1 OCID-VLG Vocabulary

We visualize a word cloud of the concept vocabulary of OCID-VLG in Fig. 1. Besides common sub-phrases such *"box", "food", "product"*, the wordcloud demonstrates that the most frequent concepts used to disambiguate objects are spatial predicates, both as pair-wise relations (*"front", "right", etc.*) and as absolute location (e.g. *"leftmost", "closest"*). Certain object names (e.g. *"kleenex", "tissues", "cereal"*) appear more frequently, as those are the objects that are most commonly ambiguous in OCID scenes, hence they spawn a lot of expressions referring to them. Finally, colors and brand names appear also frequently, as they are the most common discriminating attribute between objects of the same category.

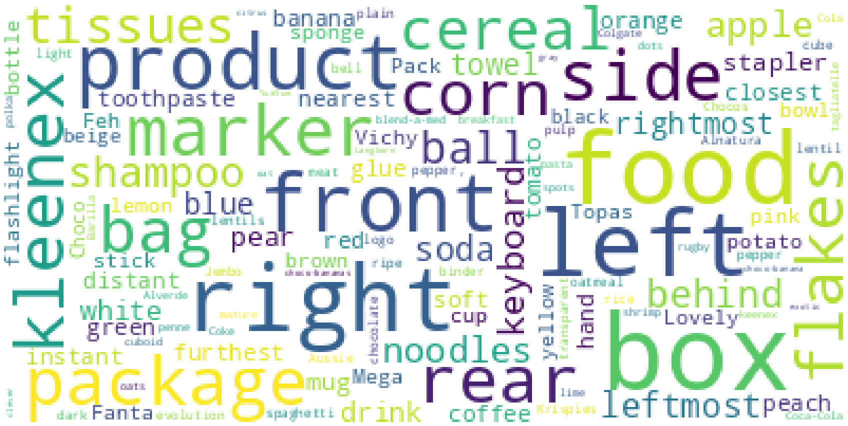


Figure 1: Wordcloud of OCID-VLG Vocabulary

The number of unique concepts per concept type, as well as the total number including paraphrases are presented in Table 1. Paraphrases include synonyms (e.g. *"Coca-Cola", "Coke"*) as well as different phrasings of relations (e.g. *"left of", "to the left side of"*).

Concept	Num.Unique	Num.Total
Category	30	55
Color	27	27
Instance	31	93
Relation	9	24
Location	4	8

Table 1: Number of concepts in OCID-VLG

Referring expressions might use instance-level names, attributes, relations, locations or combinations of all the above to disambiguate objects. We study the frequency of referring expressions on the OCID-VLG data splits in Table 2. Most frequent type is name (which includes a lot of variety in concepts such as brand, flavor etc.) with pair-wise relations following closely. Spatial relations can always refer to the target uniquely by querying for a relation to a neighbouring object. Location

Type	Train	Validation	Test
Name	20678	3014	5809
Attribute	2739	348	781
Relation	20501	2792	5769
Location	9306	1285	2672
Mixed	9997	1230	2718

