DetermiNet: A Large-Scale Diagnostic Dataset for Complex Visually-Grounded Referencing using Determiners

Clarence Lee*[1], M Ganesh Kumar*[2], Cheston Tan[2]
Design and Artificial Intelligence, SUTD [1], Centre for Frontier AI Research, A*STAR [2]
clarence_leesheng@mymail.sutd.edu.sg, m_ganeshkumar@u.nus.edu, cheston_tan@cfar.a-star.edu.sg

Abstract

State-of-the-art visual grounding models can achieve high detection accuracy, but they are not designed to distinguish between all objects versus only certain objects of interest. In natural language, in order to specify a particular object or set of objects of interest, humans use determiners such as "my", "either" and "those". Determiners, as an important word class, are a type of schema in natural language about the reference or quantity of the noun. Existing grounded referencing datasets place much less emphasis on determiners, compared to other word classes such as nouns, verbs and adjectives. This makes it difficult to develop models that understand the full variety and complexity of object referencing. Thus, we have developed and released the DetermiNet dataset [1], which comprises 250,000 synthetically generated images and captions based on 25 determiners. The task is to predict bounding boxes to identify objects of interest, constrained by the semantics of the given determiner. We find that current state-of-the-art visual grounding models do not perform well on the dataset, highlighting the limitations of existing models on reference and quantification tasks.

1. Introduction

Humans combine visual and linguistic cues to perform object localization, referencing and quantification tasks on a daily basis. For example, when someone says "pass me a cup", we first locate any cups present, and then select one cup based on other criterias, such as the nearest or cleanest one. Deep learning models [5, 9, 10, 11, 16, 19, 29, 37, 39, 41] can localize object impressively to achieve the first part of the task. However, the ability to deal with a variety of complex referencing and quantification to achieve the second part of the tasks has yet to be properly investigated.

A *determiner* is an English part-of-speech (word class) that quantifies or references the noun following it. For instance, the determiner in "my apple" versus "your apple" takes reference from different owners. The number of apples being referenced differs for "some apples" versus "all apples". Such semantic differences are succinctly captured by determiners, and not by other word classes.

Determiners like "a", "the" and "my" are ubiquitous and among the most common English words [1, 22]. Most children learn to use determiners at a near-mastery level by 3 years of age [3, 6]. Since determiners play an important role in the semantics of a phrase, they are distinctly classified in natural language processing libraries [26, 35].

Unlike numerous nouns, verbs and adjectives, there are only about 50 determiners in the English language [22]. Nevertheless, determiners can be highly complex, and a hardcoded or fixed-rule approach to using or understanding determiners simply will not work. For instance, take the determiner "some" – in its simplest form, "some" refers to a relatively small number or quantity. However, this can be highly noun-specific and context-specific, *e.g*. the absolute physical quantities for "add some salt" versus "drink some water" are very different. Furthermore, determiners that describe ownership or possession, such as "my" and "your", are highly context-dependent and dynamic, as possession can change on the fly, *e.g*. after handing over an object. In general, there are many such subtleties and complexities for determiners. Hence, a learning-based approach is needed, along with suitable training data.

If state-of-the-art models could learn a schema of determiners [33, 20, 34], it could facilitate flexible combination in novel contexts [21, 17, 28] and improve visual reasoning. However, existing vision-language models such as CLIP [31] and BLIP-2 [23] do not capture the semantic organization of determiners well (see Supplementary Material), and there is no visual grounding dataset that focuses on Determiners. Existing grounded referring expression datasets [4, 13, 15, 18, 27, 36, 38] exclusively focus on "the" and "a", making an unambiguous reference to a specific single object. Some examples include "bottle with a lid", "the

*Equal contribution

[1] https://github.com/clarence-lee-sheng/DetermiNet contains the dataset and code

blue truck in the bottom right corner" and "a bird that is close to the baby in a pink shirt". In other words, existing datasets focus on the noun, verb and adjective aspects of referring expressions, with "the" and "a" as the main determiners used.

Hence, as a first step towards bridging this gap, we developed the *DetermiNet* diagnostic dataset [15] to benchmark current state-of-the-art (SOTA) algorithms on their potential for learning determiner concepts. As with CLEVR [15], good performance on DetermiNet is not an end-goal in itself, as knowledge of the dataset generation process can be used to hand-craft toy models that will not generalize to real-world determiner usage. The dataset uses a bounding box localization task, set in a highly-constrained instruction task context, and deals only with simplified determiner definitions. Even with all these simplifications, we find that SOTA methods do not perform well.

DetermiNet contains 250,000 synthetic images and captions covering 25 determiners. The images are designed with the premise of two avatars interacting at a table with objects. The captions consist of a determiner followed by a noun; the task context is that the viewer is asking the avatar in the image to "pass me {*determiner noun*}".

The task is to choose a set of objects that is consistent with the given {*determiner noun*} pair. Examples are "those apples" or "either orange". Beyond just object detection, the task tests the ability to understand the logical semantics that define various determiners (see Fig. 1), such as selecting the correct number of requested objects. Simply returning all or random instances of the queried noun would not lead to high performance. Since the focus of DetermiNet is on the logical schema of determiners, high levels of visual realism and diversity are not crucial for benchmarking the ability of algorithms to learn determiners.

Finally, we analyze the performance of SOTA models that were pre-trained to perform visual grounding, so as to see if SOTA deep learning models can learn to understand the logical schema governing determiners.

In summary, our contributions are as follows:

1. We developed DetermiNet, the first large-scale diagnostic dataset covering the *determiners* word class, with 250,000 examples across 25 determiners from all four main types of determiners (Articles, Possessives, Demonstratives and Quantifiers).

2. We show that the core task of learning determiners is very challenging – even an oracle model struggles to learn the determiner schema from a few hundred examples and requires a large dataset.

3. We find that state-of-the-art visually-grounded models show only moderate results on DetermiNet, hence much more work is needed to perform well on the end-to-end task.

2. Related work

2.1. Datasets

There has been substantial work in developing datasets for visual question answering and referring expressions. However, referring expression datasets which include egocentric points of view and focus on the full coverage of the determiner class for referring is limited (see Table 1). While a dataset like Flickr30k Entities [30] contains some determiners, its coverage is narrow, with only 5.33% being non-articles. Furthermore, the captions do not consistently capture the semantics of the determiner. For example, although one particular caption specifies "some people ...", all the people (*i.e.* many) are labelled instead of just a relatively small number of people. Lastly, Flickr30k Entities is used as a phrase grounding dataset rather than a referring expression dataset, hence it is excluded from Table 1.

Table 1. Comparison of datasets for referring expressions [14]. A, P, D, Q, Exo and Ego stand for Articles, Possessives, Demonstratives, Quantifiers, Exocentric and Egocentric respectively.

Datasets	**A**	**P**	**D**	**Q**	**View**	**Images**	**Type**
RefCOCO [18]	Y	N	N	N	Exo	19,994	Real
RefCOCO+ [18]	Y	N	N	N	Exo	19,992	Real
RefCOCOg [27]	Y	N	N	N	Exo	26,711	Real
CLEVR-Ref+ [25]	Y	N	N	N	Exo	99,992	Synth
YouRefIt [8]	Y	N	N	N	Exo	497,348	Real
DetermiNet	Y	**Y**	**Y**	**Y**	**Ego**	250,000	Synth

2.2. Tasks

A greater confluence between computer vision and natural language processing research has given rise to increasingly complex mixed-modality tasks such as Visual Question Answering (VQA) [4, 15, 36] and Referring Expression Comprehension (REC) [18, 27, 38].

Both the datasets for VQA and REC are similar in that the input comprises of images or videos, and a language query is given as a caption. For VQA tasks, the model has to respond to the query by classifying the correct answer out of several potential choices. REC tasks are considered to be a harder problem as the model has to respond by predicting the bounding box coordinates or segmentation masks that identify the object of interest. Nevertheless, both tasks require a combined understanding of language attributes such as colour, shape and size, and visual attributes such as of object classes and location.

The DetermiNet dataset is related to the REC task, where the model needs to identify the object of interest either using bounding boxes or segmentation masks. However, our task defers from existing REC tasks in two ways.

Firstly, DetermiNet's captions involve only two components, a determiner followed by a noun, instead of descriptive adjectives such as colours or shapes [15], or other nouns such as people or objects [18]. This forces models to learn

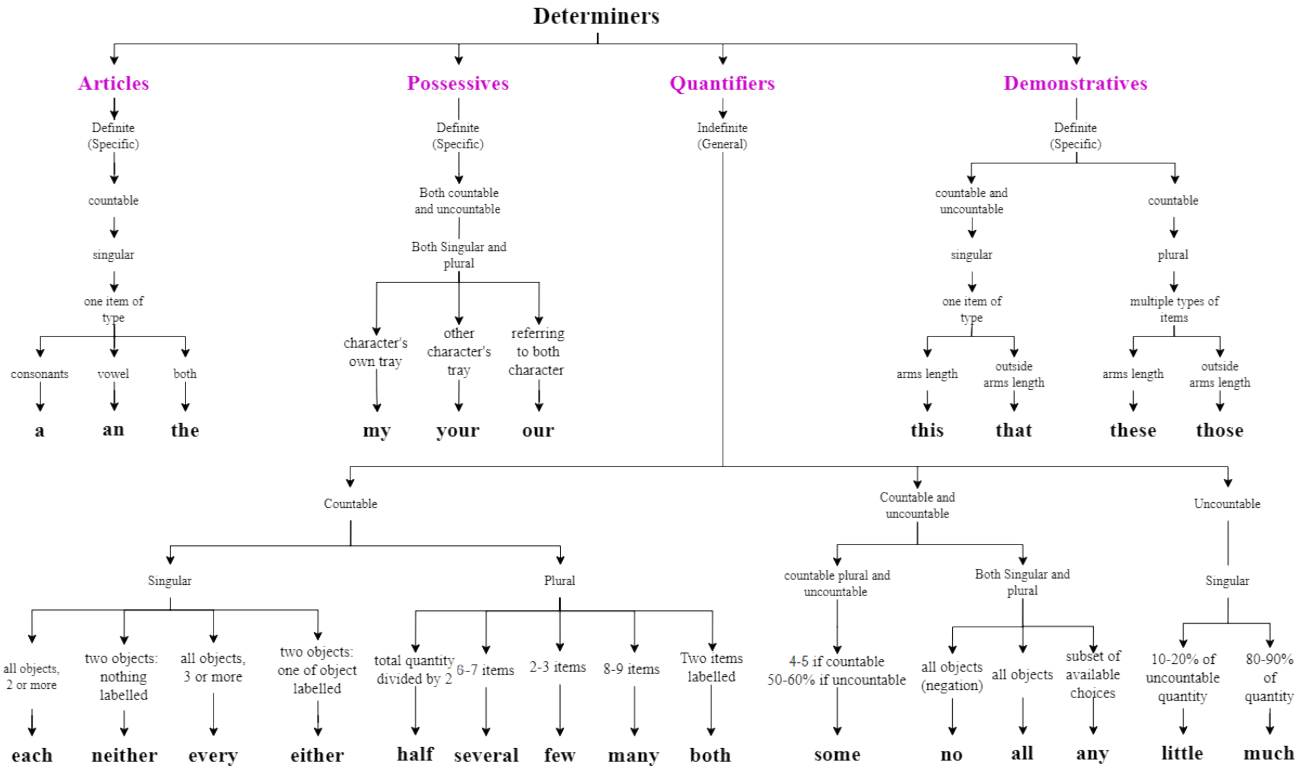


Figure 1. Organization and characteristics of the 25 determiners in DetermiNet.

and reason using a new word class, instead of using visual features and spatial representations pre-learned from other visual datasets.

