation-object pairs, not only for predicting their correct relation but also object locations.

6. Experiment Settings

Implementation Details. For the image encoder F_v, the text encoder F_t and the multimodal fusion encoder F_m we adopt the architecture and pre-trained encoders of PEVL [44], which is a joint vision-language transformer which is in turn based on the ALBEF model [27] but using additional grounded supervision with box coordinates as input tokens. Our transformer-based model contains one 12-layer vision transformer [12] as the image encoder, one 6-layer text transformer [11, 40] as the text encoder, one 6-layer multimodal transformer encoder to combine image and text information and one 6-layer transformer decoder to generate target sequences. The hidden size for each layer is 768 and the number of attention heads is 12. The decoder follows the same architecture as the multimodal transformer encoder but using masked self-attention layers, and its parameters are initialized using the weights from the multimodal transformer encoder. Code and data are available[1].

Evaluation metrics. We report Recall@K for all the experiments. For each subject and its bounding box, we keep different numbers of sequences from beam search – which is our mechanism for decoding. For example, Recall@3 indicates we keep 3 relation, object, object-boxes with the highest scores. We split the outputs into two parts: relation-object (Rel-Object) and object-location (Object-Loc). If one predicted text part among K returned sequences belongs to the ground truth text part or its synsets, we count this sample as positive for text evaluation. If one predicted text is "correct", we keep looking at the predicted box, and if the predicted box and ground truth box have IoU ≥ 0.5, this sample is counted as positive for the bounding box part.

		Rel-Object		Object-Loc	
IoU	**Methods**	*R@1*	*R@3*	*R@1*	*R@3*
0.5	RelTR [8]	29.71	49.75	24.64	42.92
	SCoRDNet	65.60	83.80	24.88	40.81
	Motif [48]	25.83	41.79	23.22	37.05
	SCoRDNet	65.17	82.31	23.38	39.97
0.4	RelTR [8]	29.71	49.75	26.64	44.84
	SCoRDNet	65.60	83.80	30.71	49.02
	Motif [48]	25.83	41.79	24.17	38.55
	SCoRDNet	65.17	82.31	30.33	49.37
0.3	RelTR [8]	29.71	49.75	27.18	46.03
	SCoRDNet	65.60	83.80	35.47	55.70
	Motif [48]	25.83	41.79	29.72	39.73
	SCoRDNet	65.17	82.31	34.60	55.61

Table 3. Comparison with general scene-graph generation models on OIv6-SCoRD. SCoRDNet compares favorably against the strong Motif model [48] and the recently proposed RelTR [8]. For this experiment all methods were trained on Visual Genome [25].

7. Results and Discussion

Table 2 shows the main results from our experiments using the training and testing splits defined in Section 4. The base model is trained on the *Base Training Split*. For our Text-augmented model results, we train another model using the *Text-Augmented Training Split*. We report our results on *Test A*, *Test B*, and *Full Test* with R@1 and R@3 for both predicted relation-object pairs and object-location coordinates. The results in Table 2 show considerable improvements to the overall prediction of relation-object pairs and object-locations using our text-augmented model across all tests. These results demonstrate the usefulness of adding text-augmented samples even if they do not have object locations. Suppose that our training set does not have many training samples for *cutting orange*, which causes our model to be weaker for this pair. However, our model has already learned how to localize oranges from relation-object pairs such as *eating orange* or *holding orange*. Any inability to predict *cutting orange* is due solely to not seeing enough examples for this relation-object pair. Hence, by simply adding more image-text data to show images with *cutting orange*, even if we do not have fully grounded samples, the model can quickly learn to predict this association.

Next, our results for the relation-object pairs in *Test B* are a lot better than the base model. The base results are understandably quite low since this model has not seen any example for relation-object pairs present in *Test B*. However, it is remarkable how our model regains the ability to not only predict these pairs but also ground them by solely

[1] https://github.com/uvavision/SCoRD

relying on text-augmented data for these classes. We reiterate that our model has *never seen* any grounded data for these relation-object pairs. Hence, the results come solely from compositional transfer from other pairs and the weak signal from our text augmentation. While our results already show a huge improvement via text augmentation, we want to emphasize that the augmentation is solely coming from out-of-domain image sources compared to our test set, which makes it even more attractive. However, text augmentation done from an in-domain source image provides even higher performance. We show an experiment comparing in-domain vs out-of-domain ungrounded sample augmentation in the supplementary materials.