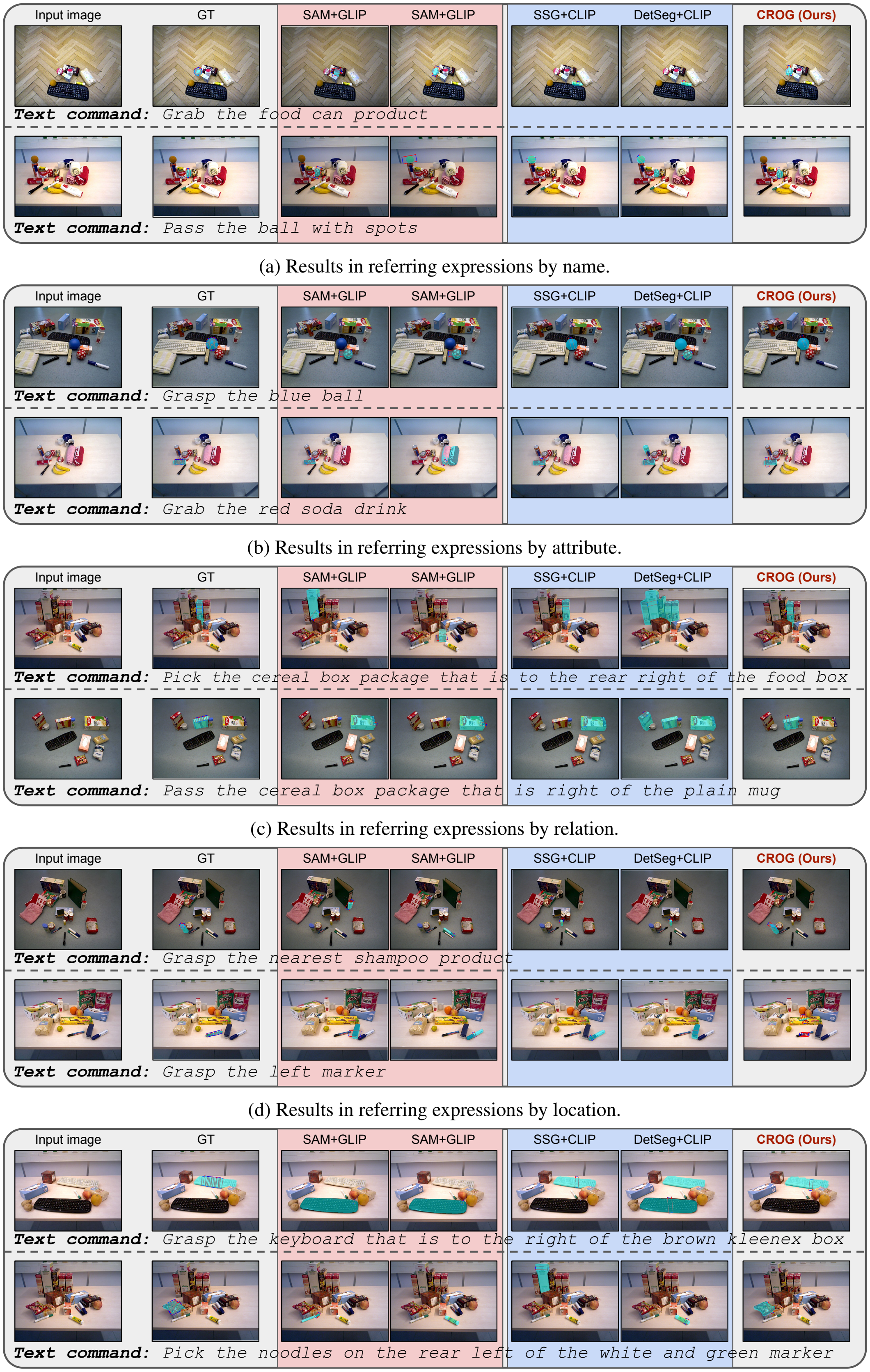
Table 2: Number of referring expressions in OCID-VLG organized by type

and mixed follow at about half frequency, while color is last, as several objects in OCID share color between different instances of the same category.

2 Qualitative Results

We visualize predicted masks and grasp poses from the implemented baselines and the proposed CROG model in Fig. 2. We include two examples per referring expression type for test scenes of OCID-VLG dataset. Zero-shot baselines based on pretrained GR-ConvNet provide poor grasp proposals, while supervised baselines + CLIP (Det-Seg, SSG) are constrained by the ranking errors of CLIP. Due to segment-then-rank pipeline, spatial information about other objects is lost when considering only the mask of a single object. As a result, CLIP-based baselines struggles with grounding spatial relations. CROG is robust across referring expression types.

In Fig. 3, we visualize outputs of the CROG model during real robot experiments. The plots include predicted mask and grasp proposal, as well as the three decoded masks from CROG's grasp projectors (quality, angle and width masks). It should be noted that the corresponding input command is shown atop each image.


Input image
GT
SAM+GLIP
SAM+GLIP
SSG+CLIP
DetSeg+CLIP
CROG (Ours)
Text command: Grab the food can product
Text command: Pass the ball with spots
Text command: Grasp the blue ball
Text command: Grab the red soda drink
Text command: Pick the cereal box package that is to the rear right of the food box
Text command: Pass the cereal box package that is right of the plain mug
Text command: Grasp the nearest shampoo product
Text command: Grasp the left marker
Text command: Grasp the keyboard that is to the right of the brown kleenex box
Text command: Pick the noodles on the rear left of the white and green marker

(a) Results in referring expressions by name.

