

MDETR - Modulated Detection for End-to-End Multi-Modal Understanding

Aishwarya Kamath
NYU

Mannat Singh
Facebook

Yann LeCun
NYU

Ishan Misra
Facebook

Gabriel Synnaeve
Facebook

Nicolas Carion
NYU

Abstract

Multi-modal reasoning systems rely on a pre-trained object detector to extract regions of interest from the image. However, this crucial module is typically used as a black box, trained independently of the downstream task and on a fixed vocabulary of objects and attributes. This makes it challenging for such systems to capture the long tail of visual concepts expressed in free form text. In this paper we propose MDETR, an end-to-end modulated detector that detects objects in an image conditioned on a raw text query like a caption or a question. We use a transformer-based architecture to reason jointly over text and image by fusing the two modalities at an early stage of the model. We pre-train the network on 1.3M text-image pairs, mined from pre-existing multi-modal datasets having explicit alignment between phrases in text and objects in the image. We then fine-tune on several downstream tasks such as phrase grounding, referring expression comprehension and segmentation, achieving state-of-the-art results on popular benchmarks. We also investigate the utility of our model as an object detector on a given label set when fine-tuned in a few-shot setting. We show that our pre-training approach provides a way to handle the long tail of object categories which have very few labelled instances. Our approach can be easily extended for visual question answering, achieving competitive performance on GQA and CLEVR. The code and models are available at <https://github.com/ashkamath/mdeatr>.

1. Introduction

Object detection forms an integral component of most state-of-the-art multi-modal understanding systems [6, 28], typically used as a black-box to detect a fixed vocabulary of concepts in an image. This popular approach to using detection systems does not account for downstream multi-modal understanding tasks, and is often a bottleneck for performance [73]. In addition, this system is usually frozen,

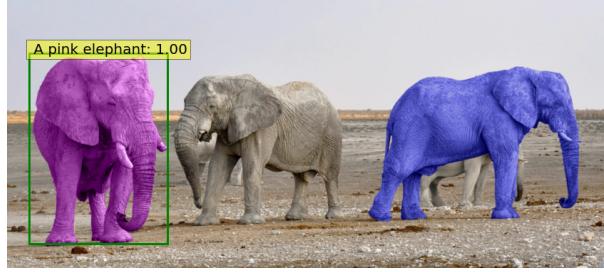


Figure 1: Output of MDETR for the query “A pink elephant”. The colors are not segmentation masks but the real colors of the pixels. The model has never seen a pink nor a blue elephant in training.

which not only prevents further refinement of the model’s perceptive capability, but also restricts it to only have access to the detected objects and not the whole image. This “pipelined” approach limits co-training with other modalities as context. In the vision-language setting, it implies restricting the vocabulary of the resulting system to the categories and attributes of the detector. As a result, such a system cannot recognize novel combinations of concepts that can be expressed in free-form text.

A recent line of work [67, 45, 13] considers the problem of text-conditioned object detection. These methods extend mainstream one-stage and two-stage detection architectures to achieve this goal. However, to the best of our knowledge, it has not been demonstrated that such detectors can improve performance on downstream tasks that require reasoning over the detected objects such as VQA. We believe this is because these detectors are not end-to-end differentiable and thus cannot be trained in synergy with downstream tasks.

Our method, MDETR, is an end-to-end *modulated* detector based on the recent DETR [2] detection framework, and performs object detection in conjunction with natural language understanding, enabling truly end-to-end multi-modal reasoning. MDETR relies solely on text and aligned boxes as a form of supervision for concepts in an image. Thus, unlike current detection methods, MDETR detects nuanced concepts from free-form text, and generalizes to

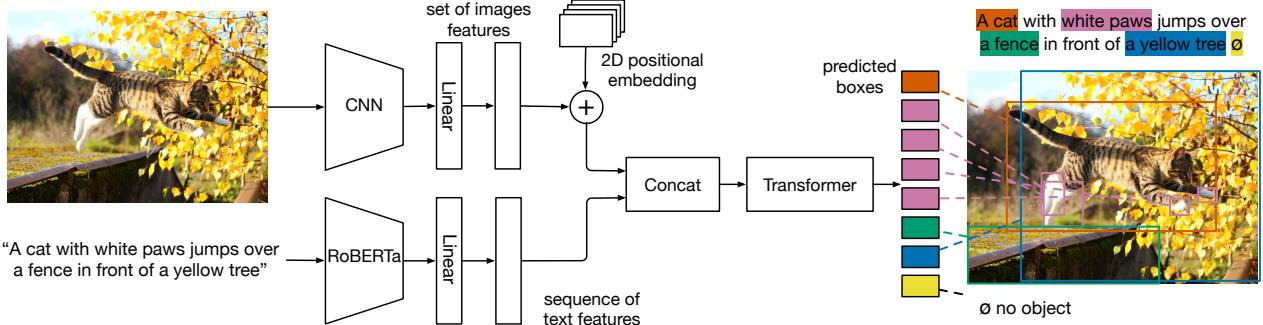


Figure 2: MDETR uses a convolutional backbone to extract visual features, and a language model such as RoBERTa to extract text features. The features of both modalities are projected to a shared embedding space, concatenated and fed to a transformer encoder-decoder that predicts the bounding boxes of the objects and their grounding in text.

unseen combinations of categories and attributes. We showcase such a combination as well as modulated detection in Fig. 1. By design, our predictions are grounded in text, which is a key requirement for visual reasoning [66]. When pre-trained using a dataset of 200,000 images and aligned text with box annotations, we achieve best reported results on the Flickr30k dataset for phrase grounding, ReFCOCO/+g datasets for referring expression comprehension, and referring expression segmentation on PhraseCut, as well as competitive performance on the GQA and CLEVR benchmarks for visual question answering.

Our contributions are as follows:

- We introduce an end-to-end text-modulated detection system derived from the DETR detector.
- We demonstrate that the modulated detection approach can be applied seamlessly to solve tasks such as phrase grounding and referring expression comprehension, setting new state of the art performance on both these tasks using datasets having synthetic as well as real images.
- We show that good modulated detection performance naturally translates to downstream task performance, for instance achieving competitive performance on visual question answering, referring expression segmentation, and on few-shot long-tailed object detection.

2. Related work

The CLEVR dataset [20] is a popular vision-language benchmark for reasoning on objects, their relations, and the composition of such relations. A prominent line of work [21, 68, 37, 15] make use of the functional programs annotations that are part of the CLEVR dataset. Such approaches tend to dominate on the question answering benchmark, but fail to generalize beyond synthetic data. Conversely, many approaches [42, 51, 60, 17] learn directly from images or

pre-detected objects, with varying amounts of inductive bias tailored to the QA task. Our method can be seen as an in-between: while not explicitly using the program supervision, it is trained to detect objects that are required for performing intermediate reasoning steps.

Recent progress in multi-modal understanding has been mainly powered by pre-training large transformer models to learn generic multi-modal representations from enormous amounts of aligned image-text data [53], then fine-tuning them on downstream tasks. These methods can be divided into single stream [6, 28, 73, 26] and two-stream [55, 34, 35, 54] architectures depending on whether the text and images are processed by a single combined transformer or two separate transformers followed by some cross attention layers. For both these types, the prevalent approach is to extract visual and textual features independently and then use the attention mechanism of the transformers to learn an alignment between the two. While this approach has improved state of the art results on a wide variety of tasks such as image-text retrieval [73], phrase grounding [26], image captioning [28] and visual question answering [25], it leaves opportunity for a more tightly knit architecture, such as MDETR, in which information flows between the two modalities at an even earlier stage of the model. Some previous attempts at achieving this using modulated architectures such as [42] and [39] show improvements on counting tasks and visual question answering.