Secondly, REC tasks usually tests the identification of a single object. However, DetermiNet requires models to predict multiple objects based on the query given, instead of identifying only a single instance. For example, if an image has three apples and two carrots and the query is "all apples", the model needs to predict all three bounding boxes instead of a single one. This is the biggest difference between DetermiNet and other REC tasks. DetermiNet allows the development of models to identify multiple objects that correspond to the determiner schema.

Since DetermiNet allows multiple solutions to be proposed, there can also be multiple combinations of possible solutions. For example, given the same image with three apples and two carrots, and the query "any apples", the total number of correct solutions quickly increases to $C(3,1)+C(3,2)+C(3,3) = 7$. The task evaluation metric should not penalise possible solutions and should accommodate the model prediction accordingly. To our knowledge, there are no REC or VQA tasks that support multiple combinations of solutions.

2.3. Models

Existing visually grounded models can combine language and visual modalities to achieve superior performance on many downstream tasks such as those in Grounded Language and Visual Question and Answering.

Dual Encoder models such as MDETR [16] and GLIP [42] use an image and text encoder model to encode the inputs before implementing a deep fusion or transformer layer to train the model on the image caption pairs. Seq2Seq models such as OFA [39] follow the likes of GPT [32] by processing multimodal inputs using byte sequence representation. A unified vocabulary approach to vision and language tasks is taken to perform the grounding tasks. SOTA models such as MDETR and OFA perform really well on visual grounding tasks by achieving 87.5% AP and 92.0% AP respectively on the RefCOCO dataset.

However, these models have been largely evaluated against referring expression datasets that are dependent on the spatial and visual attributes of objects. Hence, a more challenging dataset is needed to determine if these SOTA models are robust to solve language and egocentric-based object referencing, like in natural language.

3. The DetermiNet dataset

DetermiNet is the first visuo-linguistic dataset based on the determiners word class. Fig. 1 describes our determiner schema that describes which object and how many of those objects should be selected. The dataset was generated synthetically using this schema and focuses on the referencing and quantification of noun phrases. Determiners are largely used from an egocentric perspective, and their properties requires models to perform deeper and more complex reasoning to accomplish the visual grounding task. Careful curation of the dataset was conducted to account for these complexities.

To provide a comprehensive coverage, our dataset includes all four main types of determiners [2, 22], namely:
– **Articles:** identify nouns which the speaker is referring to
– **Possessives:** signify ownership of the noun
– **Demonstratives:** isolate nouns that are being referred to
– **Quantifiers:** describe the amount of the referred noun

3.1. Dataset design and construction

DetermiNet is a synthetically generated dataset based on an end-to-end-pipeline developed in Unity. Scene and phrase generations were done through predefined scene configurations based on the scene chart. Since the logic governing determiners is unrelated to the level of visual realism, DetermiNet follows the approach of synthetic data with visual simplicity [12, 15, 24, 40]. For example, CLEVR [15] and CLEVRER [40] use only 3 shapes, 2 materials and 8 colors; the background is uniform.

3.2. Dataset statistics

DetermiNet has a comprehensive coverage of 25 determiners. We generated 10,000 image-caption pairs per determiner, totaling 250,000 samples. We describe the breakdown of our train, test, and validation splits in Table 2.

Table 2. Statistics for train, test and validation splits

Splits	**Samples**	**Objects**	**Ground truth b-boxes**
Train	175000	2799790	460200
Validation	25000	399654	66023
Test	50000	799756	131460

In total, our dataset includes a variation of 15 object classes, including 5 countables starting with consonant sounds (*e.g.* "a lemon"), 5 countables starting with vowel sounds (*e.g.* "an apple") as well as 5 uncountable substances (*e.g.* "some grape juice"). Ground truths are determined by the object which the determiner is referring to. This referred object will then be labelled as part of the ground truth annotations (Fig. 2). Variations indicate the number of different permutations of the object, while the number of objects spawned indicate the possible count of that particular item spawned in the scene. A summary of the scene and object variations is shown in Table 3.

Table 3. Scene variations

Object	**Variations**	**No. spawned in scene**
Referred objects	15	1-9
All objects	15	10-20
Countables (consonant)	5	1-20
Countables (vowels)	5	1-20
Uncountables	5	1-20
Trays	2	2
Tables	2	2
Tray positions	3	-
Camera positions	3	-

3.3. Scene generation and ground truth annotation

DetermiNet is based on the interaction of two avatars at a table. We randomly spawn the positions of objects, as well as generate different perspectives. Configuration parameters were used to determine the construction of each scene, providing a unified interface for scene generation. These configuration parameters follow the tree in Fig. 1, and can be adjusted to the user's own definitions. Attributes include type of object (countability), number of referred objects (plurality), spawn locations and distance from the viewer. Egocentric viewpoints of the viewer were generated by attaching the camera to the viewer's head and directing the camera to focus towards the center of the table. We varied the avatars' positions to generate multiple perspectives.

Images were rendered using Unity3D. Camera projections were used to check for visibility of the spawned objects and collision detectors were put in place to ensure that objects did not intersect. Different objects (tray, tables) were also sampled to be used as random spawn locations.

Mesh vertices were projected onto the camera's 2D space to extract bounding boxes for all objects, modeling a perfect object detector. Unity's Image Synthesis module was used to generate object segmentation masks.

3.4. Phrase generation

DetermiNet uses the task context of "pass me $\{$*determiner noun*$\}$", *e.g.* "pass me an apple", "pass me that apple". For simplicity, we omitted "pass me". Hence, the phrases are simple captions with only a determiner and its noun phrase (Fig. 2), *e.g.* "an apple", "this apple", "some grape juice". Additionally, we follow this phrasing format while keeping errors in grammatical structure minimal. For example, "pass me all apples" is sufficient to capture the task instead of "pass me all the apples".

3.5. Evaluation metric for DetermiNet

Since the task is to evaluate bounding box predictions, we used the detection evaluation metric used by COCO,

Figure 2. Examples from DetermiNet, with image, phrase, target bounding boxes and segmentation masks shown.

specifically the average precision (AP) metric with IoU thresholds ranging from 0.5 to 0.95.

The DetermiNet dataset contains scenarios where different combinations of solutions can be correct. For instance, for an image with three apples and a query specifying "an apple", there are three equally correct solutions. However, a correct bounding box prediction should only contain one bounding box instead of three. If all three bounding boxes are predicted, the evaluation metric should evaluate the prediction as one true positive and two false positives.

To account for multiple correct solutions during evaluation, we developed a ground truth correction function that compares the model's predicted bounding boxes against all the relevant bounding boxes that satisfy both the determiner and noun conditions. The function chooses the ground truth bounding box that has the highest IoU with the predicted bounding box, and discards the rest of the relevant ground truth bounding boxes based on the quantity specified by the determiner.

The modified ground truth annotations are then used to evaluate the predictions. This way, if a model predicts three bounding boxes instead of one, the prediction with the highest IoU and prediction score will be treated as true positive, and the other two predictions treated as false positive.

4. Experiments

In this section, we verify the challenge posed by the dataset to refer or quantify objects of interest using five models. Since the DetermiNet task is similar to the REC task, models need to predict bounding boxes which were evaluated using the Average Precision (AP) evaluation metric. Before evaluation, the ground truth bounding box annotations were modified to account for multiple combinations of correct solutions.

4.1. Random selection model

The first model is a random bounding box selection model (Fig. 3). This model has two components. The first is a perfect object detector (see 3.3) that tags all objects with class labels and their corresponding bounding boxes.

The second component sampled prediction scores between 0 to 1 from a uniform distribution and generated positive and negative masks based on a threshold of 0.5 which was used to select bounding boxes as predictions.

In short, the perfect object detector generated a list of bounding boxes and the attention mask randomly selected a subset of bounding boxes as predictions without using information of either determiner or noun.

4.2. Neuro-Symbolic oracle model

The neuro-symbolic model (Fig. 3) was developed to isolate the main challenge of the dataset, which is to classify objects of interest based on the concept specified by the determiner. Hence, this model tackles the DetermiNet dataset as a classification problem, similar to VQA models.

Like the random selection model, a perfect object detector was used to identify all the object bounding boxes, class labels and volume of liquid within the object. These three information were fed to a single feedforward layer with 128 units to embed the visual information.

A perfect text encoder converted the two-part caption specifying the determiner and the noun into two one-hot encoded vectors. The first one-hot vector of length 25 represented the determiner, and the second one-hot vector of length 16 represented the noun. The two vectors were concatenated and fed to another feedforward layer with 128 units to embed textual information.

The output of the two embedding layers were concatenated and fed to two feedforward layers, each with 256 units, followed by a final classification layer with sigmoid activation function.

A ground truth attention mask was generated by comparing all the objects detected in the image against the ground truth bounding boxes such that masking the list of object bounding boxes detected by the perfect object detector will provide the ground truth bounding boxes. The model was trained to predict the ground truth attention mask using binary cross entropy for 30 epochs.

The model's prediction scores from the classification layer and bounding boxes extracted by the perfect object detector were used for evaluation. The neuro-symbolic model can be considered to be an **oracle model**, as it received ground-truth information about all the objects in the image, and it only needs to learn to predict the correct bounding boxes given the determiner and noun.

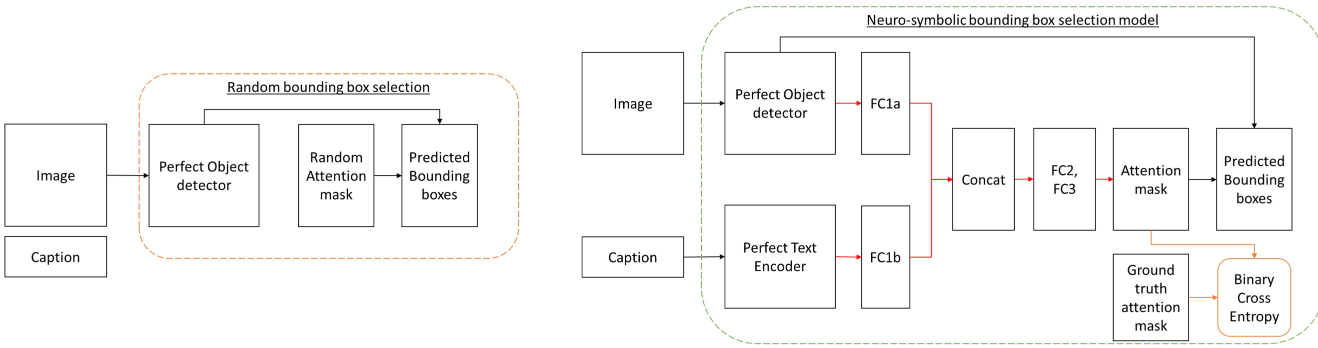


Figure 3. Random and neuro-symbolic model architectures. Weights of fully connected (FC) layers were optimized by backpropagation.

