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 enbeddings 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.

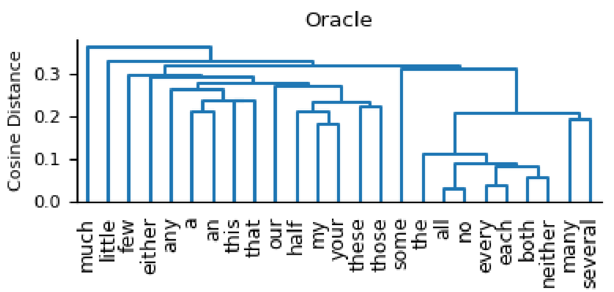


Figure 1. Dendrogram of determiner word embeddings by oracle model's FC1b layer.

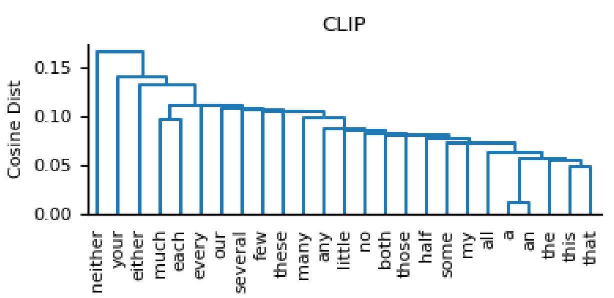


Figure 2. Dendrogram of determiner word embeddings by CLIP's text encoder.

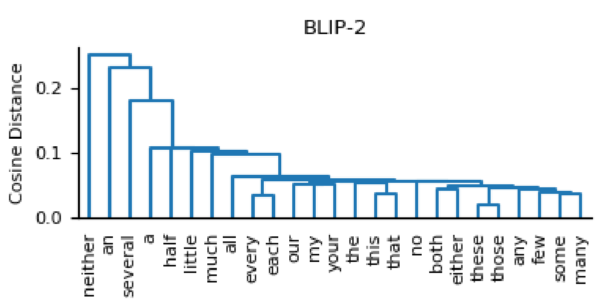


Figure 3. Dendrogram of determiner word embeddings of BLIP-2's 6.7 billion parameter text encoder.

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.

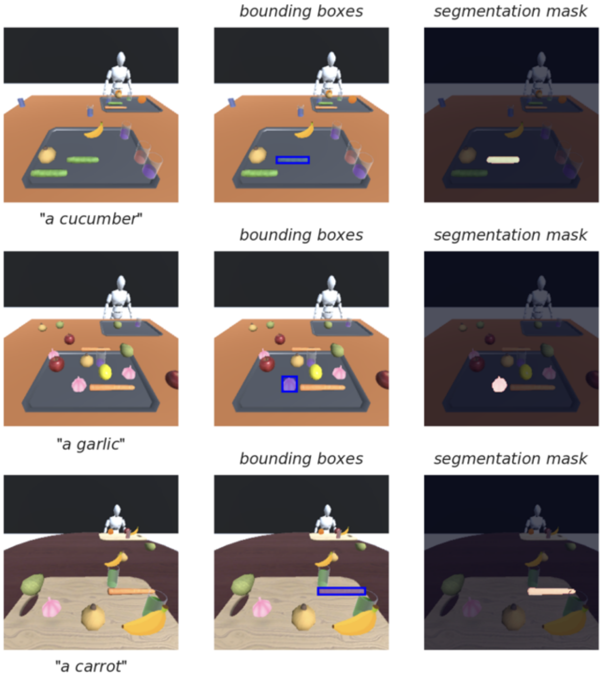
Therefore, a dynamic dataset with complex sentence structure and different contexts needs to be created for models to learn the complexities underlying possession, specific articles and negation of objects.

8. Examples for each determiner

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

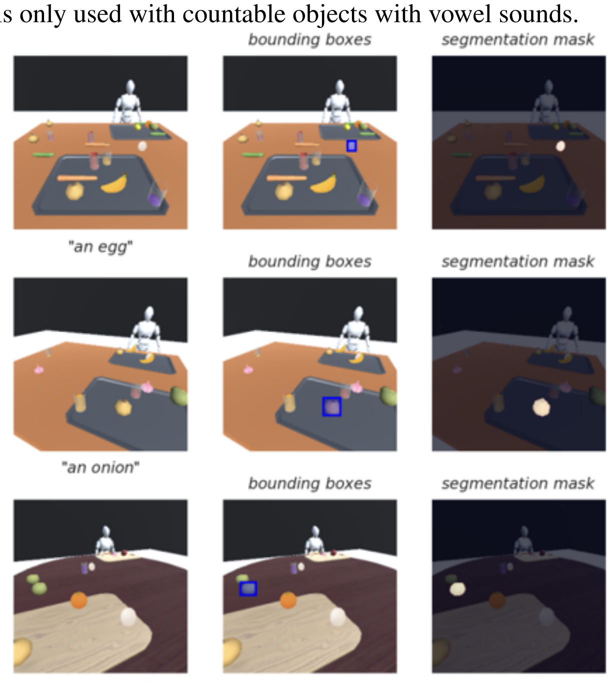
8.1. “A”