7.1. Comparisons to existing work.

We have chosen to use a unique input-output paradigm for relationship detection and grounding, which as explained in Section 5, offers many benefits and offers a more convenient setup for collecting more data in the future. This setup also allows easy use of image-text data for almost limitless expansion. However, our setup also makes it relatively difficult to put our contributions in the context of existing work. Here we show that our method is actually comparable to scene graph generation methods and other related methods when using the same data sources.

Comparisons to Scene Graph Generation. To compare with scene graph generation (SGG) methods, we keep the same 104 relation-object pairs from our *Rel-Obj Set A* and *Rel-Obj Set B* (since we already have parsed results for text-based augmentation for these). Then, we obtain predictions from both our method and existing work on this set and measure R@K for both relation-object pairs and object-locations. However, there are some key differences between our model compared with existing SGG methods. Hence, we have made a few changes in the evaluation to allow for a fair comparison of our model. There are two main concerns: (1) *Our method gets the subject box as input whereas SGG methods do not:* To equalize for this, we only consider predictions from samples where SGG methods correctly predict the subject box. So, for these samples, both methods have the same level of information; (2) *Our method is trained and evaluated on OIv6, whereas SGG methods are only trained on VG:* For this experiment, we trained a model variant on a dataset that excludes all training samples from OIv6 and we only use grounded data from VG + ungrounded expansion from COCO+CC. This reduces the total possible relation-object pairs from 104 to just 28; This will be our test bed for this experiment.

The results of this experiment are shown in Table 3 where we show considerably better results in almost all settings compared to the popular Neural Motif [48] model for scene graph prediction as well as the recent detector-free model RelTR [8]. Since our method is not trained to predict tight bounding boxes using box proposal networks or additional box refinement objectives, we report numbers at varying levels of IOU instead of just a default value of 0.5, showing that a potential combination of our method with box refinement techniques could potentially lead to more improvements.

Methods	**Rel-Object**		**Object-Loc**	
	R@1	R@3	R@1	R@3
TAP [32] + GLIP	16.31	29.18	11.64	21.31
SCoRDNet (Ours)	63.16	91.53	49.50	71.46

Table 4. We compare our results against a specialized attribute prediction model combined with an open-vocabulary object detection model (GLIP [28]) on 100 relation-object pairs in from the Full Test set that are in the vocabulary for TAP.

We observe that our relation-object pair prediction is almost twice as good compared to both of these methods, and at a lower IOU threshold, our method far outperforms box prediction as well. This shows that our method has learned to use text supervision effectively for relation grounding (even if at a coarser grounding threshold, such as IOU $<$ 0.5). While there is some room for improvement in our model in terms of refining the box predictions to better adhere to object boundaries, we believe that these results already show comparable, if not better, results for our model even when compared fairly to traditional scene graph methods. In addition, our decoder-based unification comes with many other benefits as described earlier.

Other Comparisons. Another closely related work is the recently proposed TAP [32] model for attribute prediction. This model is trained to predict related attributes for a given query of object names and boxes which is similar to our setup. This work also considers relation-object pairs as attributes of an object, so it is possible to obtain relation-object pair predictions from this model. By combining TAP with GLIP [28], we can obtain both text and box prediction. If GLIP generates multiple bounding boxes for one object, we keep the box closest to the ground truth subject bounding box. We find that TAP can predict 100 relation-object pairs in the Full Test set, and we report results on these 100 relation-object pairs in Table 4. We find that in both comparison settings, our work can far outperform this potential solution in both text and box predictions. The TAP model was generously provided by the original authors in order to conduct this experiment.

7.2. Ablation Experiments

Effect of adding more image-text data. Our earlier results have already shown that adding text data can help improve prediction and grounding for both previously seen and unseen relation-object pairs. Here we want to explore