4.3. SOTA deep learning models

To verify the full challenge posed by DetermiNet, we fine-tuned three SOTA visual grounding models, OFA[39] with ResNet-152 backbone, GLIP [42] and MDETR [16] with ResNet-101 backbone for 5 epochs on our dataset.

OFA's weights were pretrained on RefCOCO and VG datasets, GLIP's weights were pretrained on O365, GoldC, CC3M and SBU datasets while MDETR's weights were pretrained on the RefCOCO, VG and Flickr datasets. Both image and captions were passed as inputs to the SOTA models, and the bounding box predictions were obtained as outputs. The object class prediction was not relevant to our DetermiNet task, so we set category ID to 1 for all predictions. While GLIP and MDETR models returned multiple bounding box predictions and scores, OFA is designed to predict only one bounding box per image.

5. Benchmarking models on DetermiNet

After correcting the ground truth annotations to account for multiple solutions, the random bounding box selection model demonstrates the worst performance of 9.8% AP. Even though the random model has the perfect object detection module, randomly selecting different quantities of different objects without considering the textual information leads to poor performance. This can be treated as the lower-bound performance for the DetermiNet dataset.

In contrast, the oracle demonstrates the highest performance of 93.5% AP (Table 4) as it receives object class and textual information while only needing to learn the determiner schema. Since the oracle model is only tested on semantics to provide a **rough upper-bound** for DetermiNet, its performance **should not be directly compared** against end-to-end models which learn both object detection and determiner semantics, and whose learning performance is difficult to disentangle. When the oracle uses MDETR object detection outputs instead of perfect detection, overall AP fell to 62.8%.

Table 4. Model performance after correcting ground truth annotations. *OFA only predicts one bounding box.

Models	AP@IoU=0.5:0.95
Random	9.8
Oracle	93.5
OFA [39]	20.6*
GLIP [42]	55.0
MDETR [16]	70.6

When comparing end-to-end finetuned models, OFA performs the worst, as it is only able to predict one bounding box, similar to the REC task condition, contributing to high false negatives. GLIP achieves 55.0% while MDETR achieves the best performance of 70.6% AP (Table 4). Although MDETR's bounding box predictions are impressive to identify the reference objects, the model does not constrain its predictions according to the determiners schema, incurring high false positive predictions. Conversely, MDETR performs well on uncountable quantifiers and possessives (Fig. 4). This is likely because MDETR gets the raw RGB image as input, allowing it to understand and reason about volume levels within a cup or the presence of the referred object on the tray.

Figure 4. MDETR suffers from high false positives as bounding box predictions (blue) are not constrained (left) but shows learning of uncountable quantifiers (middle) and possessives (right).

Refer to the **Supplementary Material** for the confusion matrix broken down for each determiner and class, as well as the performance when constraining each model's predictions to just the top-1 bounding box (similar to OFA's constraint).

5.1. Embedding of determiners

To study how the dataset is represented in both an untrained and trained network, we extracted the neural activity of the layer before the attention mask classifier. The neural activity was clustered using Linear Discriminent Analysis, with the determiner labels as targets. Before training, the neural representations corresponding to 25 determiners were highly overlapped and the centroid coordinates for each determiner class occupied the same space (Fig. 5, left).

As training progressed, the embedding of the 25 determiners evolved into clusters (Fig. 5, middle). The dendrogram (Fig. 5, right) represents the euclidean distance between centroids after training. With training, the network learns a representation that seemingly corresponds to the organization of determiners in Figure 1.

Neural representations for "a" and "an" occupy the same subspace as they obey the same articles determiner schema. We can see similar clustering of determiner subclasses such as "both" and "neither" which fall under quantifiers and "this" and "that" which fall under demonstratives. However, some determiners such as "the" and "our" do not occupy the same subspaces as articles or possessives, suggesting that the model struggles to disentangle them. Surprisingly, unlike the oracle model, text encoders in established vision-language models such as CLIP [31] and BLIP-2 [23] do not demonstrate distinct organization of determiners **(see Supplementary Material).**

5.2. Ablation study

To determine the importance of determiners and nouns in the DetermiNet task, we conducted ablation studies using oracle and MDETR models where the determiner, noun or both determiner and noun were masked during evaluation.

Table 5. Ablation study with masked captions. Performance reported AP@IoU=0.5:0.95

Ablation condition	Oracle	MDETR
Noun+ / Det+	93.5	70.6
Noun+ / Det–	71.3	56.3
Noun– / Det+	11.3	11.3
Noun– / Det–	9.8	0.2

Masking determiners while feeding in the noun is similar to a query-based object detection task. The decrease in performance for the oracle model was 22.2% while MDETR suffered a decrease of 14.3%, suggesting that MDETR learnt to predict bounding boxes using most of the determiner concepts, though not as well as the oracle model.

When the determiner was given but the noun was masked, AP dropped significantly since the object to be identified was not known. Finally, when both determiner and noun were omitted, the oracle performed similarly to the lower bound random model while MDETR performed much worse since it also had to perform object detection.

Nevertheless, SOTA models do learn some determiner concepts, and lower performance can be attributed to errors in both object detection and bounding box classification.

5.3. Dataset efficiency

Since 10,000 examples per determiner in the full dataset is presumably way beyond what humans require to learn determiners well, we trained the oracle and MDETR models on randomly sampled subsets (N=6) of DetermiNet training samples to determine how much data is needed for the models to learn the determiner schema.

Table 6. AP@IoU=0.5:0.95 with standard deviation attained after training models on 10, 100, 1000 samples per determiner.

Samples	10	100	1000
Oracle	17.9±0.6	29.6±0.4	44.7±3.3
MDETR	2.8±1.0	33.5±1.2	55.0±0.8

Since the oracle has a perfect object detector and text encoder, the increase in oracle performance is attributed solely to the learning of determiner schema. Despite the isolation of training, the oracle model struggles to learn the concept of determiners even with 1,000 examples per determiner. This could be because the oracle model has 188,308 trainable parameters and a large dataset is needed to optimize the weights accordingly. Conversely, MDETR has 185 million parameters but was pre-trained to perform object detection. After fine-tuning MDETR with 1,000 examples per determiner, its performance matches the ablation condition where the model can achieve 56.3% without needing to learn determiners (Table 5), suggesting that the faster improvement is likely due to improved object detection in DetermiNet, rather than learning about determiners. Nevertheless, DetermiNet follows a scaling law that is consistent with other visual recognition tasks.

6. Transfer of learning to real images

We curated a dataset with 100 real world images and captions using images from COCO [7]. The oracle model achieved decent zero-shot performance on the real-image samples (78.1%), demonstrating a neural network's ability to generalize to real images if object detection works well.

Although MDETR was pre-trained on RefCOCO, it struggled to refer and quantify individual objects according to the determiner schema (10.4%) since RefCOCO did not account for such determiner concepts (Table 1) and instead predicted single bounding boxes for a collection of objects (Fig. 6). Fine-tuning MDETR on the synthetic DetermiNet significantly increased performance to 19.5% as the model learned to identify and quantify each object (Fig. 6, top row), suggesting that the determiner concepts learned

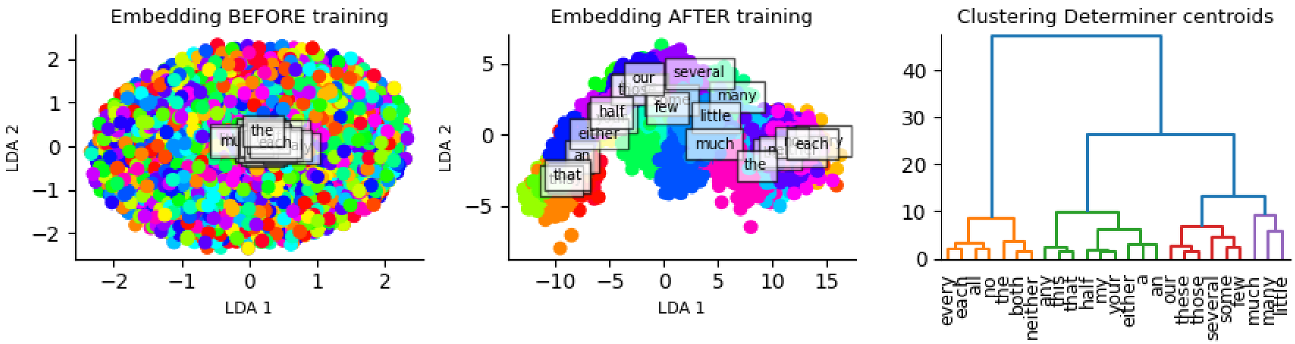


Figure 5. Clustering 25 determiners represented in the last feature layer of the oracle model using LDA.

from the synthetic dataset transferred to real images to a certain extent. However, MDETR still struggles with some determiner concepts such as "half" (Fig. 6, bottom row). The far lower MDETR performance could be due to poor object detection, separate from learning the semantics of determiners. The real-image test samples will be made available along with the synthetic DetermiNet.

Table 7. Zero-shot evaluation on real-image dataset

Models (Tasks pretrained on)	**AP@IoU=0.5:0.95**
Oracle	78.1
MDETR (Pretrained)	10.4
MDETR (Finetuned on DetermiNet)	19.5

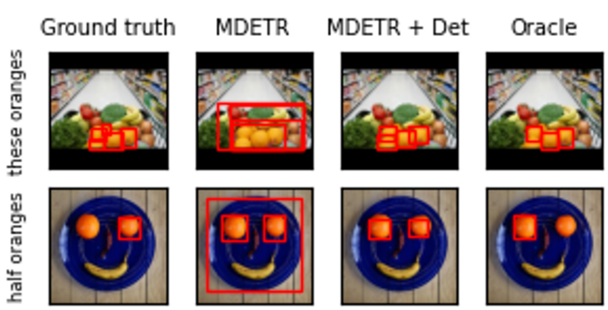


Figure 6. Ground truth, pretrained MDETR, MDETR fine-tuned on DetermiNet and Oracle model predictions on 100 real images.

7. Current limitations

Since the dataset focuses on referencing and quantification, we omitted the use of wh-determiners (*e.g.* "where", "what"), which are mainly used in question answering tasks. Since we constrained our captions to fit the task context of "pass me {determiner, noun}", comparison determiners such as "more" and "less" were left out for now, as they require multiple sets of nouns. Furthermore, gender-specific possessives such as "his" and "her" were omitted, as gender recognition is not the focus of this work. Additionally, the composition of multiple determiners (*e.g.* "pass me some of those apples) will be explored in future work.