The visual features used by the current state-of-the-art models are extracted using an external pre-trained detector [1], which outputs regions that are noisy, often over-sampled and ambiguous. [28] attempts to alleviate the problem of noisy image features by using tags as anchors between the text and images. This is still a weaker form of supervision than in MDETR where we have explicit alignment between words or phrases in text and the objects in the images. To alleviate the constraints implied by fixed vocabulary of concepts, [73] trains on a collection of much

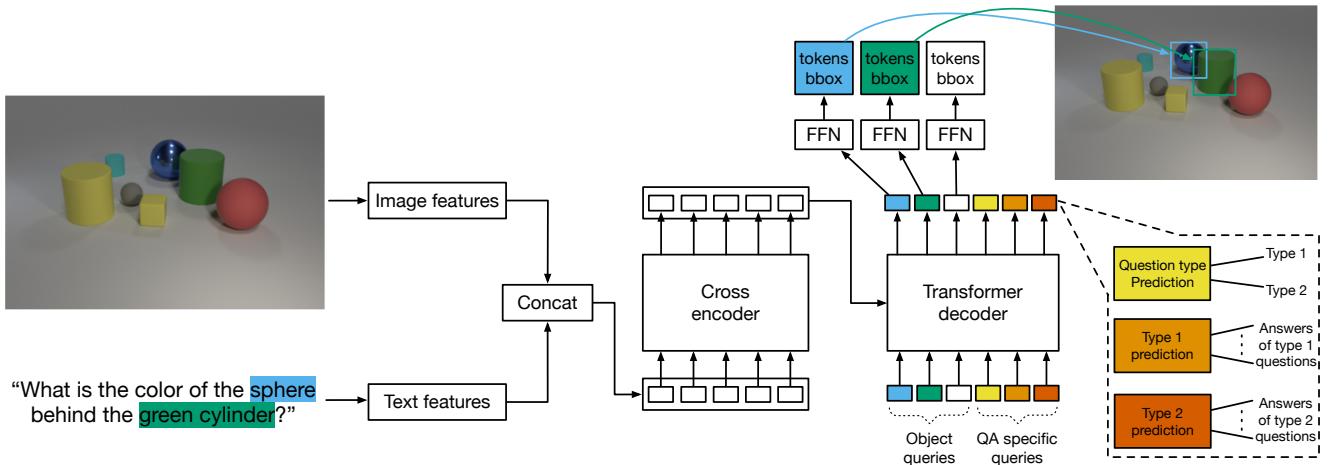


Figure 3: During MDETR pre-training, the model is trained to detect all objects mentioned in the question. To extend it for question answering, we provide QA specific queries in addition to the object queries as input to the transformer decoder. We use specialized heads for different question types.

larger object detection datasets in pursuit of better coverage. [9] conduct adversarial training on top of existing high performing models pushing performance even higher. Other approaches [69] attempt to incorporate scene graph prediction as part of their pre-training to learn more robust representations. Some recent work also attempts to build multi-purpose multi-modal architectures that are able to tackle a variety of vision-language [7] as well as pure language tasks in a single architecture [16]. A separate line of work that attacks a similar problem to ours but with a much more task specialized model architectures is the single [67, 3, 27] and two stage [45, 13] referring expression segmentation and phrase detection models which are designed specifically for this task.

3. Method

In this section we first briefly summarize the object detection pipeline [2] based on which we build our model in §3.1 and then describe how we extend it for modulated detection in §3.2.

3.1. Background

DETR Our approach to modulated detection builds on the DETR system [2], which we briefly review here. We refer the readers to the original paper for additional details. DETR is an end-to-end detection model composed of a backbone (typically a convolutional residual network [12]), followed by a Transformer Encoder-Decoder [59].

The DETR encoder operates on 2D flattened image features from the backbone and applies a series of transformer layers. The decoder takes as input a set of N learned embeddings called *object queries*, that can be viewed as slots that the model needs to fill with detected objects. All the

object queries are fed in parallel to the decoder, which uses cross-attention layers to look at the encoded image and predicts the output embeddings for each of the queries. The final representation of each object query is independently decoded into box coordinates and class labels using a shared feed-forward layer. The number of object queries acts as a de facto upper-bound on the number of objects the model can detect simultaneously. It has to be set to a sufficiently large upper-bound on the number of objects one may expect to encounter in a given image. Since the actual number of objects in a particular image may be less than the number of queries N , we use an extra class label corresponding to “no object”, denoted by \emptyset . The model is trained to output this class for every query that doesn’t correspond to an object.

DETR is trained using a Hungarian matching loss, where a bipartite matching is computed between the N proposed objects and the ground-truth objects. Each matched object is supervised using the corresponding target as ground-truth, while the un-matched objects are supervised to predict the “no object” label \emptyset . The classification head is supervised using standard cross-entropy, while the bounding box head is supervised using a combination of absolute error (L1 loss) and Generalized IoU [48].

3.2. MDETR

3.2.1 Architecture

We depict the architecture for MDETR in Fig. 2. As in DETR, the image is encoded by a convolutional backbone and flattened. 2-D positional embeddings are added to this flattened vector to conserve the spatial information. The text is encoded using a pre-trained transformer language model to produce a sequence of hidden vectors of same size as the input. A modality dependent linear projection is applied

to both the image and text features to project them into a **shared embedding space**. These feature vectors are then concatenated on the sequence dimension to yield a single sequence of image and text features. This sequence is fed to a **joint transformer encoder termed as the *cross encoder***. Following DETR, we apply a transformer decoder on the object queries while cross attending to the final hidden state of the cross encoder. **The decoder’s output is used for predicting the actual boxes.**

3.2.2 Training

We present the **two additional loss functions** used by MDETR, which encourage alignment between the image and the text. They both use the same source of annotations: free form text with aligned bounding boxes. The first loss function that **we term as the *soft token prediction*** loss is a **non parametric alignment loss**. The second, termed as the ***text-query contrastive alignment*** is a parametric loss function enforcing **similarity between aligned object queries and tokens**.

Soft token prediction For modulated detection, unlike in the standard detection setting, we are **not interested in predicting a categorical class for each detected object**. Instead, we predict the span of tokens from the original text that refers to each matched object. Concretely, we first set the **maximum number of tokens** for any given sentence to be **L = 256**. For each predicted box that is matched to a ground truth box using the bi-partite matching, **the model is trained to predict a uniform distribution over all *token positions* that correspond to the object**. Fig. 2 shows an example where the box for cat is trained to predict a uniform distribution over the first two words. In Fig. 9, we show a simplified visualization of the loss for this example, in terms of a distribution over words for each box, but in practice we use token spans after tokenization using a BPE scheme [52]. **Any query that is not matched to a target is trained to predict the “no object” label \emptyset** . Note that several words in the text could correspond to the same object in the image, and conversely several objects could correspond to the same text. For example, “a couple” referred to by two boxes in the image, could further be referred to individually in the same caption. By designing the loss function in this way, our model is able to learn about co-referenced objects from the same referring expression.

Contrastive alignment While the soft token prediction uses *positional* information to align the objects to text, the contrastive alignment loss enforces alignment between the *embedded representations* of the object at the output of the decoder, and the text representation at the output of the cross encoder. This additional contrastive alignment loss ensures that the embeddings of a (visual) object and its corresponding (text) token are closer in the feature space compared to

embeddings of unrelated tokens. This constraint is stronger than the soft token prediction loss as it directly operates on the representations and is not solely based on positional information. More concretely, consider the maximum number of tokens to be L and maximum number of objects to be N . Let T_i^+ be the set of tokens that a given object o_i should be aligned to, and O_i^+ be the set of objects to be aligned with a given token t_i .

The contrastive loss for all objects, inspired by InfoNCE [40] is normalized by number of positive tokens for each object and can be written as follows:

$$l_o = \sum_{i=0}^{N-1} \frac{1}{|T_i^+|} \sum_{j \in T_i^+} -\log \left(\frac{\exp(o_i^\top t_j / \tau)}{\sum_{k=0}^{L-1} \exp(o_i^\top t_k / \tau)} \right) \quad (1)$$

where τ is a temperature parameter that we set to 0.07 following literature [64, 47]. By symmetry, the contrastive loss for all tokens, normalized by the number of positive objects for each token is given by:

$$l_t = \sum_{i=0}^{L-1} \frac{1}{|O_i^+|} \sum_{j \in O_i^+} -\log \left(\frac{\exp(t_i^\top o_j / \tau)}{\sum_{k=0}^{N-1} \exp(t_i^\top o_k / \tau)} \right) \quad (2)$$

We take the average of these two loss functions as our contrastive alignment loss.

4. Experiments

In this section we describe the data and training used for pre-training MDETR, and provide details and results on the tasks that we use to evaluate our approach. Experiments and results on synthetic data from the CLEVR dataset are described in §4.1 followed by experiments on real world images in §4.2

4.1. Synthetic Images

CLEVR [20] is a synthetic dataset designed to evaluate multi-modal models in a controlled setting. While the visual understanding component of this task is trivial by modern standards, the main challenge is in understanding the questions, which can contain up to 20 steps of reasoning, and aligning the key phrases in them to the correct objects in the image. Unlike several successful approaches to CLEVR [21, 17, 37, 68], MDETR doesn’t incorporate any special inductive bias to cope with such complex reasoning tasks. In this section, we show that despite its relatively straight-forward formulation, our approach competes with state-of-the-art models on the question answering task.

