Scene-Graph ViT: End-to-End Open-Vocabulary Visual Relationship Detection

Tim Salzmann^{†1,2}, Markus Ryll², Alex Bewley^{*1}, and Matthias Minderer^{*1}

¹ Google DeepMind
{tsal, bewley, mjlm}@google.com
² Technical University Munich
{tim.salzmann, markus.ryll}@tum.de



Figure 1: Relationships detected by our method on an unseen image. The top relationships by confidence score are shown. Photo by Spacejoy on Unsplash.

Abstract. Visual relationship detection aims to identify objects and their relationships in images. Prior methods approach this task by adding separate relationship modules or decoders to existing object detection architectures. This separation increases complexity and hinders end-to-end training, which limits performance. We propose a simple and highly efficient decoder-free architecture for open-vocabulary visual relationship detection. Our model consists of a Transformer-based image encoder that represents objects as tokens and models their relationships implicitly. To extract relationship information, we introduce an attention mechanism that selects object pairs likely to form a relationship. We provide a single-stage recipe to train this model on a mixture of object and relationship detection data. Our approach achieves state-of-the-art relationship detection performance on Visual Genome and on the large-vocabulary GQA benchmark at real-time inference speeds. We provide ablations, real-world qualitative examples, and analyses of zero-shot performance.

[†]Work done while at Google DeepMind.

^{*}Advising project leads in alphabetical order.

2 T. Salzmann et al.

1 Introduction

A fundamental goal of computer vision is to decompose visual scenes into structured semantic representations. A commonly studied task towards this goal is object detection, in which objects in an image are localized by bounding boxes and classified into semantic categories. However, a scene description also includes the semantic *relationships* between objects. This gives rise to the task of visual relationship detection (VRD, [4, 6, 32, 48, 49, 55]). In VRD, the model detects objects and infers pairwise relationships between them in the form of <subject-predicate-object> triplets³ [32].

Detecting both objects and their relationships allows the construction of a *scene graph* [19, 30, 48] in which objects are represented as nodes and their relationships as edges. Scene graph generation (SGG) has wide-ranging applications in robotics [1,5,11,17,38,42] and image retrieval [19,27]. Increasingly, structured scene representations are also employed to provide grounding and explainability to multimodal large language models [15, 36, 59].

Prior work typically draws a distinction between object detection and relationship prediction. Detection is performed either as a wholly separate step before relationship prediction [8, 32, 48, 52, 55, 57], or by separate model parts such as "relationship decoders" that are responsible for modeling the interactions between objects [7,49,61]. This distinction makes it hard to optimize such models end-to-end for VRD [61]. In contrast, we propose an encoder-only architecture that models objects and relationships jointly, directly in the image encoder. Our architecture performs open-vocabulary relationship detection and can be trained end-to-end on arbitrary mixtures of object detection and relationship annotations.

Our model builds on a Transformer-based encoder-only object detector [34] in which the output tokens of the image encoder directly represent object proposals. From these tokens, class embeddings and bounding boxes are decoded with lightweight heads. Our insight is that this architecture is perfectly set up to learn relationships between objects directly in the image encoder, without the need for additional relationship-specific stages. This is because the existing self-attention in the encoder already models all-to-all pairwise interactions between the object proposal tokens.

To access information about the relationship between two of these tokens, we combine the embeddings corresponding to the <subject> and <object> token using a new *Relationship Attention* layer. Obtaining relationship embeddings for all possible pairwise combinations of object proposal tokens would be computationally infeasible. To reduce the number of combinations, we introduce a self-supervised hard attention mechanism that selects the highest-confidence <subject-object> pairs at a computational cost comparable to a single self-attention layer. We show how to directly supervise the attention scores of this mechanism without the need to propagate gradients through the hard selection.

³ We use <fixed width font> to distinguish the grammatical terms <subject> and <object> from generic use of the word *object* as in *object detection*.

An additional benefit of our design is that it disentangles object names from relationship predicates during inference. In contrast to prior open-vocabulary methods [53, 54, 61], our model can embed object and predicate texts separately and generate confidence scores efficiently for all possible <subject-predicate-object> combinations.

In summary, we make the following contributions:

- 1. An efficient architecture for open-vocabulary visual relationship detection.
- 2. A single-stage recipe for joint object and relationship detection training.
- 3. Efficient, disentangled object and relationship inference.
- 4. Analysis of inference speed, ablations, and qualitative examples.

Our method achieves state-of-the-art visual relationship detection performance on the Visual Genome dataset (26.1% mR@100 with graph constraint) and on the large-vocabulary GQA benchmark. It provides strong results while being significantly simpler than prior approaches to visual relationship detection.

2 Related Work

Below, we review the main approaches to VRD that are relevant to our work. For a comprehensive overview of the field, we refer to [27].

Detector-Agnostic Relationship Detection. Historically, many works on VRD assume that object detections are given, such that the task reduces to inferring relationships between the objects. These "detector-agnostic" methods use off-the-shelf detectors such as Faster-RCNN [39] to obtain boxes and box embeddings and infer a scene graph from them. Early work on large-scale VRD employed word embeddings to improve relationship generalization [32] and graph inference methods such as message passing for SGG [48]. More recent methods leverage relationship co-occurrence statistics [55] and address the long-tailed nature of the training data [8, 52, 57].