Parameter-efficient finetuning, or adding the semantic module of the oracle to a trained detector such as MDETR can serve as an additional evaluation to disentangle the learning performances of object detection and determiner semantics in end-to-end models.

8. Conclusion

We present the DetermiNet dataset to determine if models can learn object referencing and quantification for all four major determiner categories. The dataset accommodates multiple combinations of possible solutions, as in a natural language context. Since the dataset images and ground truth annotations were synthetically generated, it allows for rapid reconfiguration of parameters, scenes and object classes to increase the challenge posed by the dataset.

Our experiments demonstrate that although state-of-the-art visual grounding models are able to identify objects of interest, they do not perform well on the overall task. While they can learn the semantics of some determiners and transfer the concept to real images, they require exponential amounts of data to learn the determiner schema and struggle to handle ambiguity when considering multiple combinations of possible solutions.

In summary, DetermiNet highlights determiners as important and complex but neglected, and formulates a common task framework for all 4 determiners types. It shows the current limitations of visual grounding models in learning determiner schemas in referencing and quantification. Good oracle results on real images suggests the "determiner logic module" could be used for captioning, VQA, etc.

9. Acknowledgements

This was supported by an A*STAR CRF award (C.T.), and by the Design and Artificial Intelligence department, Singapore University of Technology and Design (C.L.).

References

[1] The OEC: Facts about the language. `https://web.archive.org/web/20111226085859/http://oxforddictionaries.com/words/the-oec-facts-about-the-language`.

[2] What is a determiner? `https://www.oxfordinternationalenglish.com/what-is-a-determiner/`, Jun 2021.

[3] Ahmad Abu-Akel, Alison L Bailey, and Yeow-Meng Thum. Describing the acquisition of determiners in english: A growth modeling approach. *Journal of Psycholinguistic Research*, 33(5):407–424, 2004.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[5] Mohit Bajaj, Lanjun Wang, and Leonid Sigal. G3raphground: Graph-based language grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4281–4290, 2019.

[6] Roger Brown. A first language. In *A First Language*. Harvard University Press, 2013.

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

[8] Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Song-Chun Zhu, Tao Gao, Yixin Zhu, and Siyuan Huang. YouRefIt: Embodied reference understanding with language and gesture. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1385–1395, 2021.

[9] Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. Using syntax to ground referring expressions in natural images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

[10] Rodolfo Corona, Shizhan Zhu, Dan Klein, and Trevor Darrell. Voxel-informed language grounding. *arXiv preprint arXiv:2205.09710*, 2022.

[11] Jialin Gao, Xin Sun, Bernard Ghanem, Xi Zhou, and Shiming Ge. Efficient video grounding with which-where reading comprehension. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10), 2022.

[12] Ankit Goyal, Kaiyu Yang, Dawei Yang, and Jia Deng. Rel3D: A minimally contrastive benchmark for grounding spatial relations in 3D. *NeurIPS*, 2020.

[13] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

[14] Md Mofijul Islam, Reza Manuel Mirzaiee, Alexi Gladstone, Haley N Green, and Tariq Iqbal. CAESAR: An embodied simulator for generating multimodal referring expression datasets. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[15] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 2901–2910, 2017.

[16] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr–modulated detection for end-to-end multi-modal understanding. *arXiv preprint arXiv:2104.12763*, 2021.

[17] Ken Kansky, Tom Silver, David A Mély, Mohamed Eldawy, Miguel Lázaro-Gredilla, Xinghua Lou, Nimrod Dorfman, Szymon Sidor, Scott Phoenix, and Dileep George. Schema networks: Zero-shot transfer with a generative causal model of intuitive physics. In *International Conference on Machine Learning*, pages 1809–1818. PMLR, 2017.

[18] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, 2014.

[19] Ranjay Krishna, Ines Chami, Michael Bernstein, and Li Fei-Fei. Referring relationships. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6867–6876, 2018.

[20] M Ganesh Kumar. Biologically plausible computations underlying one-shot learning of paired associations. *ScholarBank@NUS Repository*, 2022.

[21] M Ganesh Kumar, Cheston Tan, Camilo Libedinsky, Shih-Cheng Yen, and Andrew Yong-Yi Tan. One-shot learning of paired association navigation with biologically plausible schemas. *arXiv preprint arXiv:2205.09710*, 2023.

[22] Geoffrey Leech, Paul Rayson, et al. *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge, 2014.

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[24] Nan Liu, Shuang Li, Yilun Du, Josh Tenenbaum, and Antonio Torralba. Learning to compose visual relations. *NeurIPS 2021*.

[25] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. CLEVR-Ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4185–4194, 2019.

[26] Edward Loper and Steven Bird. NLTK: The natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[27] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2016.

[28] James L McClelland. Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *Journal of Experimental Psychology: General*, 142(4):1190, 2013.

[29] Kolby Nottingham, Litian Liang, Daeyun Shin, Charless C Fowlkes, Roy Fox, and Sameer Singh. Modular framework for visuomotor language grounding. *arXiv preprint arXiv:2109.02161*, 2021.

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

[31] 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. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[32] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. Technical report, 2018.

[33] David E. Rumelhart. The building blocks of cognition. *Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence, and education*, 1980.

[34] David E Rumelhart and Andrew Ortony. The representation of knowledge in memory. *Schooling and the acquisition of knowledge*, 99:135, 1977.

[35] Beatrice Santorini. Part-of-speech tagging guidelines for the Penn Treebank project. Technical report, 1990.

[36] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10740–10749, 2020.

[37] Alessandro Suglia, Qiaozi Gao, Jesse Thomason, Govind Thattai, and Gaurav Sukhatme. Embodied BERT: A transformer model for embodied, language-guided visual task completion. *arXiv preprint arXiv:2108.04927*, 2021.

[38] Jesse Thomason, Mohit Shridhar, Yonatan Bisk, Chris Paxton, and Luke Zettlemoyer. Language grounding with 3d objects. In *Conference on Robot Learning*, pages 1691–1701. PMLR, 2022.

[39] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.

[40] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. CLEVRER: Collision events for video representation and reasoning. *ICLR*, 2020.

[41] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

[42] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. GLIPv2: Unifying localization and vision-language understanding. In *Thirty-sixth Conference on Neural Information Processing Systems*, 2022.

Supplementary Material

1. Introduction

In this supplementary material, we elaborate on the DetermiNet diagnostic dataset by detailing the ground truth correction function, providing further analysis of our Oracle, MDETR [1], GLIP[4] and OFA [3] model predictions, as well as presenting more examples for each determiner.

2. Correcting Ground Truth

Certain determiners (such as "an") afford multiple correct solutions. For example, in an image with three apples (A, B, C) and a caption specifying the query "an apple", the prediction should contain only one bounding box that identifies any one apple (A or B or C).

The ground truth annotation used during model training comprises of only one bounding box, randomly tagged to one of the three apples (*e.g*. A). During the evaluation phase, the model might predict one bounding box to identify a different apple (*e.g*. B) that might not correspond to the ground truth bounding box (*i.e*. A). Since the model correctly identified the object and quantity, this must translate to 100% AP if there is perfect object detection.

To correct for multiple possible solutions, all the possible correct ground truth annotations (A, B, C) were compared against the model prediction (B). The apple with the highest IoU while exceeding the IoU threshold of 0.5 (*i.e*. B) will be chosen to be the new ground truth instead of the original ground truth (*i.e*. A). If the maximum IoU did not cross the threshold, the ground truth annotation will not be modified.

This complexity of evaluating multiple correct solutions extends to the following determiners, whose ground truth annotations were modified according to the concept defined for each determiner – "a", "an", "the", "either", "any", "this", "that", "some", "many", "few", "several" and "half".

3. Model prediction analysis

In this section, we analyse each model's predictions. Table 1 shows the number of corrected ground truth annotations as well as the number of predicted annotations by each model. The number of bounding box predictions by the oracle is almost similar to the ground truth. However, both GLIP and MDETR predict more bounding boxes than the ground truth annotations while OFA predicts far fewer bounding boxes compared to the ground truth annotations.

Table 2 illustrates the overall confusion matrices. Confusion matrices were generated after filtering for predicted bounding boxes with prediction scores more than 0.5. The sum of the true positives and false negatives equal the number of ground truths, while the sum of false and true positives will equal the number of predictions.

Table 1. Number of ground truth annotations and predictions.

Model	Ground truth	Predictions
Random	134,775	400,031
Oracle	133,270	135,152
OFA[3]	127,856	50,000
GLIP[4]	135,562	997,545
MDETR [1]	138,613	178,869

Table 2. Confusion matrix averaged over IoU=0.50:0.95, where FN, FP, TP stand for False Negative, False Positive and True Positive respectively.

	FN	FP	TP
Random	62,908	328,164	71,867
Oracle	17,585	19,467	115,685
OFA [3]	93,458	15,692	34,398
GLIP [4]	52,871	72,242	82,691
MDETR [1]	30,299	70,555	108,314

We perform further analysis by breaking down the confusion matrix according to each determiner, as shown in Table 3. The oracle model performed fairly well on DetermiNet. However, it incurred higher false negatives on determiners "any" and "that".

The MDETR model suffers from high false positive for determiners such as "a", "an", "'either" and "half" which requires the model to select one or a few objects instead of all objects. This highlights the inability for MDETR to constrain its predictions to the correct number of objects referred to by the determiner. A similar reasoning can be used to explain the high false positives for "this" and "that" as MDETR does demonstrate spatial reasoning by achieving high true positives for "these" and "those".

The GLIP model demonstrates a poorer ability to learn the determiner scheme. Specifically, it does not predict according to the quantity specified by the determiner. For example, it predicts more than one bounding box for all articles "a", "an" and "the" and single demonstratives "this" and "that", but does not predict all bounding boxes for the objects specified by "all", "no", "both" and "neither". In addition, it does not learn possessives "my" and "your" though it learns to choose all objects on the tray with "our". Hence, although GLIP performs better than OFA, it struggles to learn the determiner scheme as well as MDETR.

The confusion matrices support our conclusion that current SOTA models struggle to learn DetermiNet as they do not constrain their predictions according to the determiner scheme. Models like MDETR and GLIP predict more bounding boxes than required, incurring high false positives. Conversely, single output models like OFA predict one instead of multiple bounding boxes and are thus unable to quantify multiple objects, incurring high false negatives.

Table 3. Confusion matrix per model averaged over IoU=0.50:0.95, FN, FP and TP refer to False Negative, False Positive, and True Positive respectively. Blue indicates highest number among FN, FP and TP.