Model We use a ResNet-18 [12] from Torchvision, pre-trained on ImageNet [49] as the convolutional backbone. For the text-encoder, we use a pre-trained DistilRoberta [50] from HuggingFace [62]. The final transformer is the same as DETR, with 6 encoder layers as well as 6 decoder layers,

Method	CLEVR						CLEVR-Humans		CoGenT		CLEVR-Ref+
	Overall	Count	Exist	Comp. Num	Query	Comp. Att	Before FT	After FT	TestA	TestB	Acc
MAttNet[70]	-	-	-	-	-	-	-	-	-	-	60.9
MGA-Net[74]	-	-	-	-	-	-	-	-	-	-	80.1
FiLM[42]	97.7	94.3	99.1	96.8	99.1	99.1	56.6	75.9	98.3	78.8	-
MAC [17]	98.9	97.1	99.5	99.1	99.5	99.5	57.4	81.5	-	-	-
NS-VQA[68]*	99.8	99.7	99.9	99.8	99.8	99.8	-	67.8	99.8	63.9	-
OCCAM [60]	99.4	98.1	99.8	99.0	99.9	99.9	-	-	-	-	-
MDETR	99.7	99.3	99.9	99.4	99.9	99.9	59.9	81.7	99.8	76.7	100

Table 1: Results on CLEVR-based datasets. We report accuracies on the test set of CLEVR, including the detail by question type. On CLEVR-Humans, we report accuracy on the test set before and after fine-tuning. On CoGenT, we report performance when the model is trained in condition A, without finetuning on condition B. On CLEVR-Ref+, we report the accuracy on the subset where the referred object is unique. *indicates method uses external program annotations

and 8 attention heads in the attention layers. We reduce the number of object queries to 25, since the maximum number of objects to be detected is low.

Supervision The first ingredient required for training MDETR is bounding box annotations for objects in the image. The original CLEVR dataset doesn’t provide them, so we use the scene graphs from the dataset to re-create them. The second ingredient required is the alignment between bounding boxes and tokens in the question. MDETR is trained to predict only objects that are referred to in the question. Further details on bounding box and span extraction can be found in Appendix B.1.

Training We extract a subset of the full dataset having only questions with a single reference to a given object. This subset that we call CLEVR-Medium is about 15% smaller than the full set. We adopt a curriculum learning strategy where we first train only on the task of modulated detection for 30 epochs on this reduced set, followed by 30 epochs on the full CLEVR with both the modulated detection as well as question answering losses. We find that without this curriculum strategy, our model converges to a local minimum with sub-par performance as evaluated on the full CLEVR dataset. As depicted in Fig. 3, we use additional queries in the transformer decoder to answer each type of question in CLEVR: numerical, binary and attributes. We supervise each of these heads using a standard cross-entropy loss. Hyperparameter along with more details on the need for a curriculum strategy as well as some ablations on the losses are provided in Appendix B.2 and B.3.

Results The results are collected in Table 1. On CLEVR, we closely match the performance of NS-VQA[68], a method that uses external supervision in the form of ground-truth program annotations, and clearly surpass the performance of methods which like us, don’t use this extra supervision signal. We then evaluate the generalization capability of our model. CLEVR-Humans [21] is a dataset of human-generated questions on CLEVR images. It tests the

robustness of the model to new vocabulary and different reasoning primitives. In a zero-shot setting, we improve substantially over the best competing model. We credit this improvement to our pre-trained language model. After fine-tuning, the gap narrows, suggesting that additional developments may be required to further improve performance.

CoGenT is a test for compositional generalization. The evaluation protocol consists in training on a set A, where the spheres can be any color but the cubes are either gray, blue, brown or yellow, and the cylinders are red, green, purple or cyan. We then evaluate in a zero-shot manner on a split B which has the opposite color-shape pairs for cubes and cylinders. Similar to other models, we observe a significant generalization gap. On closer inspection, the biggest drop in accuracy occurs on questions querying the shape of an object (from 99.98% on testA to 34.68% on testB), suggesting that the model has learnt strong spurious biases between shape and color.

Finally, we evaluate our model on CLEVR-REF+[31], a referring expression comprehension dataset built on CLEVR images. For each object query, we train an additional binary head to predict whether or not the query corresponds to an object being referred to (as opposed to an auxiliary object in the sentence, that we detect as well). Following [31], we evaluate accuracy on the subset of expressions that refer to a unique object, measured as whether the top ranked box has an IoU of at least 0.5 with the target box. Using the aforementioned binary prediction to rank the boxes, our model correctly ranks in first position a valid box for each of the examples of the validation set, leading to an accuracy of 100%, greatly outperforming prior work.

4.2. Natural Images

4.2.1 Pre-training Modulated Detection

For pre-training, we focus on the task of modulated detection where the aim is to detect all the objects that are re-

ferred to in the aligned free form text. We create a combined dataset using images from the Flickr30k [46], MS COCO [30] and Visual Genome (VG) [24] datasets. Annotations from the referring expressions datasets, VG regions, Flickr entities and GQA train balanced set are used for training. Details on the datasets can be found in Appendix C.

Combination For each image, we take all annotations from these datasets and combine the text that refers to the same image while ensuring that all images that are in the validation or testing set for all our downstream tasks are removed from our train set. The combination of sentences is done using a graph coloring algorithm which ensures that only phrases having boxes with $\text{GIoU} \leq 0.5$ are combined, and that the total length of a combined sentence is less than 250 characters. In this way, we arrive at a dataset having 1.3M aligned image - text pairs. This combination step is important for two reasons: 1) data efficiency, by packing more information into a single training example and 2) it provides a better learning signal for our soft token prediction loss since the model has to learn to disambiguate between multiple occurrences of the same object category, as depicted in Fig 4. In the single sentence case, the soft token prediction task becomes trivial since it can always predict the root of the sentence without looking at the image. Experimentally, we find that such dense annotations translate to better grounding between text and image and subsequently to better downstream performance.

Model We use a pre-trained RoBERTa-base [32] as our text encoder, having 12 transformer encoder layers, each with hidden dimension of 768 and 12 heads in the multi-head attention. We use the implementation and weights from HuggingFace [62]. For the visual backbone, we explore two options. The first is a ResNet-101 [12] pretrained on ImageNet with frozen batchnorm layers, taken from Torchvision. This is to be comparable with current literature in the space of multi-modal understanding, where the popular approach is to use the BUTD object detector with a Resnet-101 backbone from [1] trained on the VG dataset. In our work, we are not limited by the existence of pre-trained detectors, and inspired by its success in object detection [58], we choose to explore the EfficientNet family [57] for our backbone. We use a model which was trained on large amounts of unlabelled data in addition to ImageNet, using a pseudo-labelling technique called Noisy-Student [65]. We choose the EfficientNetB3, which achieves 84.1% top 1 accuracy on ImageNet with only 12M weights. We use the implementation provided by the Timm library [61], and freeze the batchnorm layers. Training hyperparameters are detailed in Appendix A.



Figure 4: Our combination of annotations results in examples such as the following: “the person in the grey shirt with a watch on their wrist. the other person wearing a blue sweater. the third person in a gray coat and scarf.” We show the predictions from our model for this caption. It is able to pay attention to all the objects in the image and then disambiguate between them based on the text. The model is trained to predict the root of the phrase as the positive token span, which as we can see in this figure, correctly refers to the three different people.

4.2.2 Downstream Tasks

We evaluate our method on 4 downstream tasks: referring expression comprehension and segmentation, visual question answering, as well as phrase grounding. Training hyperparameters for all tasks can be found in Appendix A.

Phrase grounding Given one or more phrases, which may be inter-related, the task is to provide a set of bounding boxes for each phrase. We use the Flickr30k entities dataset for this task, with the train/val/test splits as provided by [46] and evaluate our performance in terms of Recall@k. For each sentence in the test set, we predict 100 bounding boxes and use the soft token alignment prediction to rank the boxes according to the score given to the token positions that correspond to the phrase. We evaluate under two protocols which we name ANY-BOX [26, 22] and MERGED-BOXES [44]. Please see Appendix D for a discussion on the two protocols. We compare our method to existing state-of-the-art results from two types of approaches - the text conditioned detection models [45, 67] and a transformer based vision-language pre-training model [26]. In the ANY-BOX setting, we obtain a 8.5 point boost over current state of the art on this task as measured in terms of Recall@1 on the validation set, without using any pre-training (no additional data). With pre-training, we further obtain a 12.1 point boost over the best model’s performance on the test

Method	Val			Test		
	R@1	R@5	R@10	R@1	R@5	R@10
ANY-BOX-PROTOCOL						
BAN [22]	-	-	-	69.7	84.2	86.4
VisualBert[26]	68.1	84.0	86.2	-	-	-
VisualBert†[26]	70.4	84.5	86.3	71.3	85.0	86.5
MDETR-R101	78.9	88.8	90.8	-	-	-
MDETR-R101†*	82.5	92.9	94.9	83.4	93.5	95.3
MDETR-ENB3†*	82.9	93.2	95.2	84.0	93.8	95.6
MDETR-ENB5†*	83.6	93.4	95.1	84.3	93.9	95.8
MERGED-BOXES-PROTOCOL						
CITE [43]	-	-	-	61.9	-	-
FAOG [67]	-	-	-	68.7	-	-
SimNet-CCA [45]	-	-	-	71.9	-	-
DDPN [72]	72.8	-	-	73.5	-	-
MDETR-R101	79.0	86.7	88.6	-	-	-
MDETR-R101†*	82.3	91.8	93.7	83.8	92.7	94.4

Table 2: Results on the phrase grounding task on Flickr30k entities dataset [46]. Models with † are pre-trained on COCO, models with * are also pre-trained on VG and Flickr 30k. Our models (MDETR) use a RoBERTa text encoder while other models use RNNs, word2vec-based features, or BERT (comparable to RoBERTa) text encoders. All models use a ResNet101 backbone, except MDETR-ENB3 which uses EfficientNet-B3 and MDETR-ENB5 with an EfficientNet-B5.

set, while using the same backbone.