End-to-End Relationship Detection. A fundamental limitation of the detector-agnostic approach is that object detection and relationship prediction are treated separately and cannot learn from each other. Overcoming this limitation has recently motivated the development of end-to-end architectures in which objects and relationships are predicted jointly by a single model [7, 20, 29, 61]. All of these models have in common that they use a Transformer [47] decoder to predict object and relationship embeddings. The decoder represents object proposals with query embeddings that can cross-attend into image embeddings. The literature on object detection suggests that this choice may be problematic, because the query embeddings can be difficult to initialize, which can lead to unstable and slow optimization [3,34,43,50,58,63]. To avoid this issue, we design an encoder-only architecture in which both object and relationship embeddings are computed directly from the output tokens of a Vision Transformer [10] image encoder, building on an idea from object detection [34]. In contrast to prior models [61], this allows us to train the model end-to-end on object and relationship annotations simultaneously.

4 T. Salzmann et al.

Long-Tailed and Open-Vocabulary Relationship Detection. Object classes in the natural world follow a long-tailed distribution [12,62], and this is compounded in VRD due to the combination of subject, object, and predicate in a relationship [27, 48, 49]. A large number of VRD methods specifically aim to address this issue, for example with debiasing losses [52], data augmentation [57], or data resampling [8]. Additionally, while VRD models historically assumed a fixed set of object and predicate classes, recent works add open-vocabulary capabilities [13, 41, 54, 61]. These models leverage pretrained vision-language models such as CLIP [18, 37, 51, 56] to inherit their natural-language understanding. Instead of learning a fixed set of classes, open-vocabulary models predict class embeddings that can be compared to the text embeddings of arbitrary object or predicate descriptions for classification. Here, we employ a pretrained visionlanguage backbone and combine it with a carefully designed, lightweight relationship detection head to preserve and transfer semantic knowledge from the backbone pretraining to VRD. This allows our model to achieve strong results on the large-vocabulary GQA benchmark [16] (Section 4.3).

While we focus on architectural improvements, another line of work addresses the scarcity of training data for VRD. For example, RLIP [53] and RLIPv2 [54], which focus on human-object interaction [4], propose relationship-focused imagetext pretraining and self-training, which yield large improvements in VRD performance. These works are orthogonal to our contributions and can be combined.

3 Scene-Graph ViT

We propose an architecture for open-vocabulary relationship prediction in which both objects and the relationships between them are handled as "first-class citizens" directly in a single-stage process within the model backbone. We build on an encoder-only architecture for object detection [34] and adapt it to relationship prediction by adding a specialized attention layer. This layer exploits the pairwise structure between existing object embeddings to obtain relationship embeddings as schematized in Figure 2 and described below.

3.1 Encoder-Only Open-Vocabulary Object Detection

We briefly review the detection architecture that forms the basis of our model [34]. The model consists of Transformer-based [10] image and text encoders that are contrastively pretrained on a large number of image and text pairs [37]. The image encoder is adapted to detection by removing the final pooling layer and adding heads that predict bounding boxes and class embeddings directly from the output tokens of the image encoder. For open-vocabulary object classification, the embeddings computed by the class prediction head are compared to text encoder embeddings of object descriptions. This architecture achieves strong open-vocabulary object detection performance [33, 34] and does not suffer from the training instabilities observed in some encoder-decoder detection models [3, 34, 43, 50, 58, 63].



Figure 2: For relationship selection, image tokens are first projected using two lightweight MLPs to produce <subject> and <object> embeddings. A relationship score is then computed as the inner product between all <subject> and <object> embeddings. Relationships are filtered by first selecting the top object instances, using the scores along the diagonal to represent instance likelihood. Among the remaining instances, the top <subject-object> pairs are selected using the off-diagonal scores. This yields a set of relationship triplets, each consisting of a <subject> index, an <object> index, and a relationship embeddings. For classification, the relationship embeddings are compared against text embeddings of object class or predicate text descriptions.

3.2 Extension to Relationship Prediction

In the architecture described above, each image encoder token represents an object proposal that captures per-object information. Importantly, the encoder consists of self-attention layers which, by design, model all pairwise interactions between these tokens. We therefore hypothesize that information about object relationships can be learned directly in the image encoder. To extract this information in the form of <subject-predicate-object> triplets, we introduce two MLPs that transform vision encoder output tokens into <subject> embeddings s_i and <object> embeddings o_j (Figure 2). Using different MLPs for <subject> and <object> is crucial to break symmetry in subsequent processing (Appendix A.1).

An embedding representing the relationship between two object proposals (encoder tokens) can then be obtained simply by element-wise addition of their $\langle subject \rangle$ and $\langle object \rangle$ embeddings. Obtaining relationship embeddings for all $N^2 \langle subject-object \rangle$ pairs, where N is the number of object proposals, would be computationally infeasible. We therefore introduce a *Relationship Attention* layer that performs hard attention to select the $\langle subject-object \rangle$ pairs most likely to form a relationship. This layer computes an attention-like score $p_{ij} =$

 $\mathbf{s}_i \mathbf{o}_j^T$, where \mathbf{s}_i and \mathbf{o}_j are the embedding vectors of $\langle \mathtt{subject} \rangle i$ and $\langle \mathtt{object} \rangle j$, and p_{ij} represents the likelihood that a relationship between $\langle \mathtt{subject} \rangle i$ and $\langle \mathtt{object} \rangle j$ exists. By computing this score for all $\langle \mathtt{subject-object} \rangle$ pairs, we obtain an $N \times N$ matrix, from which we select the top k pairs.⁴ For these k pairs, we compute relationship embeddings $\mathbf{r}_{ij} = \text{LayerNorm}(\mathbf{s}_i + \mathbf{o}_j)$. To classify the relationship of a pair, its relationship embedding is compared to text embeddings of relationship predicates. We use the embeddings where $\langle \mathtt{subject} \rangle$ and $\langle \mathtt{object} \rangle$ are identical $(\mathbf{r}_{i=j})$ to represent object instances, and predict object classes from them. Boxes are predicted from the corresponding image encoder output tokens (see Figure 2).

