# The devil is in the fine-grained details: Evaluating open-vocabulary object detectors for fine-grained understanding
## Supplementary Material

**Lorenzo Bianchi**[1,2], **Fabio Carrara**[1], **Nicola Messina**[1], **Claudio Gennaro**[1], **Fabrizio Falchi**[1]
[1]CNR-ISTI, Pisa, Italy [2]University of Pisa, Italy
`<name.surname>@isti.cnr.it`

## A. Dataset details

The proposed benchmarks are based on PACO (Parts and Attributes of Common Objects), an attribute-based detection dataset. PACO covers 75 object categories, encompassing 456 object-part categories and 55 attributes across image and video datasets. The attributes used to describe objects and their parts are reported in the following table.

| Type | Possible Values | | |
|---|---|---|---|
| Colors | black | light blue | blue |
| | dark blue | light brown | brown |
| | dark brown | light green | green |
| | dark green | light grey | grey |
| | dark grey | light orange | orange |
| | dark orange | light pink | pink |
| | dark pink | light purple | purple |
| | dark purple | light red | red |
| | dark red | white | light yellow |
| | yellow | dark yellow | |
| Materials | text | stone | wood |
| | rattan | fabric | crochet |
| | wool | leather | velvet |
| | metal | paper | plastic |
| | glass | ceramic | |
| Patterns | plain | striped | dotted |
| | checkered | woven | studded |
| | perforated | floral | logo |
| Transp. | opaque | translucent | transparent |

We simplified the structure of annotations in PACO to make it more straightforward for the LLM, aiming for more natural captions. Notably, we removed the *plain* pattern and *opaque* transparency from the object structure, as these are basic attributes found in almost every object without a specific pattern or transparency. Keeping them could lead to awkward sentences like *A plain opaque black dog*, but we retained them for generation of negative captions.

We also streamlined the structure by removing redundant attributes from object parts present in all components. Any missing attributes were added to the main object attributes to avoid overly complex sentences, and we also removed any part without attributes. For example, *A car with a black hood, black roof, black fender, and black bumper* became *A black car*.

In the PACO dataset, only a subset of objects has attributes, while for others, the attributes are unknown. The dataset does not provide information about whether the attributes of one object also describe others in the same scene. This poses a potential problem, as a caption generated for one object may describe others not included in the ground truth of our benchmarks, confusing a potential positive as a negative and consequently poisoning the evaluation procedure. To address this issue, we initially propagated the generated caption for a given object indiscriminately to all the other objects having the same class. Then, objects inconsistent with the assigned captions were removed during manual revision.

## B. Captions generation

The generation process leverages prompt engineering and the in-context learning capabilities of LLMs [1]. We present the model with pairs of object structures and corresponding natural language descriptions as illustrative examples. The model is then prompted to generate new captions based on queries describing the structural aspects of novel objects (see Figure 1). In cases where the generated captions fail to meet predefined criteria, such as incomplete attribute utilization or excessive length, we initiate an automatic *iterative prompting* process involving posing targeted follow-up questions to address empirically identified issues and refine the generated captions (see Figure 2). We describe this approach in detail in Algorithm 1. We could apply this methodology indefinitely until all captions meet our criteria, but we limited to one iteration, as just a single one yielded

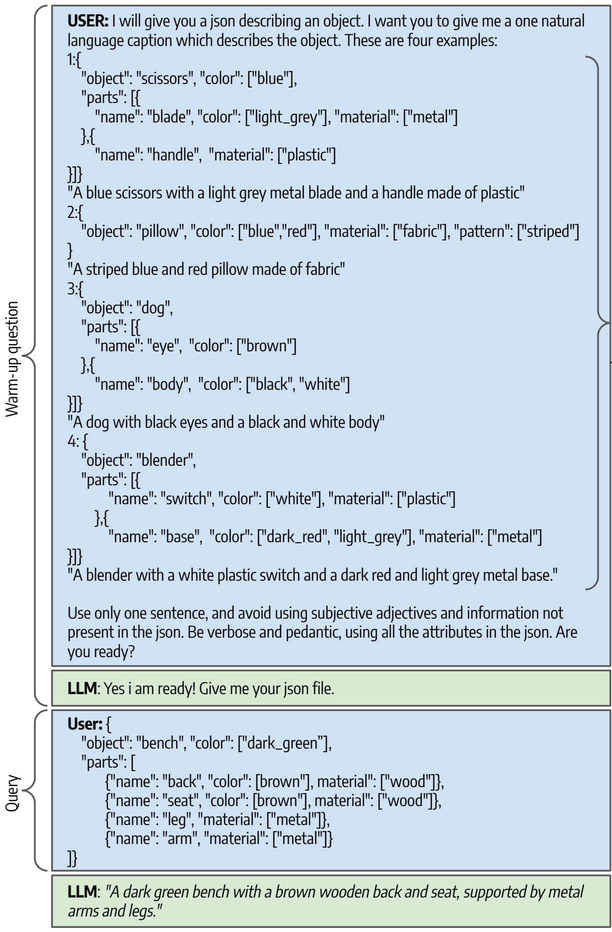


Figure 1. **Captions Generation.** We query the LLM with a JSON description outlining the object's parts and attributes and four in-context samples to enhance caption quality and mitigate hallucination risks. We adopt a *warm-up question* to separate the query from the in-context examples, as it avoids the LLM confusing the attributes of the examples with those of the query object.

sufficient data, removing the need for further iterations.