Referring expression comprehension Given an image and a referring expression in plain text, the task is to localize the object being referred to by returning a bounding box around it. The approach taken by most prior work [70, 34, 6, 69] on this task has been to rank a set of pre-extracted bounding boxes associated with an image, that are obtained using a pre-trained object detector. In this paper, we solve a much harder task - we train our model to directly predict the bounding box, given a referring expression and the associated image. There are three established datasets for this task called RefCOCO, RefCOCO+ [71] and RefCOCOg [36]. Since during pre-training we annotate every object referred to within the text, there is a slight shift in the way the model is used in this task. For example, during pre-training, given the caption “The woman wearing a blue dress standing next to the rose bush.”, MDETR would be trained to predict boxes for all referred objects such as the woman, the blue dress and the rose bush. However, for referring expressions, the task would be to only return one bounding box, which signifies the woman being referred to by the entire expression. For this reason, we finetune the model on the task specific dataset for 5 epochs. At inference time, we use the \emptyset label to rank the 100 detected boxes. Let $P(\emptyset)$ be the probability assigned to the “no object” label, we rank by decreasing order of $1 - P(\emptyset)$. We report results

in Table 3, showing large improvements over state-of-the-art across all datasets.

Referring expression segmentation Similarly to DETR, we show that our approach can be extended to perform segmentation by evaluating on the referring expression segmentation task of the recent PhraseCut [63] dataset which consists of images from VG, annotated with segmentation masks for each referring expression. These expressions comprise a wide vocabulary of objects, attributes and relations, making it a challenging benchmark. Contrary to other referring expression segmentation datasets, in PhraseCut the expression may refer to several objects. The model is expected to find all the corresponding instances. Our training occurs in two stages. In the first step, we take our pre-trained model after 40 epochs and fine-tune it for 5 epochs on this dataset, supervising the model to output correct boxes for the referred expressions. We use the box AP on the validation set for early stopping. In the second stage, following [2], we freeze the weights of the network, and only train a segmentation head for 15 epochs, supervised using a combination of the Dice/F1 loss[38] and the Focal loss [29]. At inference-time, we assign a confidence to each predicted box equal to $1 - P(\emptyset)$ where $P(\emptyset)$ is the probability assigned to the “no-object” token (see §3). We then filter the boxes with a confidence lower than 0.7. Finally, we merge the masks corresponding to each of these boxes into one binary mask corresponding to this referring expression. The results are collected in Table 4. Our model is able to produce clean masks for a wide variety of long tailed-concepts covered by PhraseCut. Example predictions from our model on this dataset are given in Appendix A.

Visual Question Answering We evaluate our hypothesis that modulated detection is a useful component for multi-modal reasoning by fine-tuning our pre-trained model on the GQA dataset. Our model architecture is depicted in Fig 3. Apart from the 100 queries that are used for detection, we use additional object queries that specialize in the type of question as well as one that is used to predict the type of question, where the types are defined in the GQA annotations as REL, OBJ, GLOBAL, CAT and ATTR. We take our pre-trained model trained for 40 epochs on our combined dataset, and initialise these object queries as well as the heads for each of them randomly, and fine-tune first for 5 epochs on the unbalanced *all* GQA split, followed by 10 epochs on the *balanced* split similar to what is done in prior work [28, 5]. During the first 5 epochs, we train the modulated detection losses along with the question answering, but put a weight on question answering loss that encourages the model to focus more on this task. For the balanced split fine-tuning, we only use the question answering loss. During inference, the type head predicts the type

Method	Detection backbone	Pre-training image data	RefCOCO			RefCOCO+			RefCOCOg	
			val	testA	testB	val	testA	testB	val	test
MAttNet[70]	R101	None	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27
ViLBERT[34]	R101	CC (3.3M)	-	-	-	72.34	78.52	62.61	-	-
VL-BERT_L [54]	R101	CC (3.3M)	-	-	-	72.59	78.57	62.30	-	-
UNITER_L[6]*	R101	CC, SBU, COCO, VG (4.6M)	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77
VILLA_L[9]*	R101	CC, SBU, COCO, VG (4.6M)	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71
ERNIE-ViLL[69]	R101	CC, SBU (4.3M)	-	-	-	75.95	82.07	66.88	-	-
MDETR	R101	COCO, VG, Flickr30k (200k)	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
MDETR	ENB3	COCO, VG, Flickr30k (200k)	87.51	90.40	82.67	81.13	85.52	72.96	83.35	83.31

Table 3: Accuracy results on referring expression comprehension. *As mentioned in UNITER [6], methods using box proposals from the BUTD detector [1] suffer from a test set leak, since the detector was trained on images including the validation and test set of the RE comprehension datasets. We report numbers for these methods from their papers using these “contaminated features” but we would like to stress that all of our pre-training excluded the images used in the val/test of any of the downstream datasets including for RE comprehension. CC refers to Conceptual Captions [53], VG to Visual Genome [24], SBU refers to the SBU Captions[41] and COCO to Microsoft COCO [30].

Method	Backbone	PhraseCut				
		M-IoU	Pr@0.5	Pr@0.7	Pr@0.9	
RMI[4]	R101	21.1	22.0	11.6	1.5	
HULANet[63]	R101	41.3	42.4	27.0	5.7	
MDETR	R101	53.1	56.1	38.9	11.9	
MDETR	ENB3	53.7	57.5	39.9	11.9	

Table 4: Following [63], we report the mean intersection-over-union (IoU) of our masks with the ground-truth masks. We also report the precision Pr@ I of our model, where success is marked when our proposed mask has an IoU with the ground-truth higher than the threshold I . With a comparable ResNet backbone, we observe consistent gains across all metrics over HULANet [63], the current state-of-the-art. The EfficientNet backbone further improves on those results.

of question and the answer is taken from that head. Using our model with a Resnet-101 backbone, we not only outperform LXMERT [55] and VL-T5 [7] which use comparable amount of data, but also OSCAR [28] which uses magnitude more data in their pre-training. MDETR with the EfficientNet-B5 backbone is able to push performance even higher as reported in Table 5. The NSM model makes use of an external scene graph generation model, while the MMN model makes use of the scene graph and functional programs during training. To train MDETR, we use the scene graph to obtain the alignment between question words and the boxes.

4.2.3 Few-shot transfer for long-tailed detection

Inspired by the success of CLIP [47], on zero-shot transfer for image classification, we explore the opportunity to construct a useful detector over a given label set from a pre-trained MDETR model. Unlike CLIP, we do not ensure our

Method	Pre-training img data	Test-dev	Test-std
MoVie [39]	-	-	57.10
LXMERT[55]	VG, COCO (180k)	60.0	60.33
VL-T5 [7]	VG, COCO (180k)	-	60.80
MMN [5]	-	-	60.83
OSCAR [28]	VG, COCO, Flickr, SBU (4.3M)	61.58	61.62
NSM [19]	-	-	63.17
VinVL [73]	VG, COCO, Objects365, SBU Flickr30k, CC, VQA, OpenImagesV5 (5.65M)	65.05	64.65
MDETR-R101	VG, COCO, Flickr30k (200k)	62.48	61.99
MDETR-ENB5	VG, COCO, Flickr30k (200k)	62.95	62.45

Table 5: Visual question answering on the GQA dataset.

pre-training dataset contains a balanced representation of all the target classes. By construction, our dataset has no training instances where there are zero boxes aligned to the text, biasing the model to always predict boxes for a given text. This prevents evaluating in a true zero-shot transfer setting, so we turn instead to a few-shot setting, where the model is trained on a fraction of the available labelled data. We conduct our experiments on the LVIS dataset [11], a detection dataset with a large vocabulary of 1.2k categories, with a long-tail that contains very few training samples, making it a challenging dataset for current approaches. Federated datasets often pose problems to standard detectors, and require developing specific loss functions [56]. However this property makes it well suited to train MDETR: for each positive category, we create a training instance composed of the image and a text version of the class name, and provide as

Method	Data	AP	AP50	AP _r	AP _c	AP _f
Mask R-CNN	100%	33.3	51.1	26.3	34.0	33.9
DETR	1%	4.2	7.0	1.9	1.1	7.3
DETR	10%	13.7	21.7	4.1	13.2	15.9
DETR	100%	17.8	27.5	3.2	12.9	24.8
MDETR	1%	16.7	25.8	11.2	14.6	19.5
MDETR	10%	24.2	38.0	20.9	24.9	24.3
MDETR	100%	22.5	35.2	7.4	22.7	25.0

Table 6: Box AP *fixed* results on LVIS-v1. Since the validation set of LVIS contains some training images from MSCOCO, we report results on the subset of 5k validation images that our model has never seen during training. We call this subset *minival*. All models use a Resnet 101 as backbone. Mask-RCNN can be regarded as a strong representative of the detection performance of current approaches on this dataset, using bells and whistles such as Repeat Factor Sampling (RFS) to address class imbalance. We use a vanilla DETR pretrained on MSCOCO as a few-shot transfer baseline, and show that our pre-training on natural text improves performance significantly, especially on rare categories.

annotations all the instances of this category. For each negative category, we provide the class name and an empty set of annotations. For inference on a given image, we query each possible class name, then merge the sets of boxes detected on each of the text prompts. This inference scheme costs about 10s/image on a GPU.