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



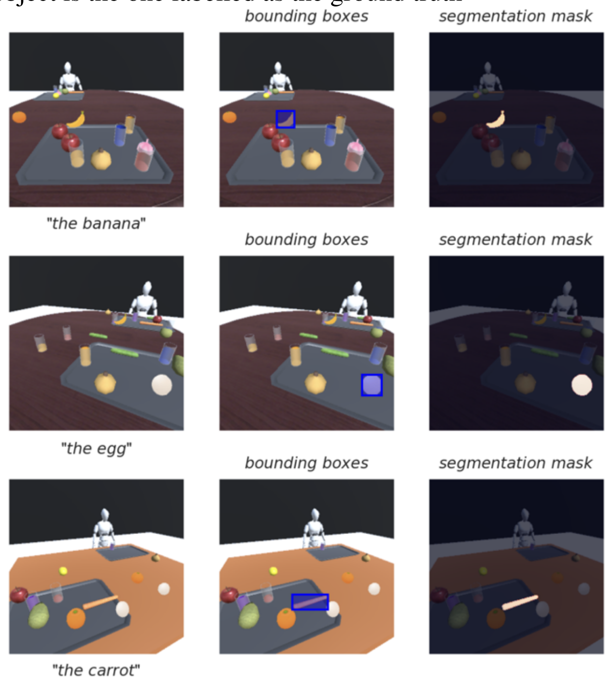
8.2. “An”

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



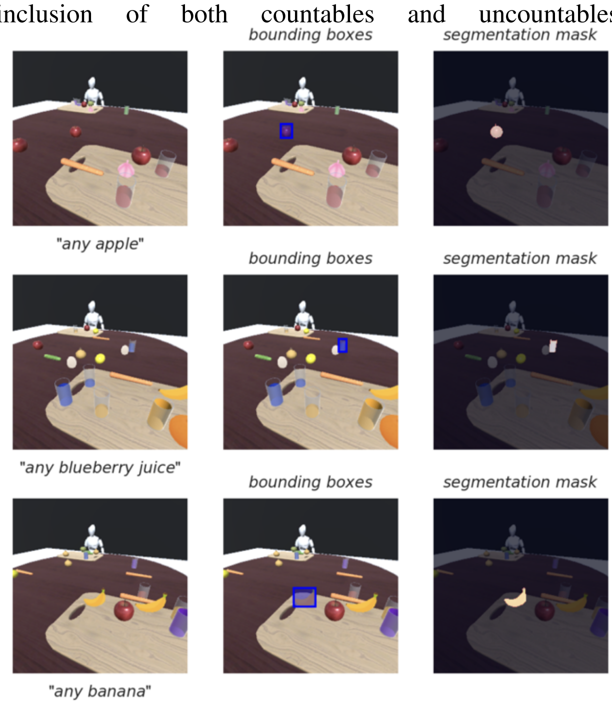
8.3. “The”

“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



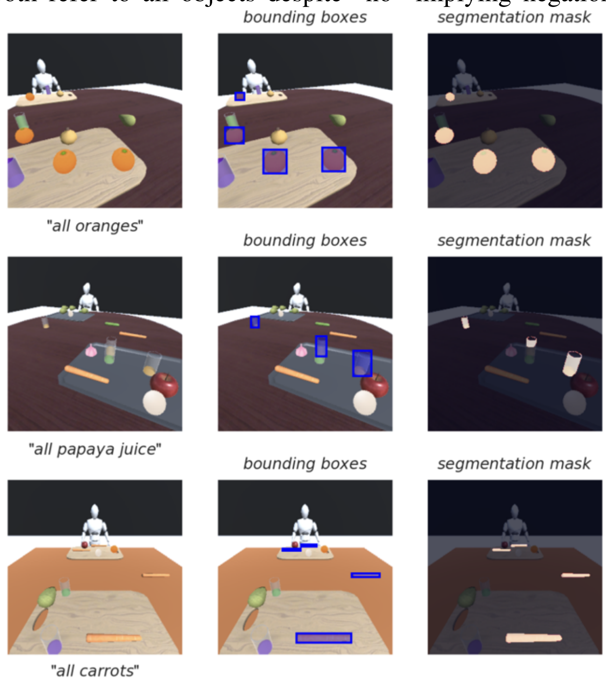
8.4. “Any”