The Relationship Attention layer therefore identifies object pairs to focus on for relationship classification, by computing a hard attention with <subject> embeddings as queries and <object> embeddings as keys. Since the hard selection is not differentiable, gradients from the relationship prediction cannot be used directly to train this layer. Instead, the relationship score is self-supervised to predict the maximum <predicate> class probability that will later be predicted for the relationship at the classification stage (Section 3.3).

3.3 Training

The image and text encoders of our model are initialized from a vision-language model that was contrastively pretrained on a large number of image-text pairs [37] After adding the Relationship Attention as well as the object bounding box and class prediction heads, the model is jointly trained on a mixture of object and relationship detection datasets in a single training stage (Section 4.1), using the losses described below.

Bipartite Matching. Bipartite matching between object predictions $(\mathbf{r}_{i=j})$ and ground-truth annotations is performed based on a cost consisting of the object classification loss and the box prediction losses as in [3]. This matching between objects also establishes a matching of predicted to ground-truth relationship predicates, since a relationship is uniquely identified by the *<subject>* and *<object>* indices *i* and *j*. Unmatched predictions are trained to predict low scores for all classes and incur no box prediction loss.

Box Prediction Losses. For bounding box regression, we use the L1 and generalized intersection-over-union (gIoU) losses described in [3].

Object and Predicate Classification Loss. For both object category and predicate classification, we use a sigmoid cross-entropy loss as in [34]. This loss is computed between the ground truth classes and logits obtained by computing the inner product of the relationship embeddings \mathbf{r}_{ij} selected by the Relationship Attention layer with the text embeddings of the class names. For embeddings corresponding to individual objects $(\mathbf{r}_{i=j})$, the class name is the object category or description. For embeddings corresponding to relationships $(\mathbf{r}_{i\neq j})$, we use the predicate text as class name. This differs from prior work [53, 54, 61], which

⁴ This selection reduces the number of relationships that need to be processed, e.g., from $N^2 = 3600^2$ to k = 16386 (99.9% reduction) for our ViT-L/14 model.

uses the full <subject-predicate-object> triplet description. Our approach has the advantage of disentangling object category and predicate names, which allows for more efficient inference since object category and predicate texts are embedded separately.

Relationship Score Loss. To select only the most promising embeddings for further processing, the Relationship Attention layer predicts a score p_{ij} that represents the likelihood that the corresponding $\langle \text{subject-object} \rangle$ pair forms a relationship (if $i \neq j$), or that the corresponding object exists in the image (if i = j). This score is trained with a sigmoid cross-entropy loss against targets provided by the model itself, namely the maximum probability predicted for any class for the corresponding \mathbf{r}_{ij} embedding [33]. In this way, the relationship score is trained to predict the class probability of a potential embedding \mathbf{r}_{ij} before actually computing and classifying the embedding. Note that this loss can only be computed for those objects and relationships that ultimately get selected for further processing. We found this to provide sufficient supervision.

The final loss is an equally weighted sum of four components: (1) classification loss, (2) L1 box loss, (3) gIoU box loss, (4) relationship score loss.

4 Experiments

4.1 Experimental Setup

Datasets. We use the following datasets for training or evaluation:

Visual Genome [26] is the largest VRD dataset, labeled with 2.3M triplet relationships across 108K images. However, as Visual Genome includes noisy annotations, the community commonly evaluates VRD on a cleaned version of the dataset [55] where the label space is reduced to the 150 most frequent object classes and the 50 most frequent predicate classes. This dataset is commonly referred to as Visual Genome 150 (VG150).

GQA [16] uses the same image corpus as Visual Genome, but is more diversely labeled, with 1703 object classes (1704 including a background class) and 310 predicates. Spatial relationships like "to the left of" are automatically labeled based on the 2-D spatial arrangement of bounding boxes. Similar to VQ150, GQA200 is a reduced and cleaned version of GQA with 200 object classes and 100 predicate classes [9].

HICO-DET [4] is a dataset which focuses on a more specialized subset of visual relationships, specifically interactions between humans and objects. HICO-DET contains 50k images and is exhaustively annotated for 600 defined human-object interaction (HOIs).

Objects365 [40] (O365) is a large scale object-detection dataset with 365 object categories across two million images.

Open X-Embodiment [35] (OXE) dataset is targeted towards learning vision language action policies for robotics. OXE lacks bounding box annotations which precludes quantitative evaluation, but it represents both a visually different setting compared to the standard VRD benchmarks and a promising application area, so we use it for qualitative evaluation.