Determiner	**Oracle**			**MDETR**			**GLIP**		
	FN	FP	TP	FN	FP	TP	FN	FP	TP
a	728	449	**1272**	438	**2487**	1562	1468	**3160**	532
an	627	487	**1373**	354	**2353**	1647	1571	**3175**	429
the	155	146	**1845**	447	558	**1553**	1996	**3372**	4
my	993	1109	**3023**	457	518	**3551**	1185	**3438**	2822
your	1416	1627	**2587**	1390	**3138**	2545	1951	**4141**	1984
our	1965	3042	**6066**	1833	4695	**6140**	2473	3856	**5500**
this	958	537	**1042**	220	**2252**	1780	682	**4759**	1318
that	**1039**	713	961	604	**2719**	1396	991	**5104**	1009
these	523	986	**5482**	685	721	**5389**	1024	3637	**5050**
those	693	1080	**5331**	1825	2974	**4148**	2226	**4556**	3747
any	**1162**	76	1150	1347	1459	**4521**	1515	2014	**2780**
all	194	241	**6841**	1515	3057	**5410**	**3532**	1739	3393
no	197	246	**5761**	1311	2403	**4635**	**3259**	2076	2687
every	51	108	**7973**	1756	2928	**6267**	2765	1418	**5258**
each	45	118	**6982**	1543	2282	**5527**	2908	1649	**4162**
few	120	59	**4771**	1093	1467	**3885**	**2672**	2081	1724
several	88	1237	**13334**	2570	6637	**9403**	2314	2249	**9092**
many	26	88	**16937**	3713	8009	**13237**	3875	1934	**12886**
some	344	1054	**7274**	1983	4994	**6982**	2625	2688	**6201**
both	57	61	**3943**	875	1143	**3125**	**3403**	2603	597
neither	51	66	**3949**	876	1305	**3124**	**3397**	2520	603
either	671	379	**1329**	372	**2420**	1628	1586	**2693**	414
half	1659	924	**2335**	865	**5085**	3070	1388	**4064**	2547
little	599	817	**4483**	1140	3024	**3846**	1125	1743	**3862**
much	1305	460	**3684**	1087	**2420**	1628	940	1573	**4092**

4. Breakdown by determiner class

Table 3 shows the performance breakdown of each determiner while Table 4 provides breakdown analysis of the four determiner classes. The number of determiners included in each class is indicated in the bracket with Articles, Demonstratives, Possessives and Quantifiers having 3, 4, 3, and 15 determiners respectively.

Table 4. Performance breakdown (AP@IoU=0.5:0.95) by determiner class. Number in brackets indicates number of determiners.

Models	All (25)	A (3)	D (4)	P (3)	Q (15)
Oracle	93.5	76.4	85.3	71.3	96.9
OFA	20.6	37.5	31.5	22.9	19.3
GLIP	55.0	1.9	33.9	44.3	63.8
MDETR	70.6	62.9	72.8	71.5	70.5

The oracle achieved the highest performance across most determiner classes while MDETR achieved slightly higher results for possessives. Understanding the concept of possessives required visual information to locate an object on a tray, and pure coordinates and bounding boxes may be misleading. For example, an apple can be in front, rather than on a tray which will cause the apple's bounding box to overlap with the tray bounding box. The oracle model only received bounding boxes and not visual information. This could be a reason why MDETR could reason slightly better than the oracle model about possessives.

5. Top-1 bounding box prediction comparison

Table 5. Model performance (AP@IoU=0.5:0.95). Right column indicates model predictions constrained to single bbox prediction.

Models	**AP** (multiple bbox)	**AP** (single bbox)
Random	9.8	1.6
Neuro-Symbolic	93.5	34.7
OFA	-	20.6
GLIP	55.0	14.3
MDETR	70.6	29.7

Table 5 shows the performance of all models when constrained to a single bounding box prediction. As DetermiNet requires detection of multiple objects, the AP dropped for all models. OFA performs slightly better than GLIP, achieving 20.6% as compared to 14.3%. MDETR is still the best end-to-end model, achieving 29.7%.

6. Determiner representations in current VLMs

The following dendrograms show the cosine distance of the 25 determiner embeddings extracted from the text encoders of the Oracle, CLIP, BLIP-2 models. The embeddings learned by the oracle in Figure 1 is similar to the organization of determiners and the four determiner classes are grouped closely together. Conversely, determiner organization and clustering is lacking in the text encoder embeddings of CLIP (Figure 2) and BLIP-2 (Figure 3). BLIP-2 is a current SOTA visual-language model with a GPT-3 equivalent text encoder with 6.7 billion parameters [2]. The poor separability between determiner classes demonstrate that existing VLMs insufficiently capture the semantics of determiners, motivating the need for a new large dataset that can explicitly teach determiner semantics to VLMs.

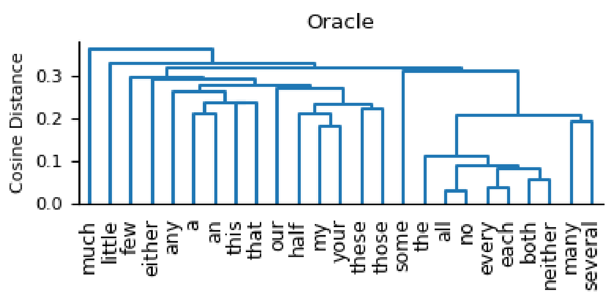


Figure 1. Dendrogram of determiner word embeddings by oracle model's FC1b layer.

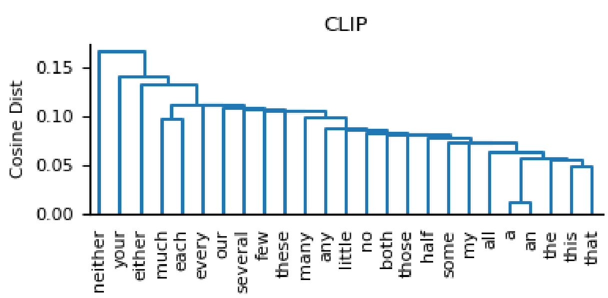


Figure 2. Dendrogram of determiner word embeddings by CLIP's text encoder.

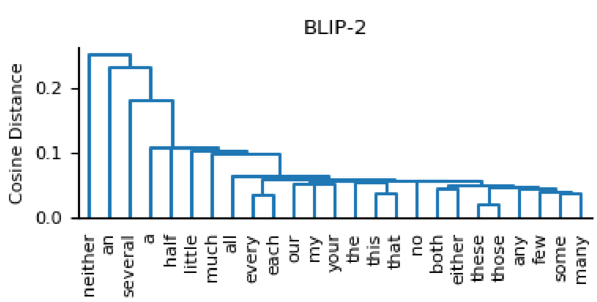


Figure 3. Dendrogram of determiner word embeddings of BLIP-2's 6.7 billion parameter text encoder.

7. Additional limitations

The caption for each sample is simply comprised of two parts, the determiner and the noun. This makes some samples ungrammatical. Examples include "all papaya juice" and "half apples". Although some of these cases can be easily fixed, we decided against it, as these fixes would be ad-hoc and only for presentation purposes, since they do not change either the logic or the learning of determiners. For example, "all" can be displayed as "all the", so that "all apples" becomes the grammatically-correct "all the apples" – but the extra "the" doesn't change the underlying logic of "all".

The possessive determiners (*e.g.* "my") are context- and noun-specific. For example, when I pass a cup to you, the possession could change from "my cup" to "your cup", but alternatively the cup could still be mine but you are borrowing it from me. It is difficult to demonstrate the various definitions and combinations of possessions using a static image. Hence, the concept of possession in DetermiNet was simplified to objects on a tray to symbolize "our", and objects on the tray closer to or further away from the camera's point of view as "my" and "your".

Everyday usage of the determiner "the" can also imply that the object was already previously mentioned, or is of common knowledge (*e.g.* "the sun"). Again, it is difficult to portray this concept using static images with no continuity between samples. Instead, we simplified the concept "the" to refer to an object that is the only one of its category in an image.

Determiners include the negative words "no" and "neither". However, the use of these within our task framing (*e.g.* "pass me no apples") is semantically incongruous. Nonetheless, we simplified these concepts and ground truth annotations to be the same as "all" and "both" respectively, and the model has to predict all or two bounding boxes for the objects of interest. Negation of an object could be conveyed using complex sentences such as "pass me all red objects but no apples", but that is beyond the current scope of this paper.

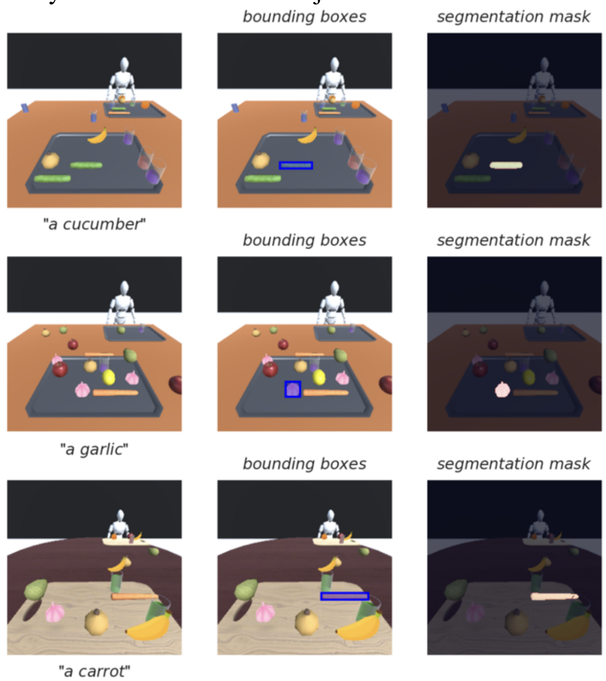
Therefore, a dynamic dataset with complex sentence structure and different contexts needs to be created for models to learn the complexities underlying possession, specific articles and negation of objects.

8. Examples for each determiner

The following section gives three examples for each determiner as well as the definitions used to generate the scenes.

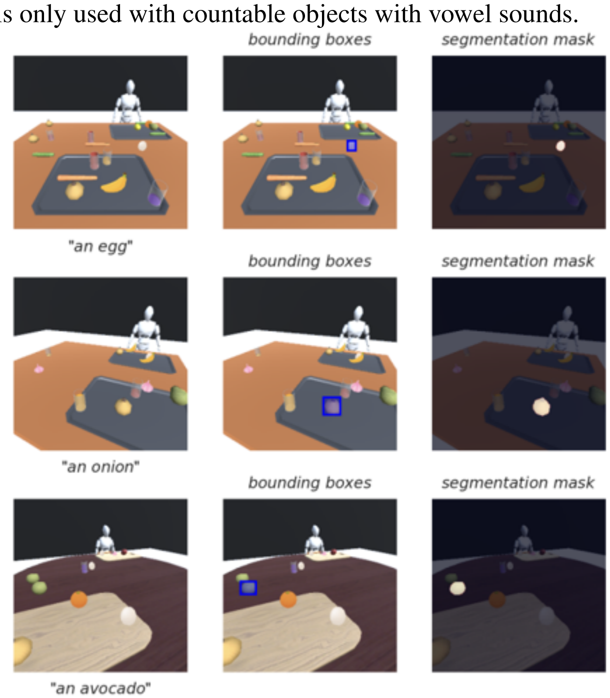
8.1. “A”

“A” selects a single object referred to in the phrase and is only used with countable objects with consonant sounds.