We fine-tune MDETR on three subsets of the LVIS train set, each containing respectively 1%, 10% and 100% of the images. We ensure a balanced sampling of the categories, such that our 1% set contains at least one positive and one negative examples from each category. We compare to two baselines: the first one is Mask-RCNN trained exclusively on the full training set of LVIS. The other is a DETR model pre-trained on MSCOCO then fine-tuned on the various subsets of the LVIS training set. Our results are shown in Table 6. Following recent recommendation [8] on AP evaluation in the context of large vocabulary, we report the box AP *fixed*, obtained by limiting the number of detections per category instead of per image. Even with as little as 1 example per class, MDETR leverages the text pre-training and outperforms a fully fine-tuned DETR on rare categories. We note however that under full fine-tuning on the whole training set, the performance on small objects drops significantly from 20.9 AP with 10% data to 7.5 with 100%, likely due the extreme class imbalance. We expect that common techniques such as Repeat Factor Sampling will improve the situation in future work.

5. Conclusion

We presented MDETR, a fully differentiable modulated detector. We established its strong performance on multi-modal understanding tasks on a variety of datasets, and

demonstrated its potential in other downstream applications such as few-shot detection and visual question answering. We hope that this work opens up new opportunities to develop fully integrated multi-modal architectures, without relying on black-box object detectors.

Acknowledgements

We would like to thank Kyunghyun Cho, Ethan Perez, Sergey Zagoruyko and Francisco Massa for helpful discussions and feedback at various points of this project. We would also like to thank Alex Kirillov and Ross Girshick for their help with the LVIS evaluations, Justin Johnson for test set evaluation on CLEVR, Bryan Plummer for discussions on best evaluation practices for phrase grounding and finally Runtao Liu and Chenxi Liu for their feedback on dataset construction and evaluation for CLEVR referring expressions.

Aishwarya Kamath was supported in part by AFOSR award FA9550-19-1-0343 and Nicolas Carion by a grant from NVIDIA.

References

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. IEEE Computer Society, 2018. [2](#), [6](#), [8](#)
- [2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. [1](#), [3](#), [7](#), [16](#)
- [3] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7453–7462. IEEE, 2019. [3](#)
- [4] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7453–7462. IEEE, 2019. [8](#)
- [5] Wenhui Chen, Zhe Gan, Linjie Li, Yu Cheng, William Wang, and Jingjing Liu. Meta module network for compositional visual reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 655–664, 2021. [7](#), [8](#)
- [6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019. [1](#), [2](#), [7](#), [8](#)

- [7] Jaemin Cho, Jie Lei, H. Tan, and M. Bansal. Unifying vision-and-language tasks via text generation. *ArXiv*, abs/2102.02779, 2021. 3, 8
- [8] Achal Dave, Piotr Dollár, Deva Ramanan, Alexander Kirillov, and Ross Girshick. Evaluating Large-Vocabulary Object Detectors: The Devil is in the Details. 9
- [9] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3, 8
- [10] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society, 2017. 21
- [11] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5356–5364. Computer Vision Foundation / IEEE, 2019. 8
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 3, 4, 6
- [13] Ryota Hinami and Shin’ichi Satoh. Discriminative learning of open-vocabulary object retrieval and localization by negative phrase augmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2605–2615, Brussels, Belgium, 2018. Association for Computational Linguistics. 1, 3
- [14] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 15, 19
- [15] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 804–813. IEEE Computer Society, 2017. 2
- [16] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. 2021. 3
- [17] Drew A. Hudson and Christopher D. Manning. Compositional attention networks for machine reasoning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 2, 4, 5
- [18] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE, 2019. 20
- [19] Drew A. Hudson and Christopher D. Manning. Learning by abstraction: The neural state machine. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5901–5914, 2019. 8
- [20] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society, 2017. 2, 4
- [21] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. Inferring and executing programs for visual reasoning. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3008–3017. IEEE Computer Society, 2017. 2, 4, 5
- [22] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1571–1581, 2018. 6, 7, 20
- [23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 14
- [24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 6, 8
- [25] Chenliang Li, Ming Yan, Haiyang Xu, Fuli Luo, Wei Wang, Bin Bi, and Songfang Huang. Semvlp: Vision-language pre-training by aligning semantics at multiple levels, 2021. 2
- [26] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2, 6, 7, 20
- [27] Ruiyu Li, Kai-Can Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5745–5753. IEEE Computer Society, 2018. 3
- [28] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for

- vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 1, 2, 7, 8
- [29] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society, 2017. 7
- [30] Tsung-Yi Lin, M. Maire, Serge J. Belongie, James Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6, 8
- [31] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L. Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4185–4194. Computer Vision Foundation / IEEE, 2019. 5
- [32] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 6
- [33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 14
- [34] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23, 2019. 2, 7, 8
- [35] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10434–10443. IEEE, 2020. 2
- [36] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 11–20. IEEE Computer Society, 2016. 7, 19
- [37] David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4942–4950. IEEE Computer Society, 2018. 2, 4
- [38] F. Milletari, N. Navab, and S. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, 2016. 7
- [39] Duy-Kien Nguyen, Vedanuj Goswami, and Xinlei Chen. MoVie: Revisiting Modulated Convolutions for Visual Counting and Beyond. *arXiv:2004.11883 [cs]*, 2020. arXiv: 2004.11883. 2, 8
- [40] A. Oord, Y. Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 4
- [41] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1143–1151, 2011. 8
- [42] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3942–3951. AAAI Press, 2018. 2, 5
- [43] Bryan A Plummer, Paige Kordas, M Hadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Svetlana Lazebnik. Conditional image-text embedding networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 249–264, 2018. 7
- [44] Bryan A. Plummer, Arun Mallya, Christopher M. Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase localization and visual relationship detection with comprehensive image-language cues. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1946–1955. IEEE Computer Society, 2017. 6
- [45] Bryan Allen Plummer, Kevin Shih, Yichen Li, Ke Xu, Svetlana Lazebnik, Stan Sclaroff, and Kate Saenko. Revisiting image-language networks for open-ended phrase detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 3, 6, 7, 20
- [46] 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 *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society, 2015. 6, 7, 20
- [47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 4, 8
- [48] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8945–8954. IEEE, 2019. 8

- 16-20, 2019, pages 658–666. Computer Vision Foundation / IEEE, 2019. 3
- [49] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115, 2015. 4
- [50] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019. 4
- [51] Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia, and Tim Lillicrap. A simple neural network module for relational reasoning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4967–4976, 2017. 2
- [52] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, 2016. Association for Computational Linguistics. 4
- [53] 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, Melbourne, Australia, 2018. Association for Computational Linguistics. 2, 8
- [54] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 2, 8
- [55] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, 2019. Association for Computational Linguistics. 2, 8
- [56] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 11659–11668. IEEE, 2020. 8
- [57] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019. 6
- [58] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10778–10787. IEEE, 2020. 6
- [59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 3
- [60] Zhonghao Wang, Mo Yu, Kai Wang, Jinjun Xiong, Wen-mei Hwu, Mark Hasegawa-Johnson, and Humphrey Shi. Interpretable Visual Reasoning via Induced Symbolic Space. 2, 5
- [61] Ross Wightman. Pytorch image models. <https://github.com/rwightman/pytorch-image-models>, 2019. 6
- [62] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, 2020. Association for Computational Linguistics. 4, 6
- [63] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasicut: Language-based image segmentation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10213–10222. IEEE, 2020. 7, 8
- [64] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3733–3742. IEEE Computer Society, 2018. 4
- [65] Qizhe Xie, Minh-Thang Luong, Eduard H. Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10684–10695. IEEE, 2020. 6
- [66] Jianwei Yang, Jiayuan Mao, Jiajun Wu, Devi Parikh, David D. Cox, Joshua B. Tenenbaum, and Chuang Gan. Object-Centric Diagnosis of Visual Reasoning. 2
- [67] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*,

Seoul, Korea (South), October 27 - November 2, 2019, pages 4682–4692. IEEE, 2019. 1, 3, 6, 7, 20