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



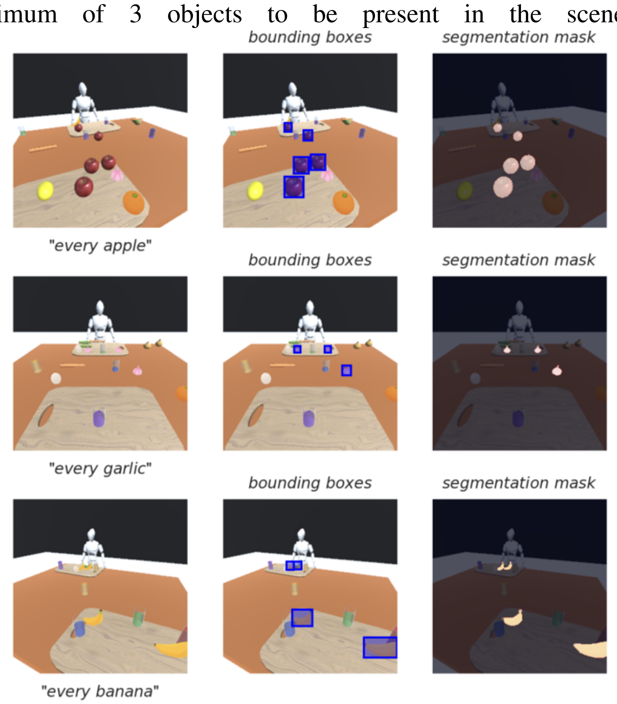
8.5. “All” / “No”

“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.



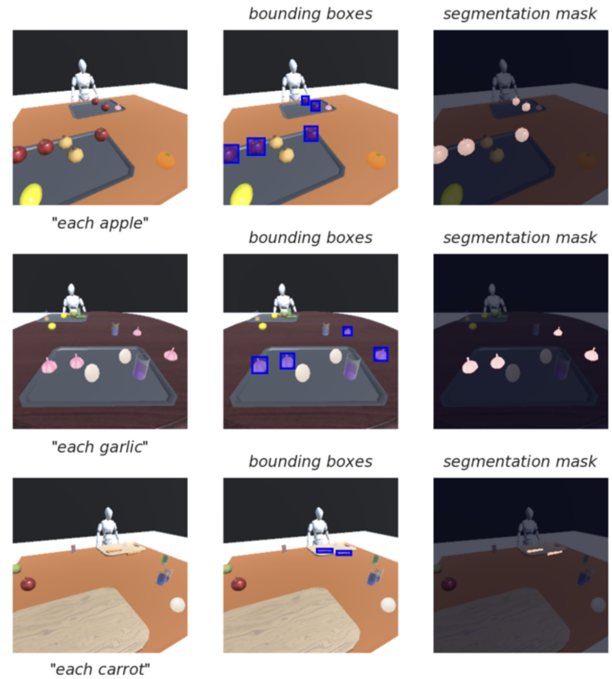
8.6. “Every”

“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



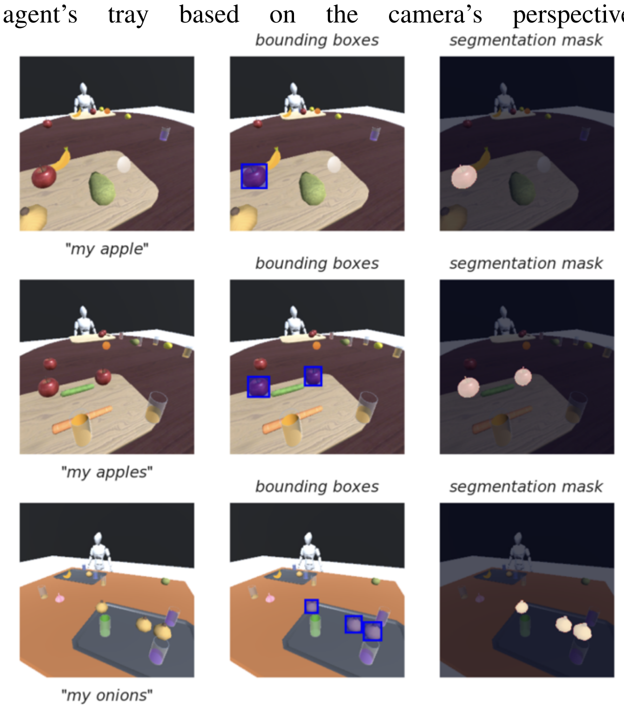
8.7. “Each”