(b) Results in referring expressions by attribute.

(c) Results in referring expressions by relation.

(d) Results in referring expressions by location.

(e) Results in referring expressions by a mix of concepts.

Figure 2: Qualitative results in OCID-VLG test scenes.

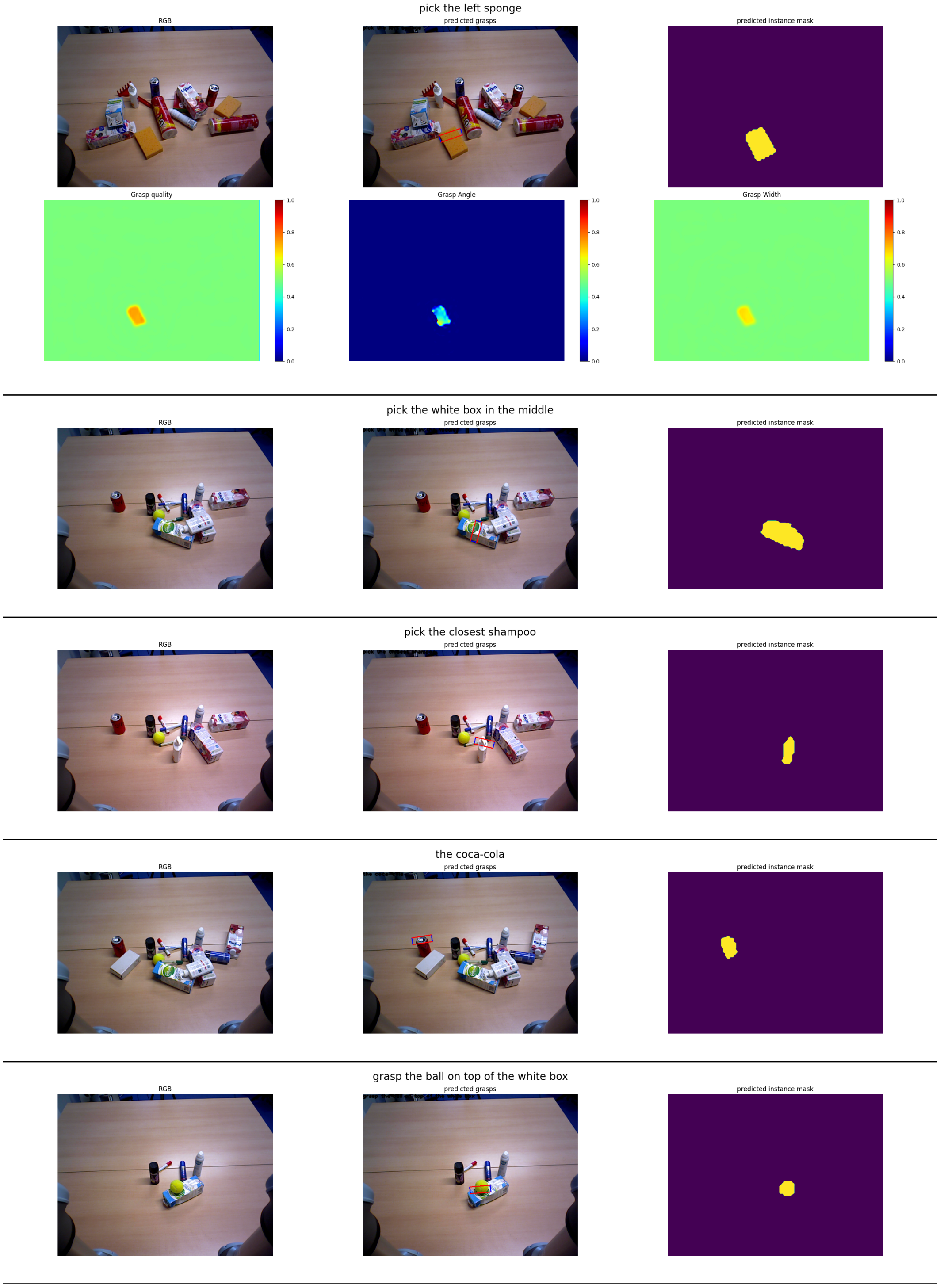


Figure 3: Qualitative results in real robot experiments.