	VG150	VG	HICO	GQA200	\mathbf{GQA}	Objects365
% in training mixture $#$ removed	12.5% 6587	12.5% 17727	$12.5\% \\ 729$	$12.5\% \\ 15361$	$\begin{array}{c} 0\% \\ 18801 \end{array}$	$50\% \\ 15306$

Table 1: Training data mixture. # removed indicates the number of images that were removed from the official training split due to overlap with data we evaluate on.

Training. Unless otherwise noted, we train models on a mixture of VG, VG150, GQA200, HICO, and O365 in the proportions show in Table 1. The B/32 model is trained for 200'000 steps at batch size 256 (further details in Appendix A.1). Since the datasets we use share similar image sources, there is some overlap between the official training and evaluation splits. To ensure that no evaluation images are used for training, we rigorously filter all training datasets to remove images present in any of the test splits for all datasets we evaluate on, i.e. VG150, GQA, GQA200, and HICO. We use an image similarity filter that also identifies non-exact matches [24]. Table 1 shows the number of images removed from each training split. Note that not all prior work performs similar deduplication.

Evaluation Procedure and Metrics. Prior work presents a diverse range of evaluation metrics, often tailored to specific datasets. For non-exhaustively labelled datasets, like Visual Genome and GQA, using precision-based metrics is inherently inconclusive. Thus, the community has adapted a Recall@K metric for these datasets, where K denotes a fixed budget of <subject-predicate-object> triplets on which the recall is computed. However, this metric is biased towards the more frequent <predicate> classes in the dataset [6,46]. Therefore, recent approaches use mean Recall@K, which calculates Recall@K separately for each <predicate> class in the test data and then averages the results.

For exhaustively labeled datasets like HICO, it is feasible to use precisionbased metrics such as mean Average Precision (mAP), a well-established metric in object detection. For HICO, the mean is taken over the 600 possible HOI triplets. Besides the overall metric, we also report results separately for rare (< 10 occurrences) and non-rare (≥ 10 occurrences) HOIs as mAPr and mAPn.

Further, evaluation can be "graph-constrained", where only a single prediction can be made per detected object pair, or "graph-unconstrained", where any number of predictions can be made per pair. Existing works evaluating on Visual Genome and GQA datasets partially lack a clear distinction between graphconstrained and unconstrained evaluation. For our models, we report graphconstrained results in the main paper, which reflects the dominant approach used in other studies. Results without graph-constraint are reported in Appendix A.2. For the HICO dataset, we follow the established evaluation procedure [4] that inherently represents a graph-unconstrained evaluation.

Multiple implementations of the aforementioned metrics exist, often showing differences in results. To compare fairly and directly against a wide range of prior approaches we identified the PyTorch evaluation procedure of [45] as the most prevalent routine for recall-based metrics and replicated their procedure in JAX with numerical accuracy. Similarly, for HICO, we numerically reproduce the original Matlab evaluation procedure outlined in [4] in JAX.

9

Model	Backbone	mR@50	mR@100
RelTR [7]	ResNet-50	6.8	10.8
Transformer $+$ CFA [28]	ResNeXt-101	12.3	14.6
VCTree $[46]$ + IETrans + Rwt $[57]$	ResNeXt-101	12.0	14.9
Motif $[55]$ + IETrans + Rwt $[57]$	ResNeXt-101	15.5	18.0
GPS-Net[31] + IETrans + Rwt[57]	ResNeXt-101	16.2	18.8
SG-Transformer [52] + IETrans + Rwt [57]	ResNeXt-101	15.0	18.1
Sgtr [29]	ResNet-101	15.8	20.1
DT2-ACBS [8]	$\operatorname{ResNet-101}$	22.0	24.4
UniVRD	CLIP: ViT-B/32	9.6	12.1
UniVRD	CLIP: ViT-B/16	10.9	13.2
UniVRD	CLIP: ViT-L/14	12.6	14.5
SG-ViT	CLIP: ViT-B/32	15.0	18.1
SG-ViT	CLIP: ViT-B/16	15.7	19.3
SG-ViT	CLIP: ViT-L/14	17.8	21.8
SG-ViT + simple predicate rebalancing	CLIP: ViT-L/14	22.3	26.1

Table 2: Performance on the Visual Genome 150 test set. We use graphconstrained evaluation (one predicate prediction per <subject-object> pair). For the last row ("+ simple text rebalancing"), the model was briefly fine-tuned on VG150 data with rebalanced predicate frequencies (see Section 4.2 and Appendix A.1).

Exhaustive Relationship Evaluation. Previous methods [53,54,61] entangle objects and relationships in a single representation. Consequently, these methods score their embeddings against the full <subject-predicate-object> triplet. While this is feasible for datasets with a low number of possible triplets, like HICO with 600 defined relationships, it quickly becomes computationally intractable for datasets with a larger vocabulary of objects and predicates (e.g. VG, GQA). Prior works [61] therefore evaluate only on triplets which are known to be present in the test split, which may inflate metrics. In contrast, our method disentangles objects and predicates (Section 3). This allows for an exhaustive evaluation that considers all object and predicate combinations and enables applications where knowledge of possible relationships is unavailable.

4.2 Relationship Detection Performance

We use Visual Genome as our main benchmark for relationship detection. Table 2 shows results for our models and prior work. Prior studies often achieve reported performance through the combination of a base model with a specialized approach to overcome the skewed data distribution (indicated as *base-model* + *specialization* in Table 2). Our models provide strong performance even without special treatment of the skewed predicate distribution. After brief fine-tuning on data in which the frequency of predicate annotations was rebalanced with simple rejection sampling (see Appendix A.1), our method improves on the prior best method DT2-ACBS [8] by 1.7 points mR@100 (26.1 vs. 24.4). This shows that our already-strong method may be further improved by combining it with more advanced rebalancing methods.