“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



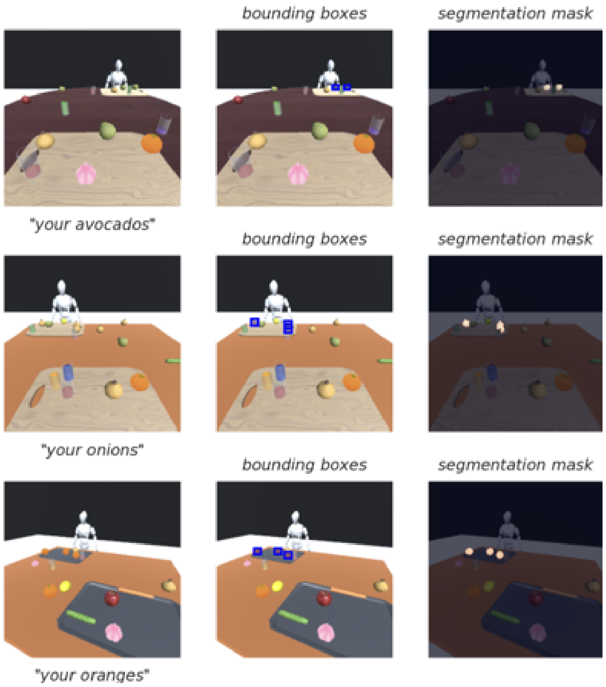
8.8. “My”

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



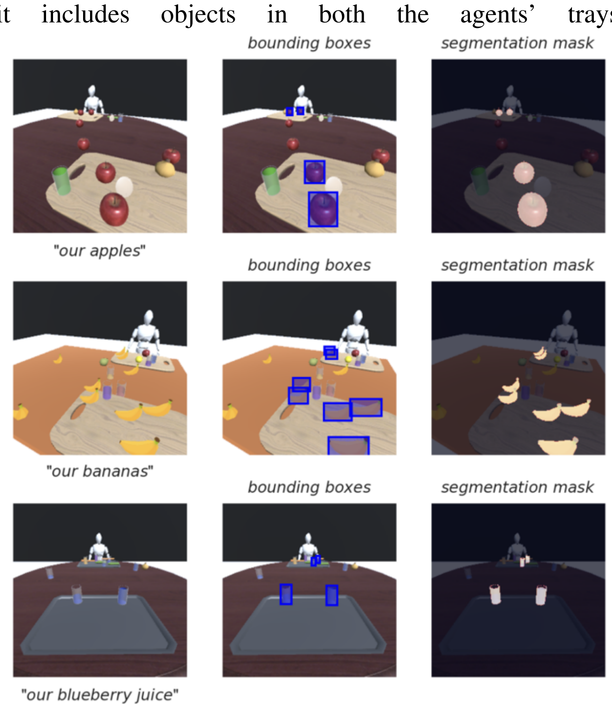
8.9. “Your”

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



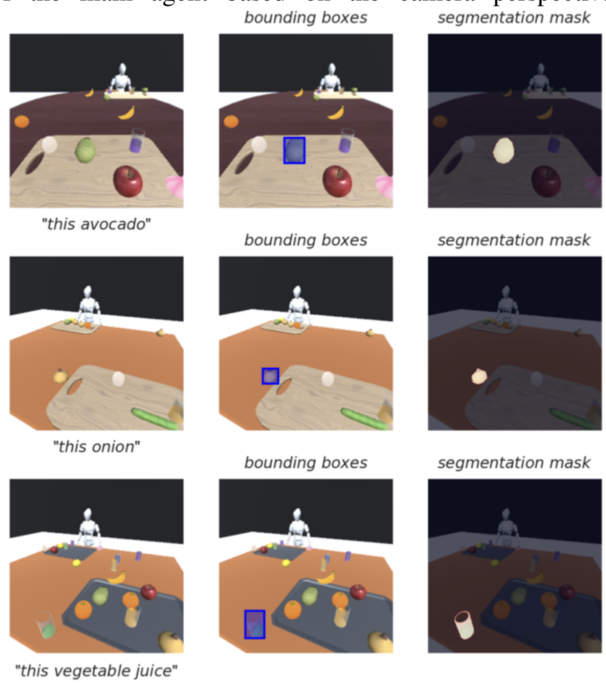
8.10. “Our”