- [68] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic VQA: disentangling reasoning from vision and language understanding. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1039–1050, 2018. 2, 4, 5
- [69] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020. 3, 7, 8
- [70] 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 *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1307–1315. IEEE Computer Society, 2018. 5, 7, 8
- [71] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling Context in Referring Expressions. *arXiv:1608.00272 [cs]*, 2016. arXiv: 1608.00272 version: 3. 7, 19
- [72] Zhou Yu, Jun Yu, Chenchao Xiang, Zhou Zhao, Qi Tian, and Dacheng Tao. Rethinking diversified and discriminative proposal generation for visual grounding. In Jérôme Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 1114–1120. ijcai.org, 2018. 7
- [73] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *arXiv preprint arXiv:2101.00529*, 2021. 1, 2, 8, 17
- [74] Yihan Zheng, Zhiqian Wen, Mingkui Tan, Runhao Zeng, Qi Chen, Yaowei Wang, and Qi Wu. Modular graph attention network for complex visual relational reasoning. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. 5

A. Model details and hyperparameters

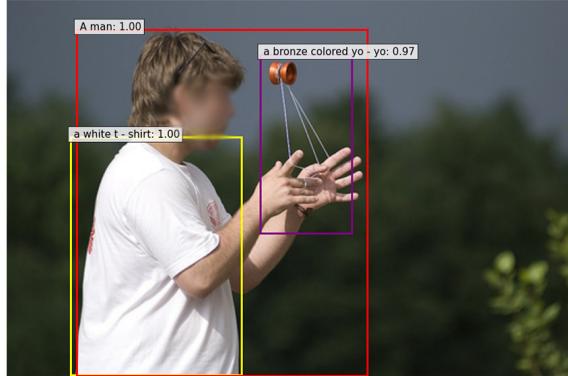
Pre-training hyperparameters MDETR follows the pre-train then fine-tune strategy by first training on our constructed combined dataset for 40 epochs followed by fine-tuning on the respective downstream task. We train our model using AdamW [33], a variant of Adam [23] better suited for weight decay. We use exponential moving average (EMA) with a decay rate of 0.9998, and a weight-decay of $1e^{-4}$. The backbone and the transformer have a constant learning rate of respectively $1e^{-4}$ and $1e^{-5}$ for 35 epochs, after which their learning rate is reduced by a factor of 10. For the language model’s learning rate, we use a linear decay with warmup schedule, increasing linearly to $5e^{-5}$ during the first 1% of the total number of steps, then decreasing linearly back to 0 for the rest of the training.

Flickr30k Our results on Flickr30k are evaluated using the pre-trained model, without any additional fine-tuning as we found that it brings no additional gains. For evaluation, we must rank the boxes associated with each phrase. Since there might be several phrase in the same sentence, we must provide a ranking for each and every such phrase. To that end, we use the prediction from the soft-token classification loss. In the example depicted in Fig. 9: to rank the boxes for the phrase “a cat”, we use the probability mass that each query assigns to the positions that correspond to “a cat” in the sentence “a cat with white paws jumps over a fence in front of a yellow tree” (in this example, the first few tokens). Through this approach, the red box is found to be the highest-ranked box. On the other hand, if we want the boxes corresponding to “the fence”, we sort them according to the corresponding token positions, and in this case we find the green box as the top-scoring one.



(a) “one small boy climbing a pole with the help of another boy on the ground”

(b) “A man talking on his cellphone next to a jewelry store”

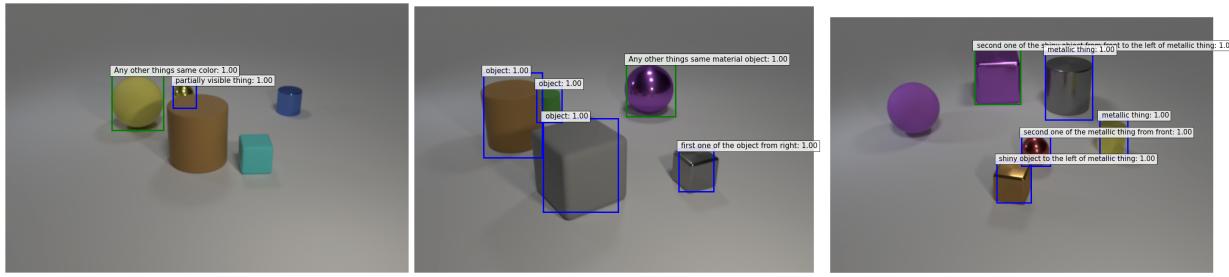


(c) “A man in a white t-shirt does a trick with a bronze colored yo-yo”

Figure 5: Examples of phrase grounding on the Flickr30k dataset



Figure 6: Qualitative segmentation examples on the phrasicut dataset



(a) Query: “Any other things that are the same color as the partially visible thing(s)” **(b)** Query: “Any other things that are the same material as the first one of the object(s) from right” **(c)** Query: “The second one of the shiny object(s) from front that are to the left of the second one of the metallic thing(s) from front”

Figure 7: Qualitative example from the CLEVR-REF+ dataset. When the model predicts a box that is referred to, we display it in green. The other boxes are intermediate reasoning steps and are depicted in blue.

Referring Expression Comprehension For the Referring Expression Comprehension task, there is a stark difference in how the data is presented to the model in terms of density of annotation. For all other datasets that we use in our pre-training, each noun phrase in the sentence is annotated with its respective box, if available. On the other hand, in RE comprehension, the task is to align the whole referring expression with the corresponding box, possibly by needing to disambiguate between different occurrences of the same category of object. In other words, our model now needs to predict one box per expression. A problem with this setting is that our box-token contrastive alignment as well as soft token prediction losses get very diluted signal if we align the whole sentence to the box. To alleviate this, we pre-process the text using Spacy [14] to extract the root of the sentence using a dependency parser. The tokens from this root phrase are used to align to the box. Fine-tuning on this dataset is therefore crucial for good performance. We fine-tune for 5 epochs on the RefCOCO, RefCOCO+ and RefCOCOg datasets with a learning rate of $1e^{-5}$ for the backbone and $5e^{-5}$ for the rest of the network. We use a learning rate drop by a factor of 10 after 3 epochs. For the text encoder we use a learning rate of $1e^{-5}$, with a linear decay with warmup schedule, warming up over the first 1% of steps and then decaying to 0 linearly. At inference time, to detect a given expression, we feed it to the model alongside the image. We then rank the 100 detected boxes according to the probability that the box corresponds to an actual object (as opposed to a “no object”). If $P(\emptyset)$ is the probability mass assigned to the “no object” label, then we rank by increasing order of $P(\emptyset)$, or equivalently by decreasing order of $1 - P(\emptyset)$. We show an example in Fig 8 of the box predicted by our model for the corresponding referring expressions. In addition there is quite some variety in the type of text annotations from the three datasets. Both RefCOCO and RefCOCO+ were collected in a timed game setting whereas RefCOCOg was not. This led to differences in the length and diversity of language used in the different datasets. RefCOCO+ disallowed usage of location words to describe objects or disambiguate between multiple occurrences of the same object, focusing more on appearance based descriptions. RefCOCOg consists of expressions more than twice the length (on average) of the others and with more flowery and descriptive language.

We also fine-tuned the EfficientNetB5 model on these datasets but did not see much improvement over the EfficientNetB3 model, and we believe this is due to the smaller size of these datasets causing the larger model to overfit.



Figure 8: Examples from RefCOCO, RefCOCO+ and RefCOCOg datasets. Fig(a) taken from RefCOCO, Fig(b) from RefCOCO+ and Fig(c) and (d) are taken from RefCOCOg, in which the expressions are much longer on average and contain more descriptive language than in RefCOCO and RefCOCO+. Even when the expressions are long, we train our model to align the box to the root of the phrase, for eg. “the man” in (c). The model however, still has access to the whole text and uses it to disambiguate between the two men in the image.