We also note the large difference in performance to UniVRD [61], which builds on the same detection architecture as our method, but adds a Transformer decoder-based relationship module. To disentangle whether the improve-

			Model
Model	mR@50	mR@100	Scene Graph Classification
LLM4SGG [21] Neural Motif w/ GCL [9] VCTree w/ GCL [9] SG-Transformer w/ CFA [28] SHA w/ GCL [9]	5.3 16.8 15.6 13.4 17.8	6.5 18.8 17.8 15.3 20.1	IMP [25, 44, 48] Neural Motif [25, 44, 55] Unbiased TDE [25, 45] RTN [25] MP [23] Transformer w/ EBM [44]
SG-ViT (CLIP: ViT-B/32) SG-ViT (CLIP: ViT-B/16) SG-ViT (CLIP: ViT-L/14)	15.7 17.7 19.3	18.6 20.5 22.9	Scene Graph Generation SG-ViT (CLIP: ViT-B/32 SG-ViT (CLIP: ViT-B/16 SG-ViT (CLIP: ViT-B/16

Table 3: Performance on the GQA200 test set. Graph-constrained evaluation.

Model	mR@50	mR@100
Scene Graph Classification	ı	
IMP [25, 44, 48]	0.5	0.7
Neural Motif [25, 44, 55]	0.8	1.2
Unbiased TDE [25, 45]	-	0.7
RTN [25]	-	1.4
MP [23]	-	2.8
Transformer w/ EBM [44]	1.3	1.8
Scene Graph Generation		
SG-ViT (CLIP: ViT-B/32)	6.2	7.4
SG-ViT (CLIP: ViT-B/16)	6.4	7.4
SG-ViT (CLIP: ViT-L/14)	8.0	9.6

Table 4: Performance on the GQA test set. Graph-constrained evaluation.

ments over UniVRD are due to differences in architecture or training data, we also trained our B/32 model on the UniVRD data mixture, and still observe a large improvement (17.4 VG mR@100 for our method vs. 12.1 for UniVRD when trained on the same data; Table 6). This suggests that our encoder-only architecture is better suited for VRD than prior decoder-based architectures [7,20,29,61].

Scaling to Large Vocabularies 4.3

Improving long-tail performance has been critical to the VRD field since the natural distribution of relationship triplets is highly skewed [27]. A substantial body of literature is devoted to this goal, often proposing complex loss debiasing or data augmentation approaches [8, 27, 52, 57]. It is therefore of interest whether our method works well in this regime without special treatment of rare classes. To evaluate the performance of our method on large object and predicate vocabularies with a long tail of rare classes, we use the Graph Question Answering (GQA) dataset [16]. This dataset builds on VG and expands the number of classes and annotations. A simplified version of GQA with 200 object and 100 predicate classes, called GQA200 [9], is part of our training mixture. Our method surpasses prior results on the GQA200 test split (Table 3), despite lacking the specialized treatment of the data distribution that some prior methods use.

The full GQA dataset has an even larger vocabulary with 1703 object and 311 predicate classes. To our knowledge, no prior methods evaluate scene graph generation on the full GQA dataset. Some prior methods report results on scene graph *classification*, where ground-truth boxes are provided. As shown in Table 4, our model achieves higher performance on the much harder task of scene graph generation than prior methods on classification. We further assess zeroshot generalization to unseen classes in Appendix A.2.

We believe that two factors contribute to the strong performance in the open vocabulary regime: First, the simple design of the method does not impede transfer of semantic knowledge from the pre-trained VLM backbone because it adds only lightweight heads. In particular, we use no decoder, which could be difficult to train and could lead to "forgetting" of pretrained representations. Instead, most of the relationship modeling happens in the backbone, which has had the benefit of large-scale vision-language pretraining. Second, the openvocabulary design allows end-to-end training on a mix of datasets, which was not possible for all prior methods.

4.4 Human-Object Interaction

We also evaluate our model on the specialized task of human-object interaction, using the HICO benchmark [4]. The performance of our model is on par with the most comparable prior method [61], but does not show similarly large improvements as for VG or GQA200 (Table 5). HICO has a much narrower and more specialized vocabulary than VG or GQA (HICO evaluates on 600 prespecified triplets, whereas VG150 has 1.1×10^6 possible triplet combinations). Performance on this specialized task may benefit less from transfer of pretrained representations and may instead be limited by the amount of task-specific training data. This is supported by the fact that the recent state-of-the-art method for HICO pretrains on large amounts of person-focused pseudo-labels [54].

4.5 Ablations

Inference Speed and Number of Predicted Relationships. In the following, we ablate training datasets and model components to study the interplay of data and architecture on the model performance.

The primary hyperparameter of our method is the number of relationships k that are selected by the Relationship Attention layer from all possible <subject-object> combinations for further processing. Figure 3 shows how VRD performance and inference speed depend on k.