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



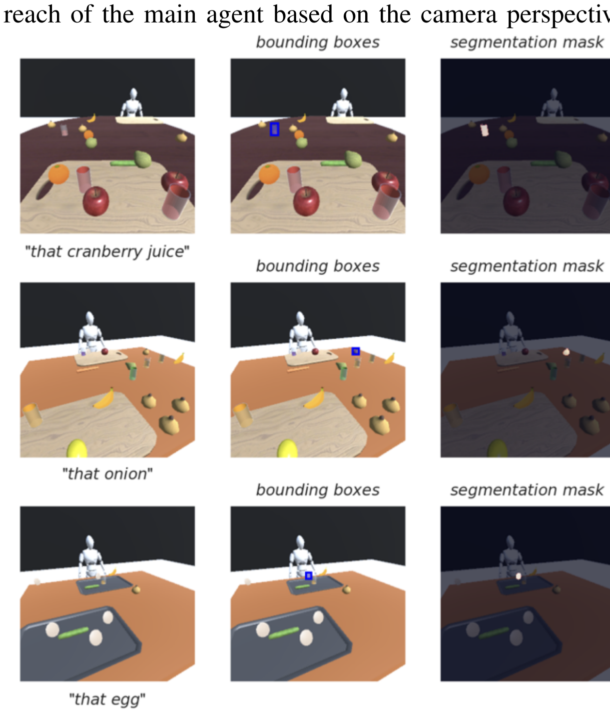
8.11. “This”

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



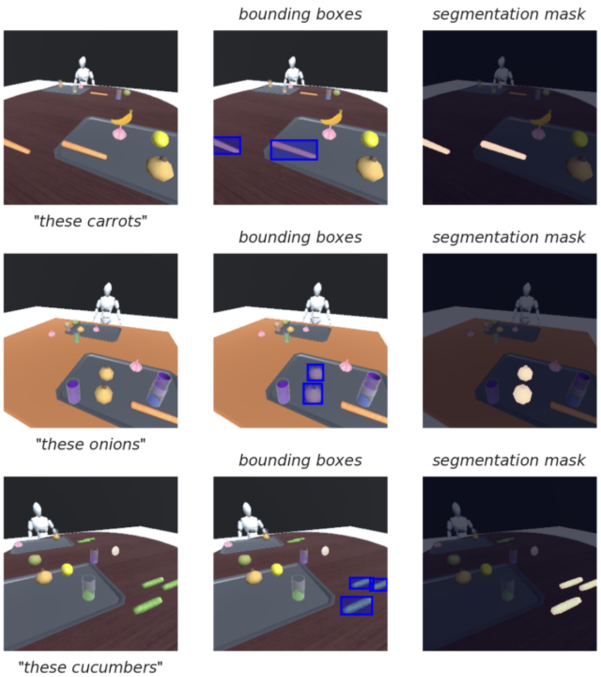
8.12. “That”

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



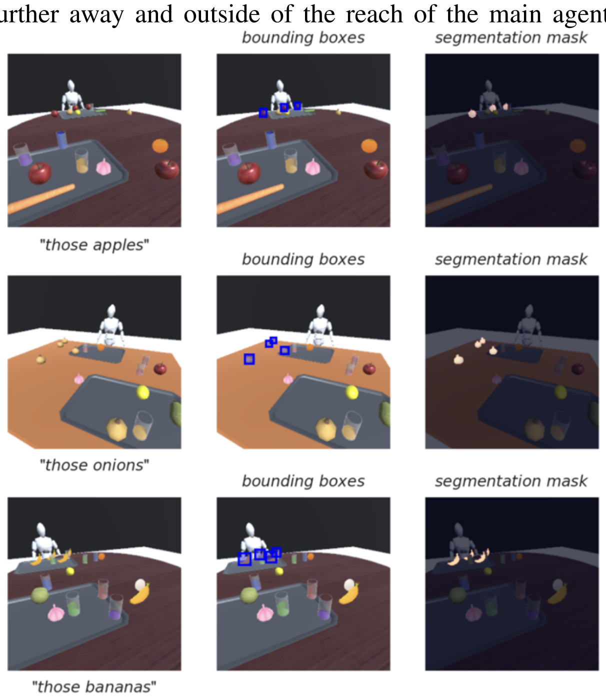
8.13. “These”