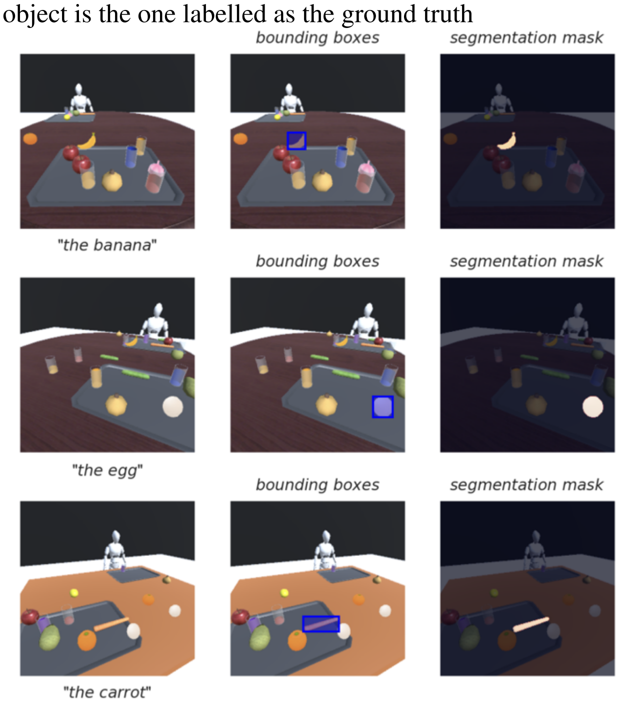
8.2. “An”

“An” selects a single object referred to in the phrase and is only used with countable objects with vowel sounds.



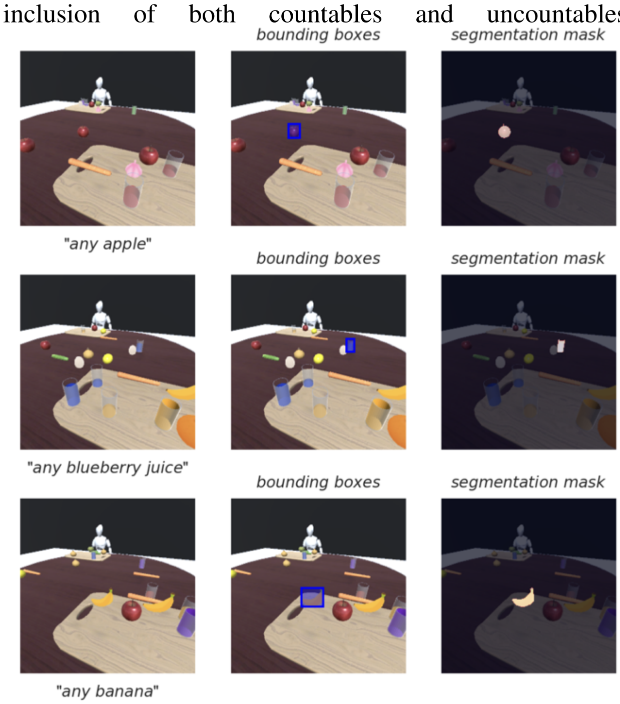
8.3. “The”

“The” is a definite article, thus only one object of the object being referred to is spawned in the scene, and that object is the one labelled as the ground truth



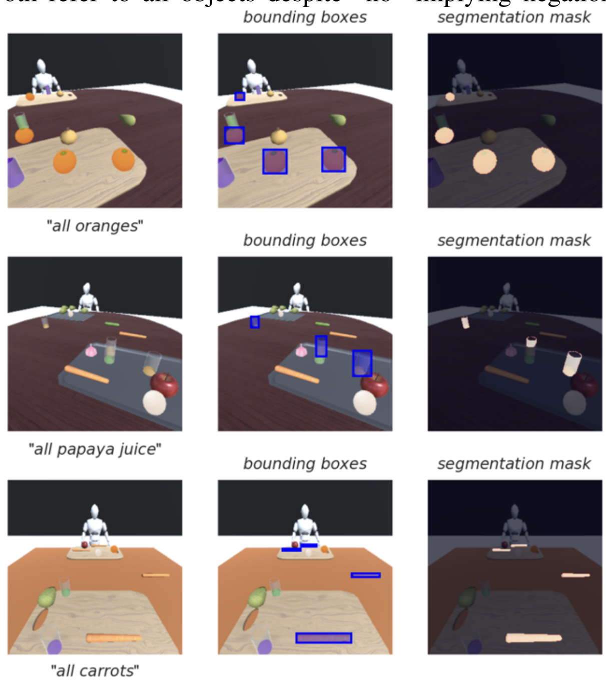
8.4. “Any”

“Any” in the singular sense such as “any apple” is similar to a/an, however, it allows the inclusion of both countables and uncountables.



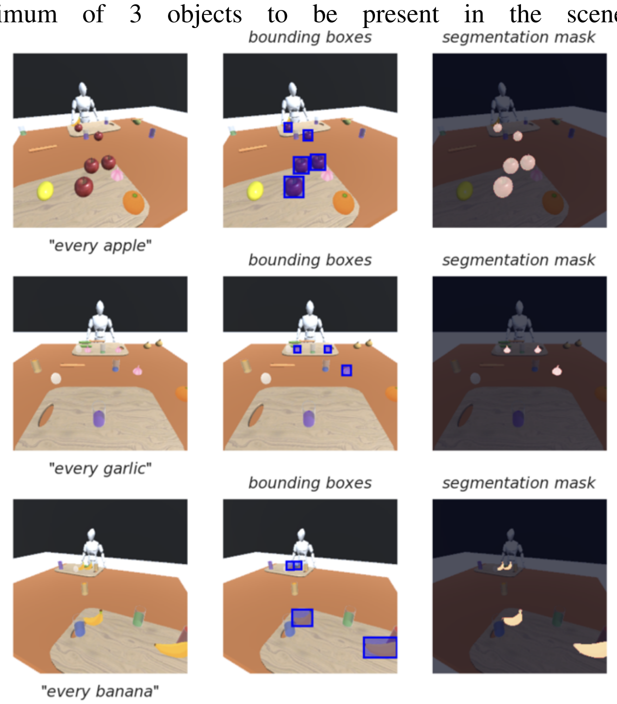
8.5. “All” / “No”

“All” and “no” are synonymous in the referencing task as “all apples are red” is equivalent to saying “no apples are not red”. Hence, in the dataset, “all” and “no” both refer to all objects despite “no” implying negation.



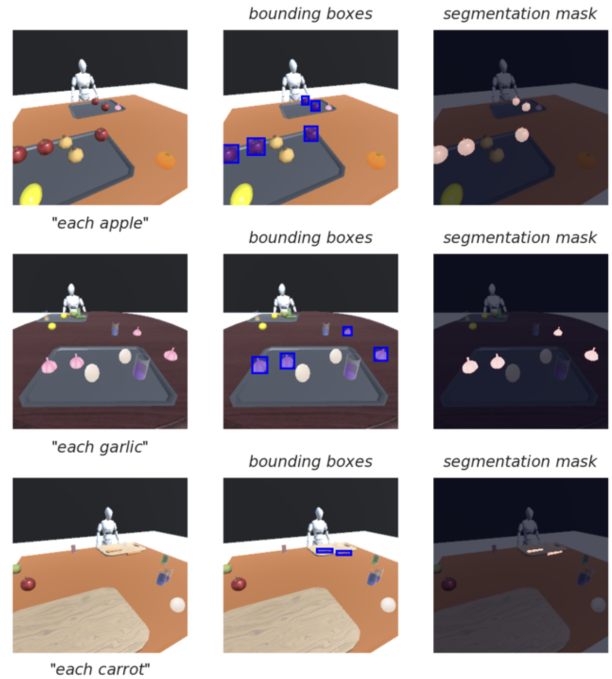
8.6. “Every”

“Every” is similar to “all” however, it only includes countable objects and also requires a minimum of 3 objects to be present in the scene



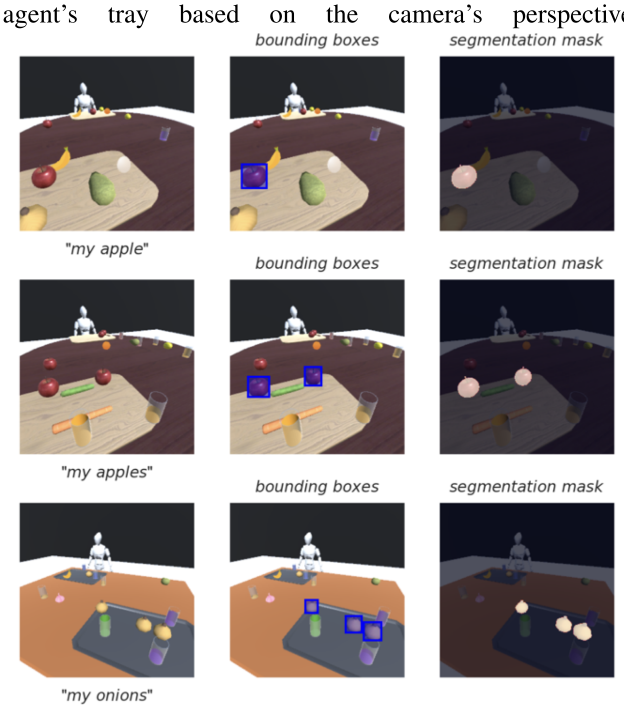
8.7. “Each”

“Each” is similar to all however, it only includes countable objects and also requires a minimum of 2 objects to be present in the scene



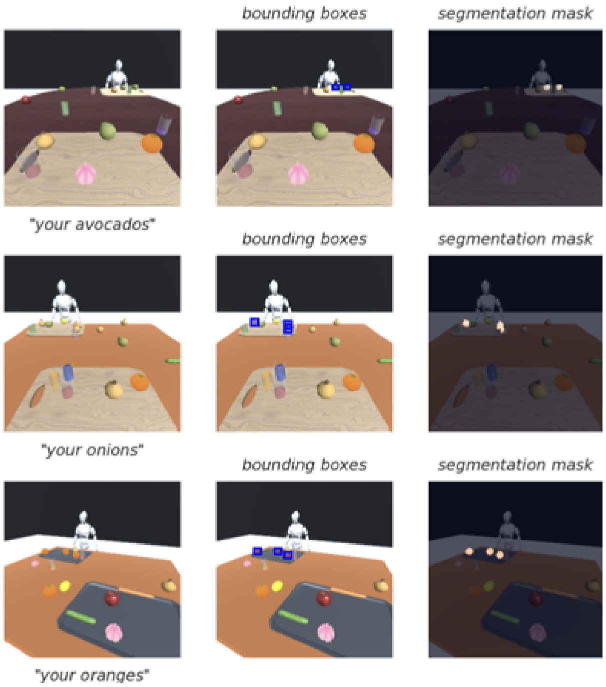
8.8. “My”

“My” selects all the objects on the main agent’s tray based on the camera’s perspective



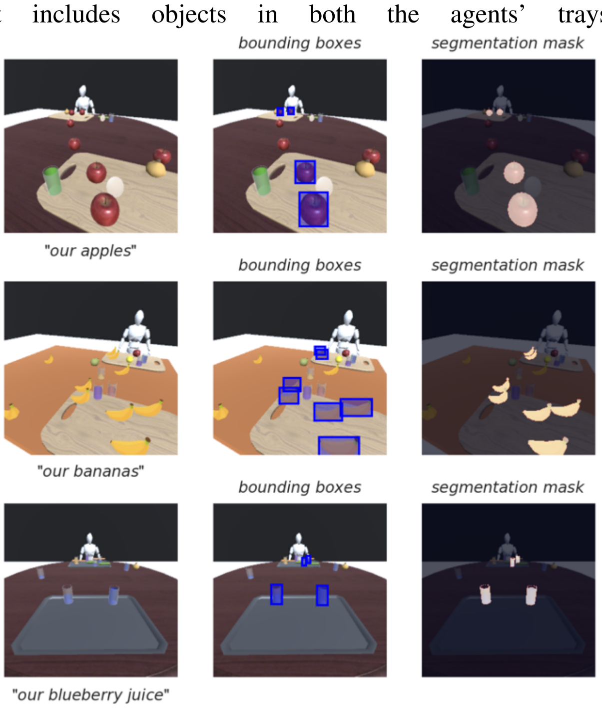
8.9. “Your”

“Your” selects all the objects on the other agent’s tray based on the camera perspective.