For training, we use $k = 2^{14}$, which results in 71% (B/32) to 81% (L/14) of the speed of a pure object detection model [34] of the same size. Without top-k selection, the model would be several times slower (right end of top plot in Figure 3). Since the vast majority of *subject-object>* pairs do not form relationships, VRD performance is essentially unaffected by top-k selection at inference. For inference, we can therefore use $k = 2^{11}$ at no loss in accuracy to achieve 90% of the speed of the object detector (for B/32 and



Figure 3: Model speed and VRD accuracy by number of selected relationships k. Speed is relative to a non-VRD object detector [34].

L/14). At this k, the B/32 model achieves 52.8 FPS at batch size 1 on an NVIDIA V100 GPU and is therefore suitable for real-time applications.

Dataset Mixture. To assess how training data and architectural improvements contribute to the performance of our model compared to prior work, we train the models on a range of dataset mixtures (Table 6). The full mixture, which we use for all experiments unless otherwise noted, includes VG, VG150, GQA200,

12 T. Salzmann et al.

Model	Backbone	mAP	mAPr	mAPn
Model HOTR [20] RLIP [53] RLIPv2 [54] RLIPv2 [54] UniVRD [60] UniVRD [60] SG-ViT SG-ViT	ResNet50 ResNet50 ResNet50 Swin-L CLIP: ViT-B/32 CLIP: ViT-B/16 CLIP: ViT-B/16 CLIP: ViT-B/22 CLIP: ViT-B/32	mAP 25.10 32.84 35.38 45.09 29.98 37.41 28.86 31.98	mAPr 20.33 26.85 29.61 43.23 22.94 22.94 28.90 23.72 26.83	mAPn 25.86 34.63 37.10 45.64 32.02 32.02 39.95 30.39 33.52
SG-ViT	CLIP: ViT-L/14	38.11	33.71	39.42

Table 5: Performance on the HICO test set. Our method is on par with the most comparable prior work, UniVRD, for overall mAP. However, our method performs better on "rare" classes (mAPr), indicating better generalization to rarely seen concepts. For reference, we also show RLIPv2, the state-of-the-art method on HICO. RLIPv2 proposes a self-training approach that is orthogonal and compatible to our method.

	VG	VG150	Mixtu GQA200	ıre HICO	O365	COCO	VG150 mR@100	GQA200 mR@100	HICO mAP
Our mixture	12.5%	12.5%	12.5%	12.5%	50%	_	18.1	18.6	28.9
UniVRD [61] mix ours w/o O365 ours w/o VG	$\begin{vmatrix} -\\ 25\%\\ - \end{vmatrix}$	$10\% \\ 25\% \\ 25\%$	$_{25\%}^{-}$ 12.5%	$20\% \\ 25\% \\ 12.5\%$	50% - 50%	20% 	$ \begin{vmatrix} 17.4 & \downarrow & 4\% \\ 15.3 & \downarrow & 15\% \\ 14.2 & \downarrow & 22\% \end{vmatrix} $	$ \begin{array}{c ccccc} 5.1 & \downarrow 73\% \\ 15.6 & \downarrow 16\% \\ 15.3 & \downarrow 18\% \end{array} $	$\begin{vmatrix} 17.4 \ \downarrow \ 8\% \\ 23.8 \ \downarrow \ 18\% \\ 26.8 \ \downarrow \ 7\% \end{vmatrix}$

Table 6: Dataset Ablation. Performance of SG-ViT (B/32) trained on different data mixtures. VG150 and GQA200 evaluation is graph-constrained, HICO is unconstrained.

HICO and O365. To allow direct comparison to UniVRD [61], we also train on their mixture (VG150, HICO, O365, COCO) and find that our model still performs better on VG150 by a large margin (18.1 mR@100 vs 12.1 for UniVRD for B/32 models). Since UniVRD uses the same image and text encoding backbones as our model, this result suggests that our encoder-only model with Relationship Attention is an advance over decoder-based relationship architectures. We further find that including both pure object detection (O365) and large-vocabulary VRD data (VG) benefit all metrics.

Object Detection Performance. A necessary step towards good relationship detection performance is accurate object detection. In our architecture, object detection is treated as a special case of relationship detection in which the <subject> is identical to the <object>. We validate this design by training the model only on detection datasets without relationship annotations, using the same dataset mixture as the OWL-ViT [34] model (Objects365 and Visual Genome, [34]). We find that our model achieves a similar object detection performance as OWL-ViT (21.9% vs. 22.1% mAP and 18.3% vs. 18.9% mAPr on LVIS [12]), indicating that the Relationship Attention design does not interfere with object detection.



Figure 4: Qualitative examples of difficult edge-cases from VG150 test split [26, 55]. From left to right: Ground Truth, SG-ViT (B/32), SG-ViT (L/14). In all cases the $\langle subject \rangle$ is lime and the $\langle object \rangle$ is red.

4.6 Qualitative Results

In Figure 4, we visualize qualitative examples to highlight difficult edge cases. Even where the model output differs from the ground truth, it is often plausibly correct, e.g. where either the extent of an object differs with the ground truth or there are multiple appropriate, but unannotated, relations. For example, both models select the headlight on the bus in Figure 4b for "light on bus" as opposed to the sun reflected off the windshield. In Figure 4c both models focus more on the predicate> "on" in "snow on hill" than the distinction between a mountain and a hill. Figure 4d shows a challenging scene where the B/32 model selects the bottle inside the open fridge while the L/14 correctly selects a water bottle in the door for "bottle in door" while also capturing the extent of the door instance.