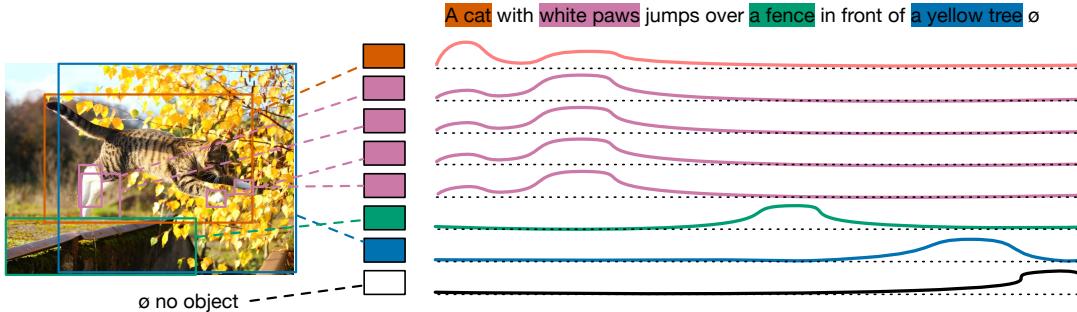


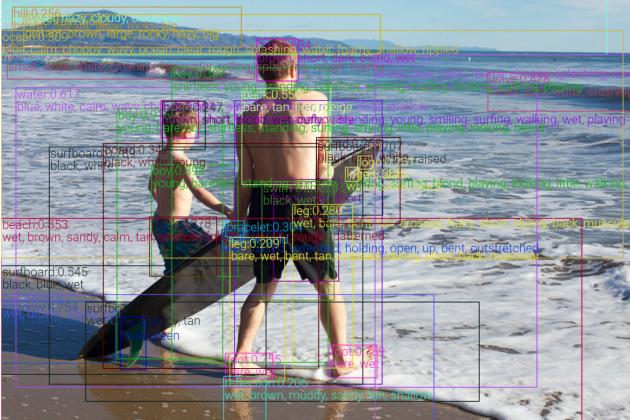
Figure 9: Illustration of the soft-token classification loss. For each object, the model predicts a distribution over the token positions in the input sequence. The weight of the distribution should be equally spread over all the tokens that refer to the predicted box.

PhraseCut Detection: For this phase, we use a batch size of 64, a learning rate of $1e^{-5}$ for the text encoder and backbone and $5e^{-5}$ for the rest of the network, and exponential moving average (EMA) of the network weights with a decay of 0.9998. segmentation: For this stage, we use a lr of $5e^{-4}$ and no EMA. See [2] for additional details.

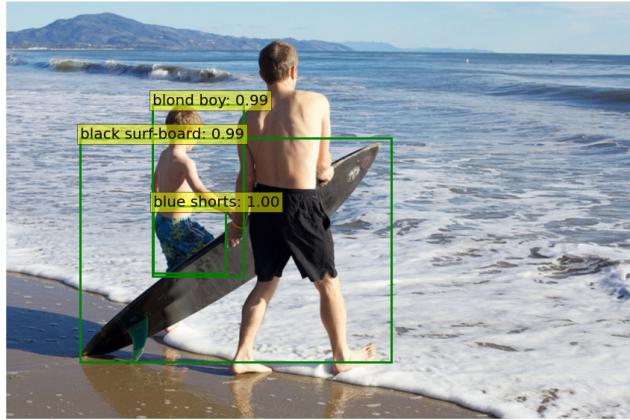
B. CLEVR Experiments

B.1. Dataset details

The CLEVR dataset consists of 3D-rendered scenes containing between 3 and 10 objects of various shapes, material, size and color. Each of these scenes is associated with about 10 questions that are formulated about the visible objects,



(a) Current object detection pipeline outputs, predicting all possible objects in the image. This extensive annotation is essential to multi-modal understanding systems that treat detection as a black box.



(b) MDETR predicts boxes relevant to the caption and labels them with the corresponding spans from the text. Here we use the caption: “blond boy wearing blue shorts. a black surf-board.”

Figure 10: Modulated detection using MDETR vs detection output for a current state-of-the-art multi-modal understanding system. Image taken from [73]

generated from a fixed set of templates. Each question is guaranteed to be answerable, and the annotations further provide a functional program that describe how to compute the answer using elementary reasoning steps. The total training set contains 70k images and slightly less than 700k questions. Overall, the visual aspect of this task, *ie* the scene parsing, is not really challenging by modern standards, since the set of objects is limited, unambiguous, and there are no visual distractors. The only challenging cases occur in the event of heavy occlusion, where it might be hard to make out the shape of the occluded object, or in some cases where the question requires comparing ambiguous spatial relations (eg. asking which object is the closest to the camera in a setting where they are visually nearly tied). On the other hand, the text understanding aspect is more involved, since the questions can be quite complex, involving up to 20 reasoning steps.

The first ingredient required for training MDETR is bounding box annotations for objects in the image. The original CLEVR dataset doesn’t provide any, so we use the scene graphs from the dataset to re-create the original scene in the 3D-renderer Blender, then use some of its functionalities to extract the segmentation masks of the visible parts of the objects, and deduce the bounding boxes from that. The main complication is that the original rendering involved some non-deterministic jittering of the camera’s position and rotation, leading to some potential discrepancies in the computed boxes. To minimize the error, we use the known 3D position of each object, as well as their known 2D location in the rendered image to optimize the camera parameters using a gradient-based approach. The final boxes obtained using this approach are accurate within a 10-pixel error margin, which we deem appropriate for our purposes.

The second ingredient required is the alignment between bounding boxes and tokens in the question. MDETR is trained to predict *only* objects that are referred to in the question. For example, in the question “What is the color of the cube in front of the small cylinder?”, we provide an annotation for both the small cylinder (an intermediate step) and the cube (the main subject), and none of the other objects present in the scene. We use the functional programs that are part of the original CLEVR annotations to extract this set of objects, along with their spans in the original question.

B.2. Training details

Pre-training We first train the model only on the modulated detection objective, on our CLEVR-Medium subset, for 30 epochs. Following DETR training procedure, the transformer and the backbone use a learning rate of respectively $1e^{-4}$ and $1e^{-5}$, and we reduce them by a factor of 10 at the 20th epoch. The text encoder uses a linear decay with warmup schedule, with a warm-up to $5e^{-5}$ over the first 1% of the training steps.

QA-finetuning For question answering, we take our pre-trained checkpoint, add the (untrained) question queries and their corresponding heads, then train on the full CLEVR dataset for 30 epochs, following the exact same learning rate schedule. We monitor the accuracy on the validation set to apply early-stopping. Finally, for CLEVR-Humans, we further fine-tune for 60 epochs, with the learning-rate drop occurring at epoch 40.

CLEVR-Ref+ For the referring expressions task, we follow the same recipe: starting from the pre-trained model, we fine-tune for 30 epochs using the same learning-rate schedule.

Model	AP
Baseline	99.0
- contrastive loss	83.2
- soft-token classification	87.7

Table 7: Ablation results on modulated detection on CLEVR-Medium. We report the class-agnostic AP. See text for details.

B.3. Ablations

We use CLEVR as a test bed to ablate several aspects of our model. Depending on the ablation, we report either the accuracy on the question answering task on the validation set of CLEVR, and/or the detection performance on this dataset, measured as a class-agnostic Average Precision (AP). When inspecting the modulated detection capabilities of the model, we use class agnostic Average Precision (AP) to evaluate the model. In the unconditional detection case, DETR is able to detect all boxes perfectly. When evaluated on the task of modulated detection, the AP metric therefore captures the model’s capability for text understanding since now only the boxes relevant to the query must be detected. To put this in context, a model that detects all boxes even when given a text query (thereby ignoring the text completely) gets an AP of around 60. The goal is to achieve an AP close to 100 which would imply the model only finds the relevant boxes.

B.3.1 Loss ablations

We first ablate the various parts of our loss. The results are summarized in Table 7. We report modulated detection results on the CLEVR-Medium subset, that we constructed by removing data-points from CLEVR where the same object is referred to by distinct parts of the question. In these ablations, we consider only the performance of the detector, and not the question answering capability. As a result, the queries related to question answering are not present and we do not propagate any QA related loss.

Contrastive loss We first ablate the impact of the contrastive loss, by training a model without it. In this situation, the alignment must occur solely through the soft-token classification loss. As shown in Table 7, removing this loss results in a drastic drop in AP. More specifically, when evaluating the model, it becomes apparent that it is able to filter the objects based on some attributes (in particular their shape and size) but not others (in particular color and texture). It is unclear what drives the model in this local sub-optima, nor what statistical shortcut it is leveraging to correctly identify shapes and sizes. However, it shows that solely predicting the spans of the text query associated with each object is not sufficient to learn proper alignment. The contrastive loss, which forces object-queries to be similar to their corresponding text-token, is thus necessary.

Soft-token classification loss We now study whether predicting the text-spans associated with each object is necessary, provided that we propagate the contrastive loss. Instead of predicting a distribution over span, we construct a simplified version of MDETR which only predicts a binary label for each object query: “object” or “no-object” (\emptyset). This formulation is equivalent to the vanilla DETR classification loss, with one object class. We observe similar results as the previous ablation, namely a sharp decline in AP and a model that only understands half of the attributes correctly. We thus conclude that both ingredients of our loss are indeed required.

B.3.2 Question answering ablations

Finally in Table 8 we ablate two aspects of our training recipe that differ with previous approaches:

- **Curriculum:** We evaluate a model trained directly on the full CLEVR training set, without our modulated detection pretraining on CLEVR-medium. Similarly as the previous section, the model learns to detect only a subset of attributes, leading to poor QA accuracy.
- **Single QA head** In our approach, we train a specialized head for each type of question (numerical, binary, or categorical over the attributes). This differs from previous approaches that usually cast it as a single classification over all possible answers. As shown in Table 8, this separation has a big impact on the final accuracy. We hypothesize that it enables the attention pattern for each question type to specialize accordingly to the task, there-by yielding better performance.