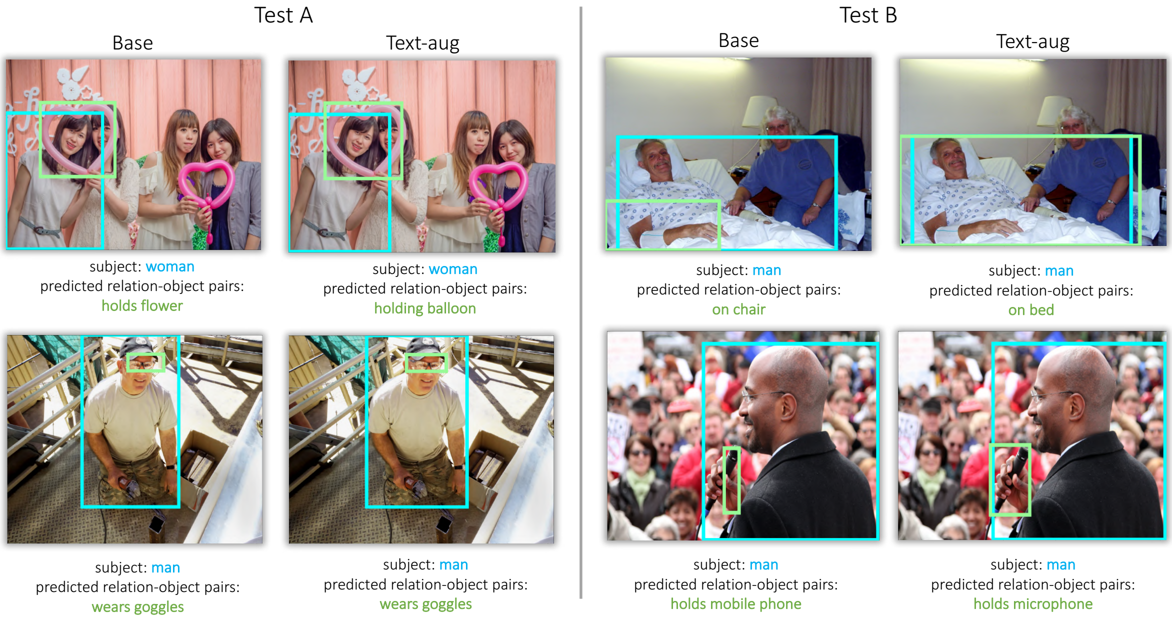
Test A
Base
Text-aug
subject: woman
predicted relation-object pairs:
holds flower
subject: woman
predicted relation-object pairs:
holding balloon
subject: man
predicted relation-object pairs:
wears goggles
subject: man
predicted relation-object pairs:
wears goggles
Test B
Base
Text-aug
subject: man
predicted relation-object pairs:
on chair
subject: man
predicted relation-object pairs:
on bed
subject: man
predicted relation-object pairs:
holds mobile phone
subject: man
predicted relation-object pairs:
holds microphone

Figure 3. Qualitative results from the models trained with and without text augmentation from COCO and CC. "Base" indicates results generated by the model trained with 50% training samples for the relation-object pairs in *Rel-Obj Set A* and no samples for the relation-object pairs in *Rel-Obj Set B*. "Text-aug." indicates results generated by the model trained with Text-Augmented Training Split.

how adding more data over time might affect this. To do this, we set our evaluation split to only contain the samples from Test A + Test B. Then, our base training split is constructed so that all the training samples of the 104 relation-object pairs belonging to *Rel-Obj Set A* and *Rel-Obj Set B* are removed. We refer to this ablation training set as ATS. Then, we incrementally add image-text augmentation with COCO and CC15M. The results are presented in Table 5, which clearly shows that the inclusion of more data has additive effects. This suggests that our model can very effectively scale to include new relation-object pairs that do not have any grounded data during training, and we can anticipate it to continue getting better with more samples from vast collections of image-text pairs on the internet. For example, resources such as LAION-2B and LAION-400M [37], COYO-700M [3], and RedCaps [10] can potentially be used which can add hundreds of millions of additional text-augmentation.

Additional ablation studies are included in the supplementary material such as studying the effect of removing more samples from the relation-object pairs from *Rel-Obj Set A*. Currently we remove 50% of the samples but we show that our results hold when removing less 25% or more samples 75%. Finally, we show in Figure 3 some qualitative results for relation-object pairs and object-locations predicted by our models with and without text augmentation.

8. Conclusion

The SCoRD task combined with generation-based relation detection offers several advantages. To expand the number and variety of relation-object pairs, we can simply provide sparse box annotations for a small amount of additional data, and combine them with a larger amount of ungrounded image-text data to obtain the same benefits as annotating a lot of expensive box annotations. In summary, we proposed subject-conditional relation detection (SCoRD) which is an alternative way to obtain a relation graph given a query object. We also propose a new transformer-based model which allows us to effectively tackle this task. SCoRD, combined with our method allows us to easily utilize image-text paired data to further enhance relation prediction, even when there are no annotations present in the base training set for these relationships.