Furthermore, to qualitatively assess the utility of this work beyond standard relation benchmark datasets, we also visualize some example relations applied to data from the robotics domain, namely the OXE dataset [35]. Figure 5 shows a real-life scene where multiple instances of the same semantic object categories are present within the scene. The model is able to match the <predicate> to the instances in the desired configuration. One subjective failure-case is that the "banana outside bowl" text query had its highest scoring triplet involving a bowl on a shelf in the background rather than the adjacent bowl. Another charac-



Figure 5: Qualitative examples showing SG-ViT (L/14) on out-of-distribution data from the OXE dataset [35]. In all cases the <subject> is lime and the <object> is red. Note how the model correctly disambiguates several instances of the same class (e.g. "bananas" and "bottle") depending on their relationships.

teristic is that the model occasionally places boxes over groups with multiple instances of the same semantic class when a singular word is used for either the <object> or <subject>. This can be attributed to the similarity between word and their plural in the CLIP embeddings and that many human annotations from VG also confuse singular and plural (see Appendix A.3).

5 Limitations

While our model performs strongly on large-vocabulary VRD, its performance on specialized human-object interaction detection is only on par with prior models. We discuss this limitation in Section 4.4. Further error modes are explored qualitatively in Section 4.6 and Appendix A.3. A challenge for open-vocabulary relationship detection models as a whole is zero-shot generalization to unseen objects and predicates. While our model improves over prior approaches (Tables 3 and 4), there is still a large gap when it comes to entirely unseen classes (Appendix A.2). Future research should focus on closing this gap.

6 Conclusion

We present an architecture for open-vocabulary visual relationship detection. By combining an encoder-only design with a novel Relationship Attention layer for efficient selection of high-confidence relationships, our architecture leverages VLM pretraining and multi-dataset VRD training. It achieves strong performance on standard and large-vocabulary VRD benchmarks while maintaining pure object detection performance and adding little extra computational cost. Due to its simplicity and strong performance, we believe that our method will be a useful basis for further research on visual relationship detection.

References

- Amiri, S., Chandan, K., Zhang, S.: Reasoning with scene graphs for robot planning under partial observability. IEEE Robotics and Automation Letters 7(2), 5560– 5567 (2022)
- Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: Endto-end object detection with transformers. In: Conference on Computer Vision (ECCV). pp. 213–229. Springer (2020)
- Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to Detect Human-Object Interactions. In: IEEE Winter Conference on Applications of Computer Vision (2018)
- Chen, B., Xia, F., Ichter, B., Rao, K., Gopalakrishnan, K., Ryoo, M.S., Stone, A., Kappler, D.: Open-vocabulary queryable scene representations for real world planning. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 11509–11522. IEEE (2023)
- Chen, T., Yu, W., Chen, R., Lin, L.: Knowledge-embedded routing network for scene graph generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6163–6171 (2019)
- Cong, Y., Yang, M.Y., Rosenhahn, B.: Reltr: Relation transformer for scene graph generation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
- Desai, A., Wu, T.Y., Tripathi, S., Vasconcelos, N.: Learning of visual relations: The devil is in the tails. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15404–15413 (2021)
- Dong, X., Gan, T., Song, X., Wu, J., Cheng, Y., Nie, L.: Stacked hybridattention and group collaborative learning for unbiased scene graph generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19427–19436 (2022)
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR (2021)
- Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., et al.: Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. arXiv preprint arXiv:2309.16650 (2023)
- Gupta, A., Dollar, P., Girshick, R.: LVIS: A Dataset for Large Vocabulary Instance Segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
- He, T., Gao, L., Song, J., Li, Y.F.: Towards Open-Vocabulary Scene Graph Generation with Prompt-Based Finetuning. In: European Conference on Computer Vision (ECCV) (2022)
- 14. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
- Herzig, R., Mendelson, A., Karlinsky, L., Arbelle, A., Feris, R., Darrell, T., Globerson, A.: Incorporating Structured Representations into Pretrained Vision \& Language Models Using Scene Graphs. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023)

- 16 T. Salzmann et al.
- Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6700–6709 (2019)
- Hughes, N., Chang, Y., Carlone, L.: Hydra: A Real-time Spatial Perception System for 3D Scene Graph Construction and Optimization. Robotics: Science and Systems (RSS) (2022)
- Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
- Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image Retrieval Using Scene Graphs. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
- Kim, B., Lee, J., Kang, J., Kim, E.S., Kim, H.J.: Hotr: End-to-end human-object interaction detection with transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 74–83 (2021)
- Kim, K., Yoon, K., Jeon, J., In, Y., Moon, J., Kim, D., Park, C.: LLM4SGG: Large Language Model for Weakly Supervised Scene Graph Generation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
- 22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
- Knyazev, B., de Vries, H., Cangea, C., Taylor, G.W., Courville, A., Belilovsky, E.: Graph Density-Aware Losses for Novel Compositions in Scene Graph Generation. In: British Machine Vision Conference (BMVC) (2020)
- Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big transfer (bit): General visual representation learning. In: European Conference on Computer Vision (ECCV). pp. 491–507 (2020)
- Koner, R., Shit, S., Tresp, V.: Relation transformer network. arXiv preprint arXiv:2004.06193 (2020)
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**, 32–73 (2017)
- Li, H., Zhu, G., Zhang, L., Jiang, Y., Dang, Y., Hou, H., Shen, P., Zhao, X., Shah, S.A.A., Bennamoun, M.: Scene Graph Generation: A comprehensive survey. Neurocomputing (2024). https://doi.org/https://doi.org/10.1016/j.neucom. 2023.127052
- Li, L., Chen, G., Xiao, J., Yang, Y., Wang, C., Chen, L.: Compositional feature augmentation for unbiased scene graph generation. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21685–21695 (2023)
- Li, R., Zhang, S., He, X.: Sgtr: End-to-end scene graph generation with transformer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19486–19496 (2022)
- Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1261–1270 (2017)
- Lin, X., Ding, C., Zeng, J., Tao, D.: Gps-net: Graph property sensing network for scene graph generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3746–3753 (2020)

- Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: European Conference on Computer Vision (ECCV). pp. 852– 869. Springer (2016)
- Minderer, M., Gritsenko, A., Houlsby, N.: Scaling open-vocabulary object detection. Advances in Neural Information Processing Systems 36 (2024)
- 34. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al.: Simple open-vocabulary object detection. In: European Conference on Computer Vision (ECCV). pp. 728–755. Springer (2022)
- 35. Open X-Embodiment Collaboration: Open X-Embodiment: Robotic Learning Datasets and RT-X Models (2023)
- Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Ye, Q., Wei, F.: Grounding Multimodal Large Language Models to the World. In: International Conference on Learning Representations (2024)
- 37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
- Rana, K., Haviland, J., Garg, S., Abou-Chakra, J., Reid, I., Suenderhauf, N.: Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. In: 7th Annual Conference on Robot Learning (2023)
- Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28 (2015)
- 40. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8430–8439 (2019)
- Shi, H., Hayat, M., Cai, J.: Open-vocabulary object detection via scene graph discovery. In: 31st ACM International Conference on Multimedia. pp. 4012–4021 (2023)
- Singh, K.P., Salvador, J., Weihs, L., Kembhavi, A.: Scene graph contrastive learning for embodied navigation. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10884–10894 (2023)
- 43. Song, H., Sun, D., Chun, S., Jampani, V., Han, D., Heo, B., Kim, W., Yang, M.H.: ViDT: An Efficient and Effective Fully Transformer-based Object Detector. In: International Conference on Learning Representations (2022)
- 44. Suhail, M., Mittal, A., Siddiquie, B., Broaddus, C., Eledath, J., Medioni, G., Sigal, L.: Energy-based learning for scene graph generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13936–13945 (2021)
- Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3716–3725 (2020)
- 46. Tang, K., Zhang, H., Wu, B., Luo, W., Liu, W.: Learning to compose dynamic tree structures for visual contexts. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6619–6628 (2019)
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
- Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5410–5419 (2017)

- 18 T. Salzmann et al.
- Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.: Graph r-cnn for scene graph generation. In: European Conference on Computer Vision (ECCV). pp. 670–685 (2018)
- Yao, Z., Ai, J., Li, B., Zhang, C.: Efficient DETR: Improving End-to-End Object Detector With Dense Prior. arXiv preprint arXiv:2104.01318 (2021)
- Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: CoCa: Contrastive Captioners are Image-Text Foundation Models. Transactions on Machine Learning Research (2022)
- Yu, J., Chai, Y., Wang, Y., Hu, Y., Wu, Q.: CogTree: Cognition Tree Loss for Unbiased Scene Graph Generation. In: International Joint Conference on Artificial Intelligence, IJCAI-21. International Joint Conferences on Artificial Intelligence Organization (8 2021)
- Yuan, H., Jiang, J., Albanie, S., Feng, T., Huang, Z., Ni, D., Tang, M.: Rlip: Relational language-image pre-training for human-object interaction detection. Advances in Neural Information Processing Systems 35, 37416–37431 (2022)
- Yuan, H., Zhang, S., Wang, X., Albanie, S., Pan, Y., Feng, T., Jiang, J., Ni, D., Zhang, Y., Zhao, D.: RLIPv2: Fast Scaling of Relational Language-Image Pretraining. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21649–21661 (2023)
- Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5831–5840 (2018)
- Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18123–18133 (2022)
- 57. Zhang, A., Yao, Y., Chen, Q., Ji, W., Liu, Z., Sun, M., Chua, T.S.: Fine-grained scene graph generation with data transfer. In: European Conference on Computer Vision (ECCV). pp. 409–424. Springer (2022)
- Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)
- 59. Zhang, Y., Ma, Z., Gao, X., Shakiah, S., Gao, Q., Chai, J.: GROUNDHOG: Grounding Large Language Models to Holistic Segmentation. In: Conference on Computer Vision and Pattern Recognition 2024 (2024), https://openreview. net/forum?id=PWVW3ux1Ju
- Zhao, L., Yuan, L., Gong, B., Cui, Y., Schroff, F., Yang, M.H., Adam, H., Liu, T.: Unified Visual Relationship Detection with Vision and Language Models. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
- Zhao, L., Yuan, L., Gong, B., Cui, Y., Schroff, F., Yang, M.H., Adam, H., Liu, T.: Unified Visual Relationship Detection with Vision and Language Models. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
- Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twentythousand classes using image-level supervision. In: European Conference on Computer Vision (ECCV). pp. 350–368 (2022)
- Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable Transformers for End-to-End Object Detection. In: International Conference on Learning Representations (2021)