Model	Detection AP	QA accuracy
Baseline	99.0	99.7
No curriculum	89.7	68.2
Single QA head	99.0	90.1

Table 8: Ablation results on the validation set of CLEVR. We report the class-agnostic AP and the question answering accuracy. See text for details.



Figure 11: Predicted boxes for the expression : “the person in the grey shirt with a watch on their wrist. the other person wearing a blue sweater. the third person in a gray coat and scarf.”

C. Dataset constructions

MS COCO On the COCO dataset, we include annotations from the referring expressions datasets (RefCOCO [71], RefCOCO+ [71] and RefCOCOg [36] datasets). By construction, in this dataset, each referring expression is a whole sentence that describes one object in the image, where the constituent noun phrases from the sentences are not themselves annotated. For example, in Figure 11, the caption would be “the person in the grey shirt with a watch on their wrist.”, where only the person would be annotated and not the grey shirt or their watch. To avoid ambiguity, we perform some text pre-processing using SpaCy [14] to extract the *root* of the referring expression. This is used in our soft token prediction as well as the contrastive alignment loss for aligning to the referred box. The auxiliary objects (in this example the shirt and the watch) are ignored altogether.

Visual-Genome We use annotations from VG regions, a dataset having diverse descriptions of a wide variety of objects, often having a very high degree of descriptive detail and covering several concepts. By construction, the VG dataset comprises a lot of redundant annotations. We detect redundant sentences by normalizing them (removing all punctuation, stop-words, and lower-casing), then testing for equality. Once we found a pair of equivalent sentences, two cases arise:

- The corresponding boxes are highly overlapping ($\text{IoU} > 0.7$). In this case, we consider both annotations to be redundant, and we keep only one of them.
- The boxes are non-overlapping. The most likely explanation is that the sentence is under-specified and actually corresponds to several distinct objects in the image. In this case, we merge the two data-points together, and the resulting annotations comports two boxes for this sentence.

We iterate recursively this process until no equivalent sentences remain.

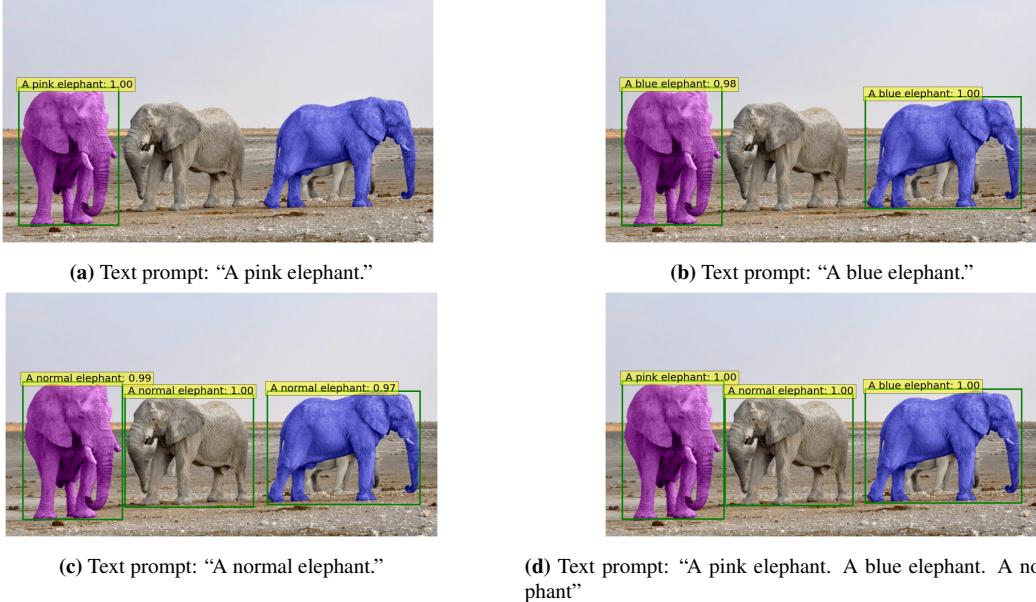


Figure 12: Qualitative results on unseen attributes combinations. While the model correctly singles out the pink elephant (a), it incorrectly includes the pink elephant when prompted about the blue one (b). In (c), we show that the model does not understand what a “normal elephant” looks like. However, in (d), when prompted about all three elephants at once, it is able to assign the correct label to each of them, by process of elimination.

In some cases, the VG annotations provide information about the object referred to in the sentence. For example, if the region is tagged “the cat on the white table”, in some cases the individual boxes for the cat and the table are available. In this case, we discard the region box and use the individual boxes instead.

We note that despite our merging strategy, it may remain some region description that do not canonically refer to a unique object in the image, but for which we don’t have ground-truth annotations for the other objects that also match the said region description. Despite the noise it introduces in the training process, we don’t pursue extra efforts to try and fix these situations.

In addition, we also use questions from the GQA dataset [18], where bounding box annotations are provided for key phrases in the questions.

D. Evaluating grounded detection

The main evaluation metric proposed to evaluate grounded detection in datasets like Flickr30K entities[46] is Recall@ k , that is measuring whether a given model is able to rank the “correct” box amongst the top k it produces. The correctness of a box is decided by computing the Intersection-over-Union (IoU) between the proposed box and the ground-truth box, and deemed correct if the IoU is above a predetermined threshold, generally 0.5. While this kind of evaluation is well-suited for tasks where there is a clear one-to-one mapping between phrases and boxes, for example in Referring Expression Comprehension tasks, we argue that in general grounded detection tasks, they fall short of properly evaluating the performance of the models. Specifically, they run into the following issues:

1. **Multiple boxes for a given phrase:** Since the recall@ k metric implies a single box per phrase, it is not clear how to extend it to situations where a given phrase refers to several distinct objects in the image . In the absence of clear guidelines, various authors have adopted divergent approaches to deal with that. Specifically, some [26, 22] consider the predicted box to be correct if it has an $\text{IoU} > 0.5$ with *any* of the ground-truth boxes. We refer to this protocol as the ANY-Protocol. Others [45, 67] first merge all the ground-truth boxes associated to the phrase by considering the smallest enclosing box. Then they proceed to compute the IoU as usual, using this merged box as the ground-truth one. We refer to this protocol as the MERGED-BOXES-Protocol.

Arguably, both methods have drawbacks: the first one keeps the atomicity of each instance but fails to evaluate whether the model found all the referred instances. The second one loses the fine-grained instance in favor of a box that may be unreasonably bloated if the instances are spread apart

Pre-training	Test-Dev				Test-Std			
	Overall	Yes-no	Other	Number	Overall	Yes-no	Other	Number
Modulated detection on combined dataset (40 epochs)	70.49	86.74	55.17	60.04	-	-	-	-
+ GQA-all (5 epochs)	70.64	86.74	55.26	60.33	70.63	86.79	55.12	60.12

Table 9: VQA v2 results

2. **Multiple phrases for a given box.** In some cases, the same object is referred to multiple times in the sentence. Sometimes, the corresponding phrases are exact duplicates of each other , or close synonyms (eg “a guy” and “a man”), but sometimes the co-references are more subtle. One such example include referring alternatively to a group (eg “a couple”) or to a sub-constituent (eg “the woman”).

Under the current evaluation protocol, each phrase is evaluated independently, even if they refer to the same object. As such, it does not test the model’s understanding of co-references, which is arguably an important aspect of learning grounded representations.

Recognising the discrepancy in the evaluation procedures in the literature and the difficulties it creates in comparing various approaches, we also evaluate MDETR under the merged-box protocol, as described in Sect D.1.

D.1. Evaluation under the MERGED-BOXES-Protocol

In MDETR, the dataset creation process operates under the assumption that a phrase is associated to all the individual boxes that correspond to it. As a result, MDETR does not naturally predict the merged-boxes that would be required by the MERGED-BOXES-Protocol. For that reason, it is necessary to fine-tune the model on a version of Flickr30k entities where the boxes have been merged appropriately. We note that the ANY-Protocol did not require such fine-tuning.

E. Experiments on VQA2

In our visual question answering experiments on the GQA dataset, we always had access to bounding box information for the questions. Only during fine-tuning on the balanced set for 10 epochs, we do not supervise the detection losses. On the other hand, for datasets such as VQA2 [10], we do not have access to any box annotations and the supervision comes solely from the question answering loss. We fine-tune two of our models — pre-trained on the joint dataset, as well as the model fine-tuned on the all-split of GQA — on the VQA v2 dataset for 25 epochs. The results are reported in Table 9. While these results are not state-of-the-art on the VQA2 benchmark, they are still quite reasonable with respect to current literature. This show that our method can be extended to tasks where we do not have the dense supervision (in the form of bounding boxes and their alignment to the text) that we otherwise expect in tasks reported in this paper.