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



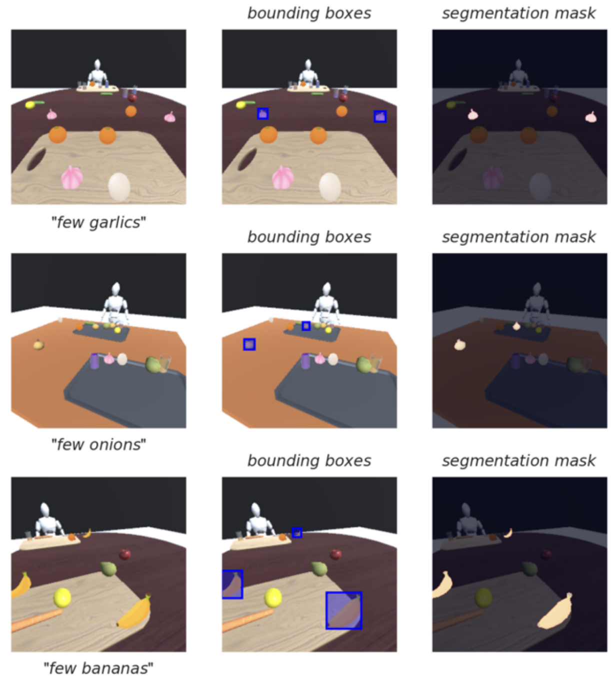
8.14. “Those”

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



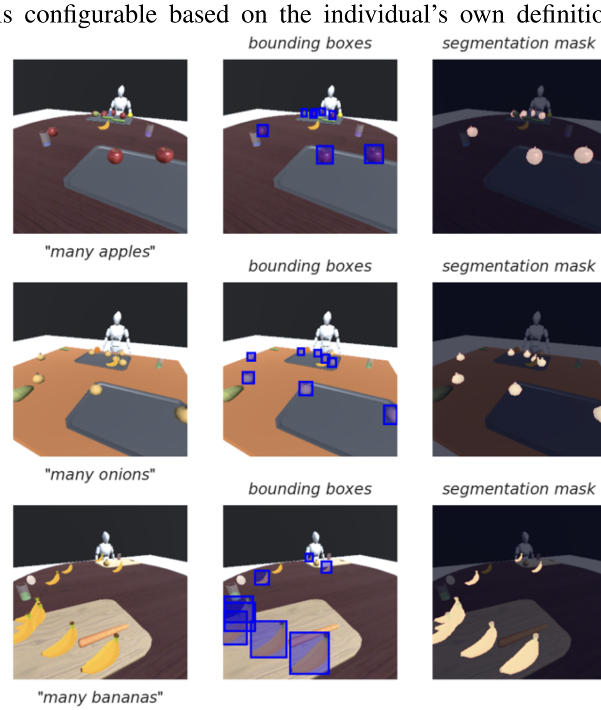
8.15. “Few”

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.



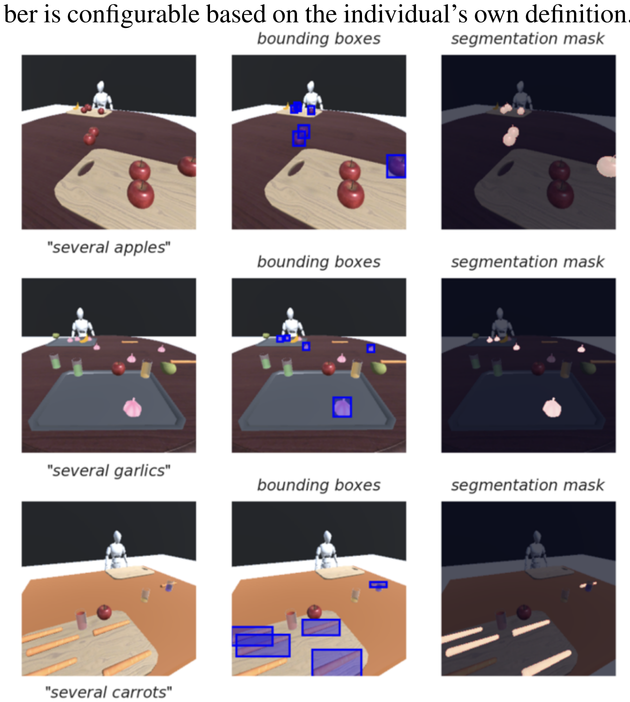
8.16. “Many”

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.