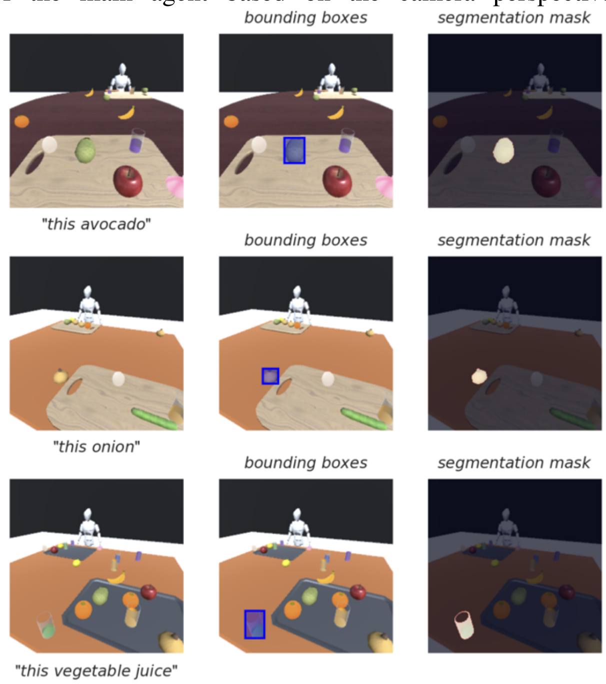
8.10. “Our”

“Our” is “your” + “my” in the scene, hence it includes objects in both the agents’ trays.



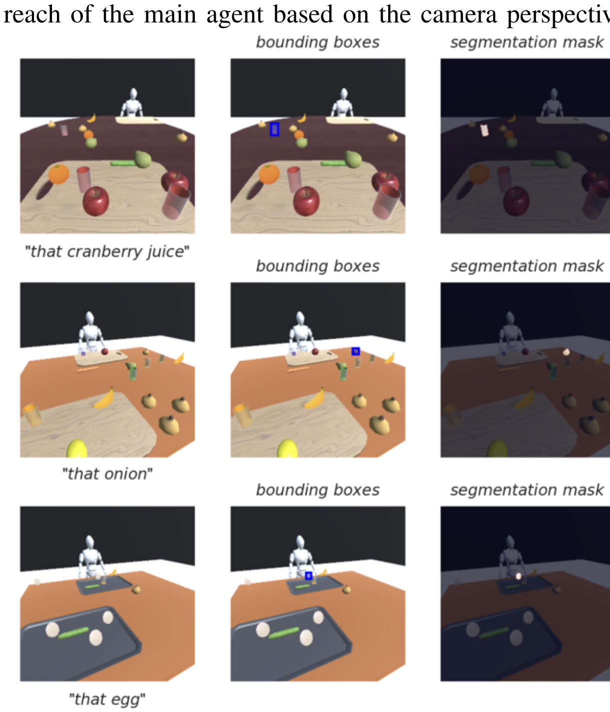
8.11. “This”

“This” refers to a single object that is within reach of the main agent based on the camera perspective.



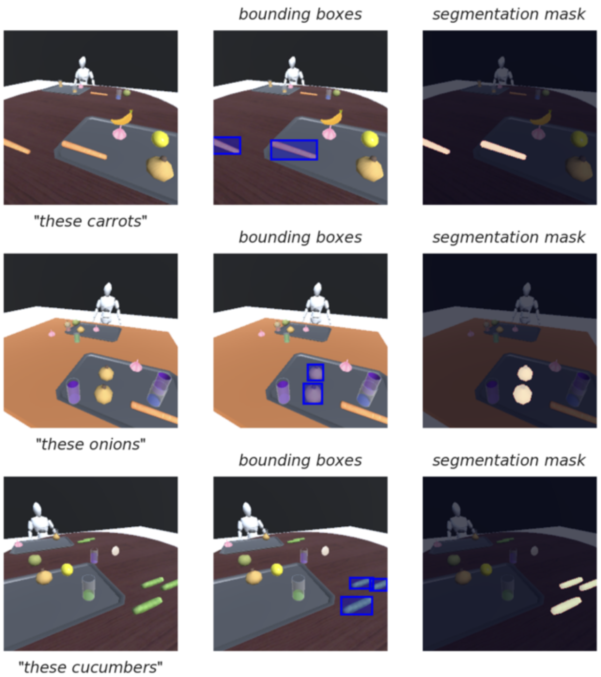
8.12. “That”

“That” refers to a single object that is outside of the reach of the main agent based on the camera perspective.



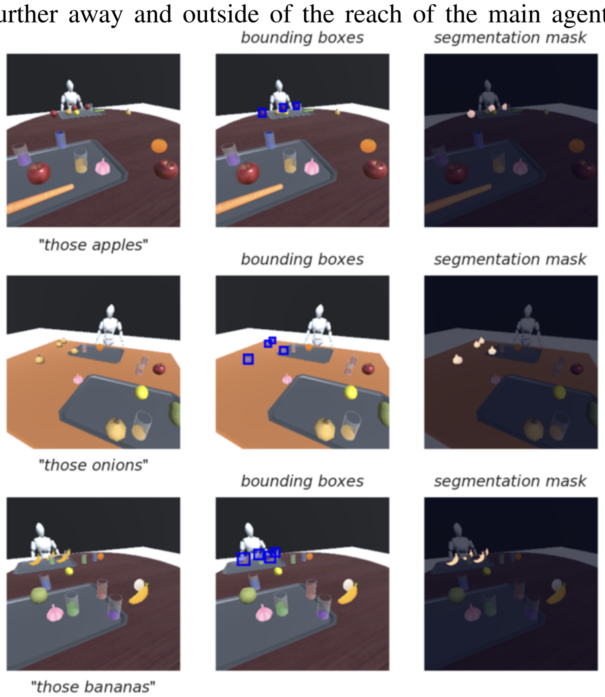
8.13. "These"

"These" refers to referencing a group of objects that are close by and within reach of the main agent.



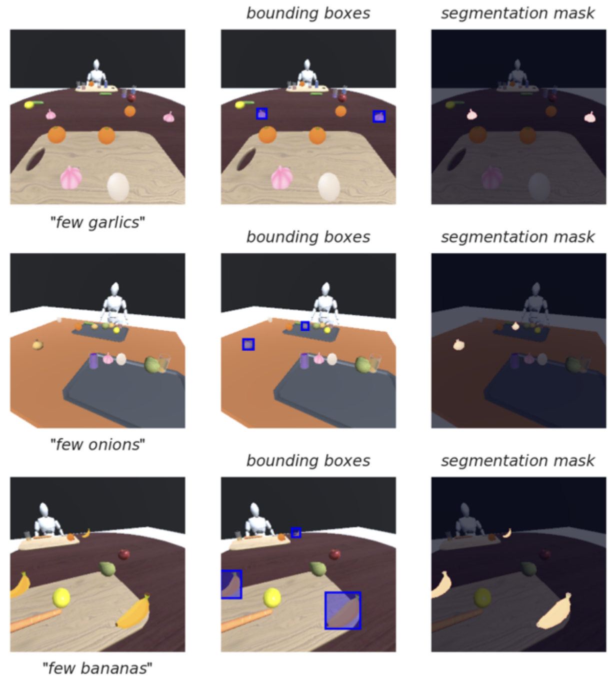
8.14. "Those"

"Those" refers to referencing a group of objects that are further away and outside of the reach of the main agent.



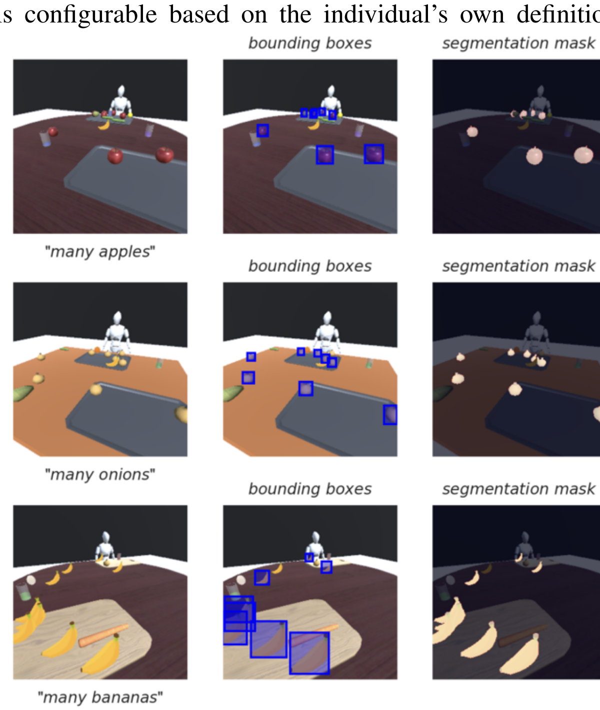
8.15. "Few"

In DetermiNet, we defined "few" as any 2-3 objects out of the all the objects mentioned in the phrase. This number is configurable based on the individual's own definition.