Data	**Rel-Object**		**Object-Loc**	
	R@1	R@3	R@1	R@3
ATS	0.64	3.87	0.62	2.82
ATS + COCO	8.74	20.24	6.74	13.95
ATS + COCO + CC	20.96	37.59	12.20	20.74

Table 5. Effect of additional text augmentation from image captions, starting from a base training on the Ablation Training Set (ATS) (See Sec. 7.2). Considerable gains are obtained by text-augmented data without any additional box information.

Acknowledgments This work was partially supported by gifts from Adobe Research and NSF Awards IIS-2201710 and IIS-2221943.

References

[1] Sherif Abdelkarim, Aniket Agarwal, Panos Achlioptas, Jun Chen, Jiaji Huang, Boyang Li, Kenneth Church, and Mohamed Elhoseiny. Exploring long tail visual relationship recognition with large vocabulary. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15921–15930, 2021. 1, 2

[2] Assaf Arbelle, Sivan Doveh, Amit Alfassy, Joseph Shtok, Guy Lev, Eli Schwartz, Hilde Kuehne, Hila Barak Levi, Prasanna Sattigeri, Rameswar Panda, et al. Detector-free weakly supervised grounding by separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1801–1812, 2021. 2

[3] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022. 8

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 3, 4

[5] Jun Chen, Aniket Agarwal, Sherif Abdelkarim, Deyao Zhu, and Mohamed Elhoseiny. Reltransformer: Balancing the visual relationship detection from local context, scene and memory. *arXiv preprint arXiv:2104.11934*, 2021. 1, 2

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

[7] Zhiyang Chen, Yousong Zhu, Zhaowen Li, Fan Yang, Wei Li, Haixin Wang, Chaoyang Zhao, Liwei Wu, Rui Zhao, Jinqiao Wang, et al. Obj2seq: Formatting objects as sequences with class prompt for visual tasks. *arXiv preprint arXiv:2209.13948*, 2022. 2

[8] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. Reltr: Relation transformer for scene graph generation. *arXiv preprint arXiv:2201.11460*, 2022. 1, 2, 6, 7

[9] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE conference on computer vision and Pattern recognition*, pages 3076–3086, 2017. 1, 2

[10] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021. 8

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 6

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[13] Hao-Shu Fang, Yichen Xie, Dian Shao, and Cewu Lu. Dirv: Dense interaction region voting for end-to-end human-object interaction detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1291–1299, 2021. 2

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

[15] 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. 2

[16] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive learning for weakly supervised phrase grounding. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 2

[17] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python. 2020. 3, 4

[18] 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. 2

[19] 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. 2

[20] 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. 2

[21] Zhi Hou, Baosheng Yu, and Dacheng Tao. Discovering human-object interaction concepts via self-compositional learning. *arXiv preprint arXiv:2203.14272*, 2022. 2

[22] 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. 1

[23] 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. 2

[24] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 2

[25] 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. 3, 4, 6

[26] 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. 3, 4

[27] 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*, 34:9694–9705, 2021. 6

[28] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2, 4, 7

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3, 4

[30] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016. 1, 2

[31] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pages 792–807. Springer, 2016. 2

[32] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Improving closed and open-vocabulary attribute prediction using transformers. In *European Conference on Computer Vision*, pages 201–219. Springer, 2022. 2, 7

[33] 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. 2, 3, 4

[34] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 4

[35] 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)*, pages 401–417, 2018. 1, 2

[36] Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440, 2020. 2

[37] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 8

[38] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 3, 4

[39] Fuwen Tan, Song Feng, and Vicente Ordonez. Text2scene: Generating compositional scenes from textual descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6710–6719, 2019. 2

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 6

[41] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2022. 2

[42] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. Deep contextual attention for human-object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5694–5702, 2019. 2

[43] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer, 2022. 2

[44] Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. PEVL: Position-enhanced pre-training and prompt tuning for vision-language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11104–11117, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. 2, 6

[45] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 322–338, 2018. 1, 2

[46] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018. 2

[47] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *Proceedings of the IEEE international conference on computer vision*, pages 1974–1982, 2017. 1, 2

[48] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840, 2018. 1, 2, 6, 7

[49] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual

relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5532–5540, 2017. 1, 2

[50] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11535–11543, 2019. 1, 2

[51] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid. Towards context-aware interaction recognition for visual relationship detection. In *Proceedings of the IEEE international conference on computer vision*, pages 589–598, 2017. 1, 2