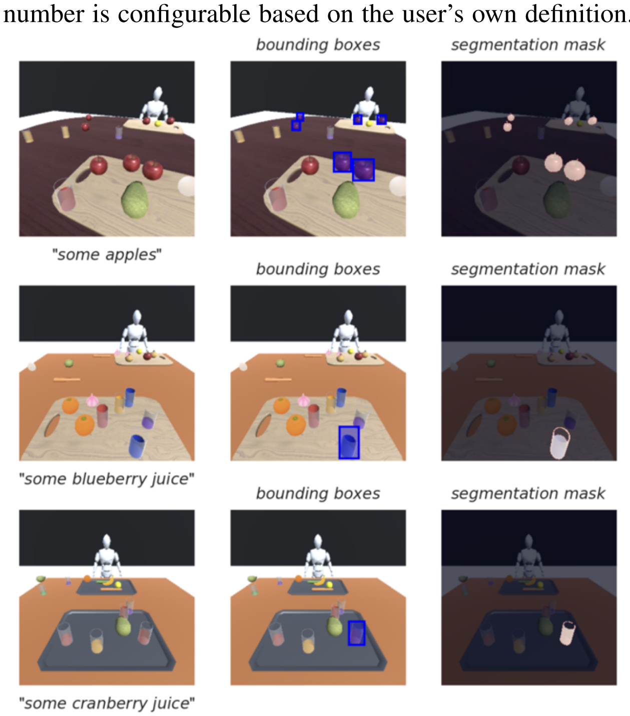
8.17. "Several"

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.



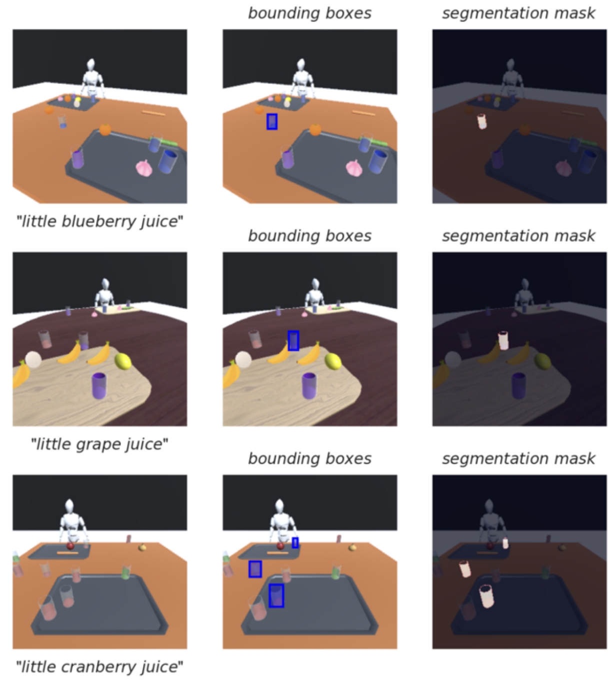
8.18. "Some"

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.



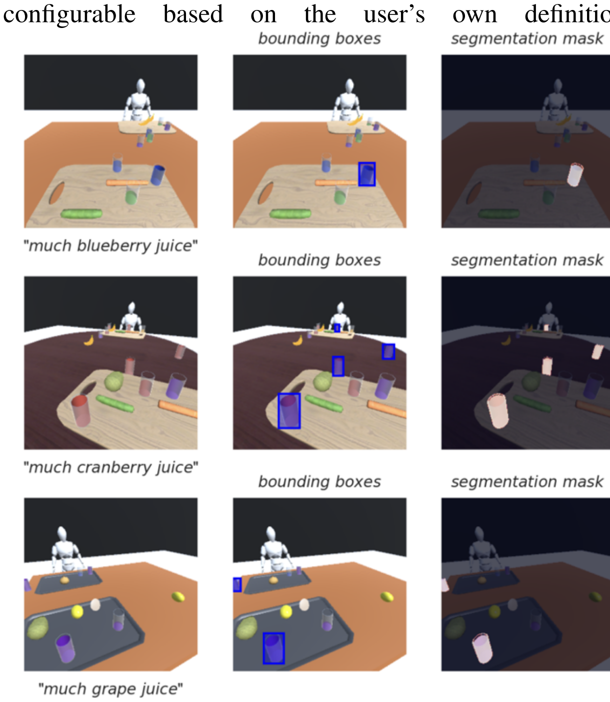
8.19. "Little"