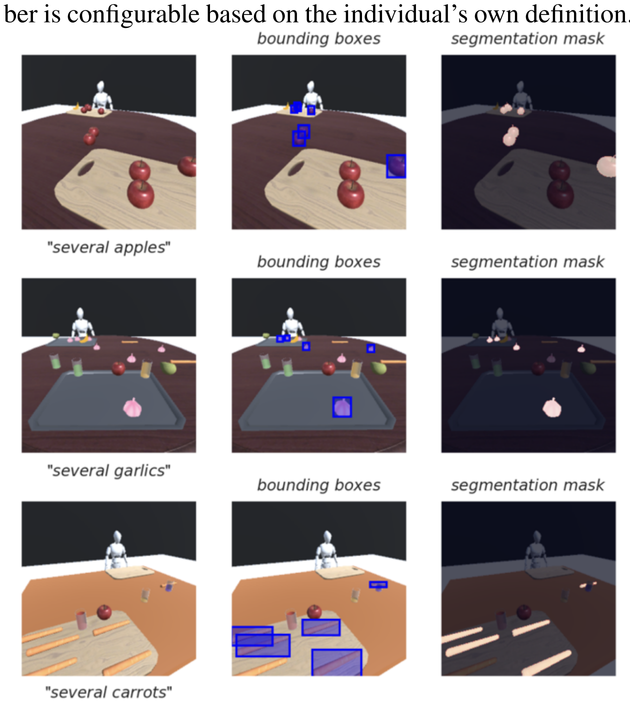
8.16. "Many"

In DetermiNet, we defined "many" as any 8-9 objects out of all mentioned in the phrase. This number is configurable based on the individual's own definition.



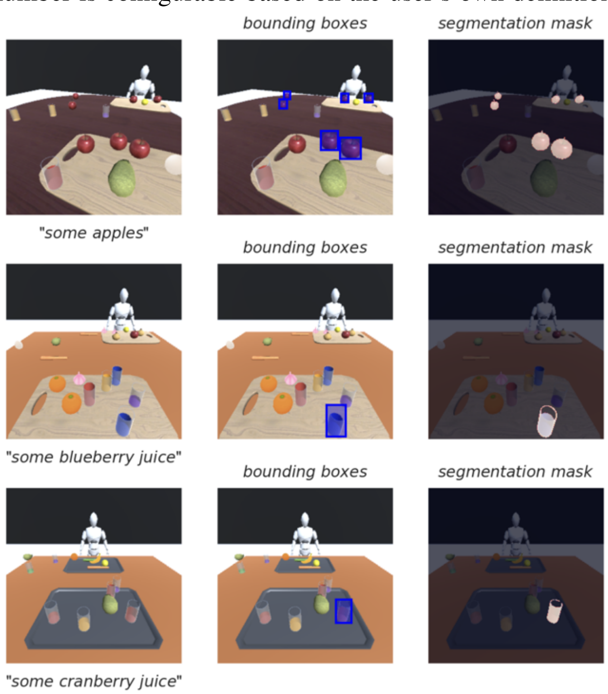
8.17. "Several"

In DetermiNet, we defined "several" as any 4-7 objects out of the all the objects mentioned in the phrase. This number is configurable based on the individual's own definition.



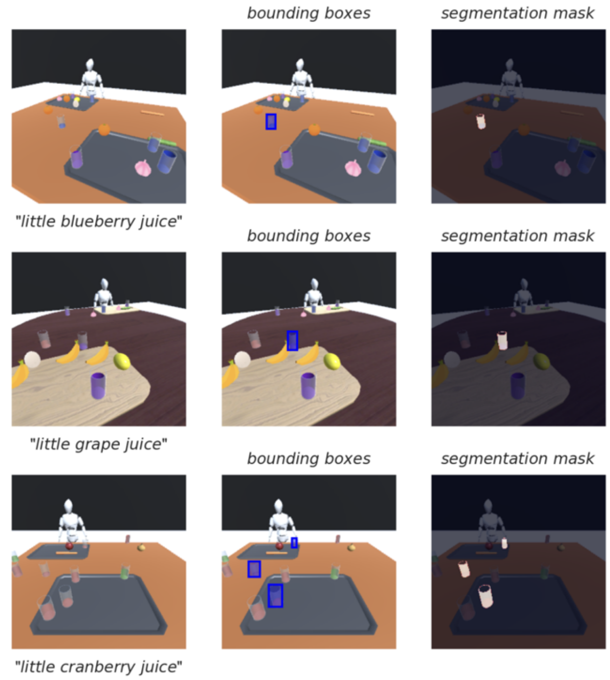
8.18. "Some"

In DetermiNet, we defined "some" as any 5-6 objects for countables and 50-60% liquids for uncountables. This number is configurable based on the user's own definition.



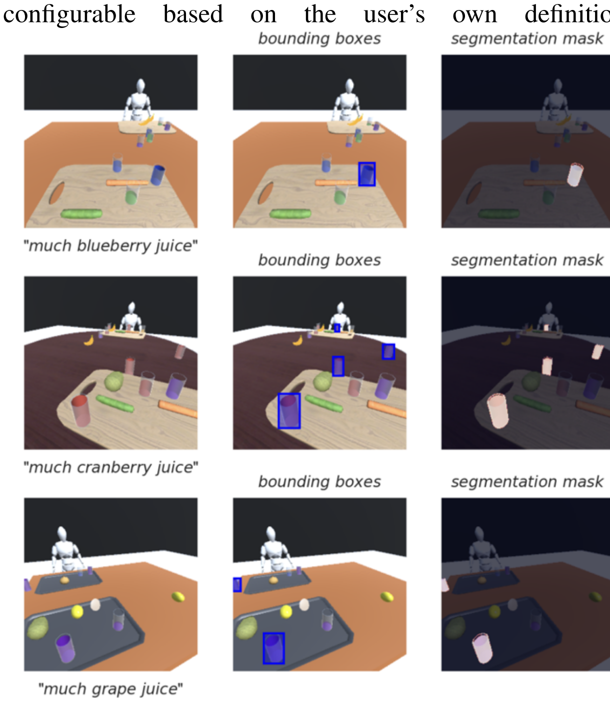
8.19. "Little"

In DetermiNet, we defined "little" as glasses being 10-20% filled for liquids, this number is configurable based on the user's own definition.



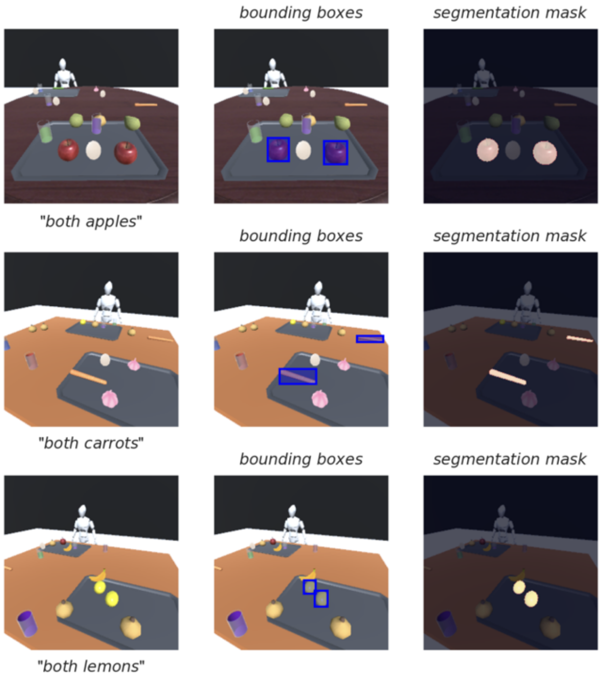
8.20. "Much"

In DetermiNet, we defined "much" as glasses being 80-90% filled for liquids, this number is configurable based on the user's own definition.



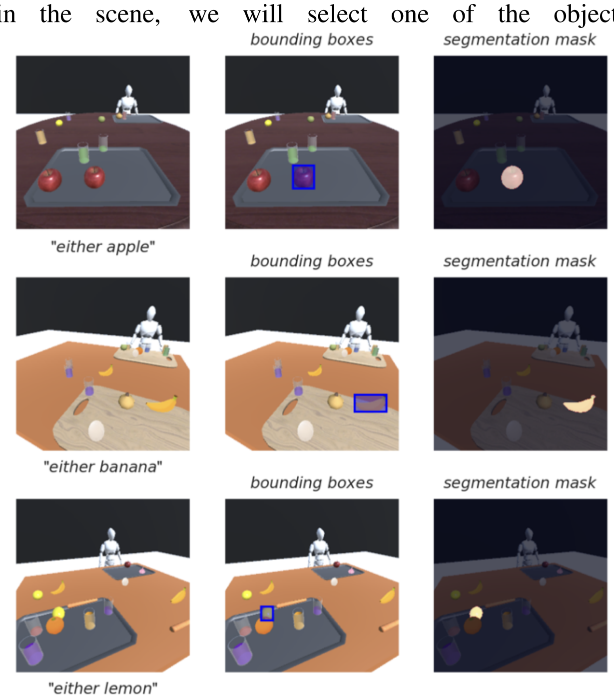
8.21. "Both" / "Neither"

"Both" and "Neither" are synonymous in the referencing task as"Both apples are red" is equivalent to saying "Neither apples are not red". Hence, "Both/Neither" indicates that out of two of the objects in the scene, we select two of them.



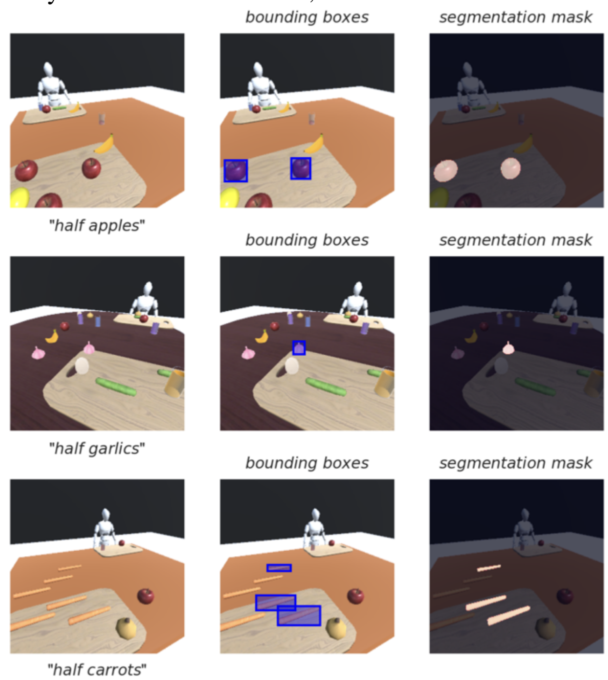
8.22. "Either"

"Either" indicates that out of two of the objects in the scene, we will select one of the object.



8.23. "Half"

"Half" selects half of the objects in the scene, typically, half would be phrased as "half the noun", however for simplicity for the determiner task, we omitted the "the"



References

[1] Aishwarya Kamath et al. "MDETR–Modulated Detection for End-to-End Multi-Modal Understanding". In: *arXiv preprint arXiv:2104.12763* (2021).

[2] Junnan Li et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models". In: *arXiv preprint arXiv:2301.12597* (2023).

[3] Peng Wang et al. "OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 23318–23340.

[4] Haotian Zhang et al. "GLIPv2: Unifying Localization and Vision-Language Understanding". In: *Thirty-sixth Conference on Neural Information Processing Systems*. 2022.