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



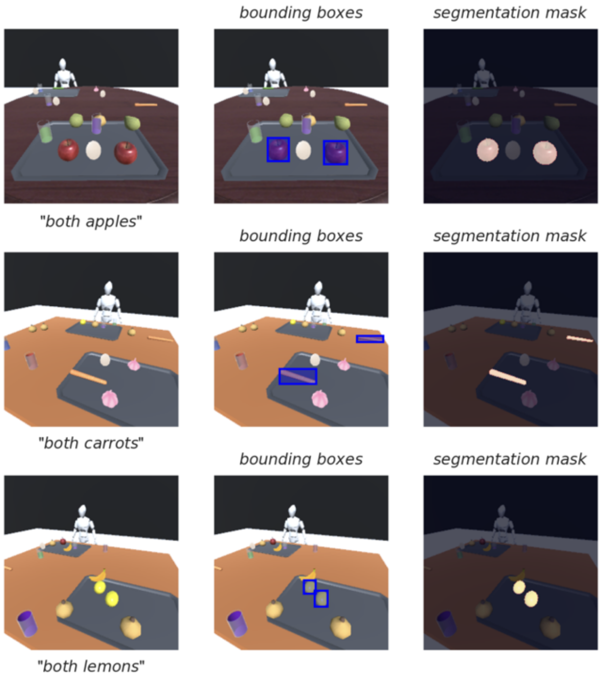
8.20. "Much"

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



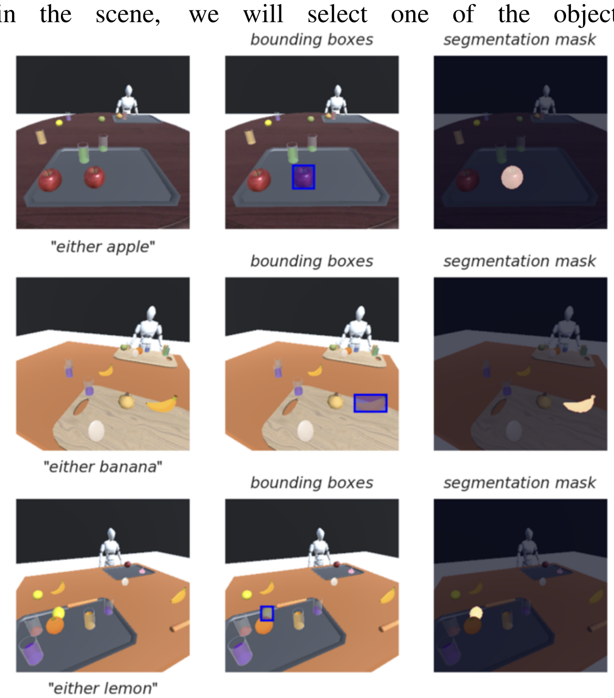
8.21. “Both” / “Neither”

“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.



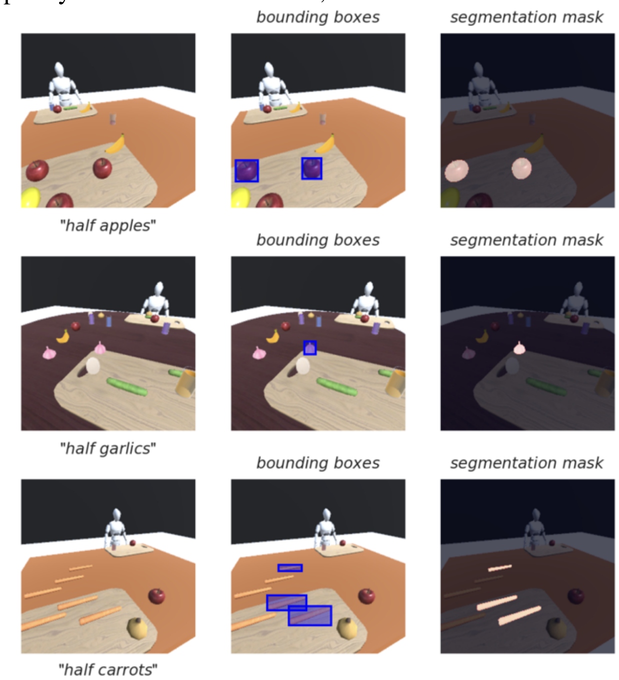
8.22. “Either”

“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.