Subsequent to caption generation, we meticulously conducted a manual revision process to uphold the benchmark's quality and accuracy.

## C. Model Architectural Details

In our experiments, we employed the following architectural configurations of detectors:

- For OWL-based models (**OWL-ViT** and **OWLv2**), we evaluated configurations with ViT B/16 and ViT L/14 backbones.
- For **ViLD**, we utilized the configuration featuring Resnet-152 as a backbone, with a distillation weight set to 0.1.
- **Detic** was configured with the larger setup, employing Swin-B as the backbone and ImageNet-21K pre-training, specifically, the `Detic_LCOCOI21k_CLIP_SwinB_896b32_4x_ft4x_max-size` configuration.
- **GroundingDino** was instantiated using the GroundingDINO-T configuration, employing Swin-T as the backbone.
- For **CORA**, we utilized the model with Resnet50x4 as the backbone.

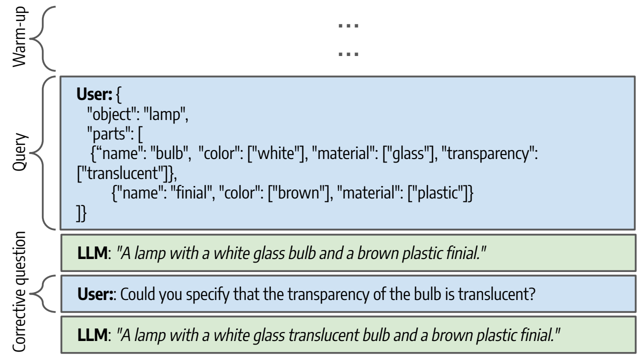


Figure 2. **Caption Correction.** If the initial caption did not meet our predefined criteria, we asked the LLM to generate an improved version, addressing the identified issues. The various follow-up questions for this iterative prompting process are outlined in Listing 1.

As the captions in our benchmark are in natural language, all inferences were conducted without any pre-appended prompts. The sole exception is observed in *CORA*, which employs an internal prompt ensemble, amalgamating 80 distinct prompts. In Figure 3, we present the results of this model with prompt ensemble disabled, processing the input caption without any modifications. These results indicate that the performances are not significantly affected by the prompt ensemble.

## D. Handling Grounding Dino

GroundingDino diverges from other detectors in its approach, as it is mainly conceived for REC and cannot be fed directly with a dictionary of captions but only with a single caption. For this reason, we made an inference for each caption within the vocabulary. Each prediction is then represented as $d_i = (\mathbf{b}_i, h_i, t_i)$, with $\mathbf{b}_i$ being the bounding box coordinate and $h_i$ denoting the score value assigned to the caption $t_i$. This differs from other open-vocabulary object detectors, where each prediction incorporates a score array $\mathbf{s}_i$, and each element reflects the score associated with the corresponding caption. Importantly, this distinction does not impact the mAP of the detector, as it solely considers the predicted label for each corresponding predicted box. However, this difference affects the rank metric, as the score array $\mathbf{s}_i$ over the vocabulary entries is needed.

For the rank calculation, where we need to rank all the possible vocabulary elements for each object, we consid-

**Algorithm 1** Caption Generation through Iterative Prompting

```
procedure CreateCaption(object, followup_prompts, n_iterations)
    prompt ← CreatePrompt(object)                    ▷ Creates a prompt with 4 in-context examples + object structure
    i ← 0
    while i < n_iterations do
        caption ← LLM(prompt)
        identified_problem ← CheckIssues(caption, object)    ▷ Empirically check for issues in caption generation
        if identified_problem is None then
            return caption                                   ▷ The caption is correct
        end if
        prompt ← prompt + caption + followup_prompts[identified_problem]
        i ← i + 1
    end while
    return None                                      ▷ Caption incorrect after n_iterations
end procedure
```

Listing 1. List of prompts employed in Iterative Prompting for correcting inaccurate captions. Each prompt is accompanied by a comment elucidating the condition that prompts the activation of that particular question.

```
followup_prompts = {
    # caption with more than 60 words
    0: "Your answer was too long. Create only one sentence for the object that describes
       what the object looks like considering its attributes",
    # ' is a ' inside the caption
    1: "Your answer is a definition of what the object is. Give me a caption that only
       describes the object and its attributes",
    # object not inserted
    2: "You did not specify that you are describing a {object_name}. Reformulate the caption
        with this addition",
    # part not inserted
    3: "You did not specify that the {object_name} has a {part_name}. Reformulate the
       caption with this addition",
    # attribute not considered
    4: "Could you specify that the {attribute_type} of the {object_name} is {
       attribute_value}?",
    # ':' in the caption
    5: "Do not list the elements of the object. Summarize the description of the object in a
        natural language caption",
    # more than 2 '"'
    6: "You gave me more than one caption. Summarize them in only one caption",
    # a number in the caption
    7: "Your answer contains a number not present in the JSON. Create a new caption
       considering only the attributes I gave you and without adding information",
    # only one '"' in the caption
    8: "Answer is not complete. Write a complete caption",
    # found an illegal character
    9: "Illegal characters in the caption. Remove them",
    # 'or' in the caption
    10: "Ensure that the attributes are described using 'and' instead of 'or' to correctly
       represent all the specified attributes.",
    # 'single' in the caption
    11: "You used the word 'single'; reformulate the caption without it"
}
```

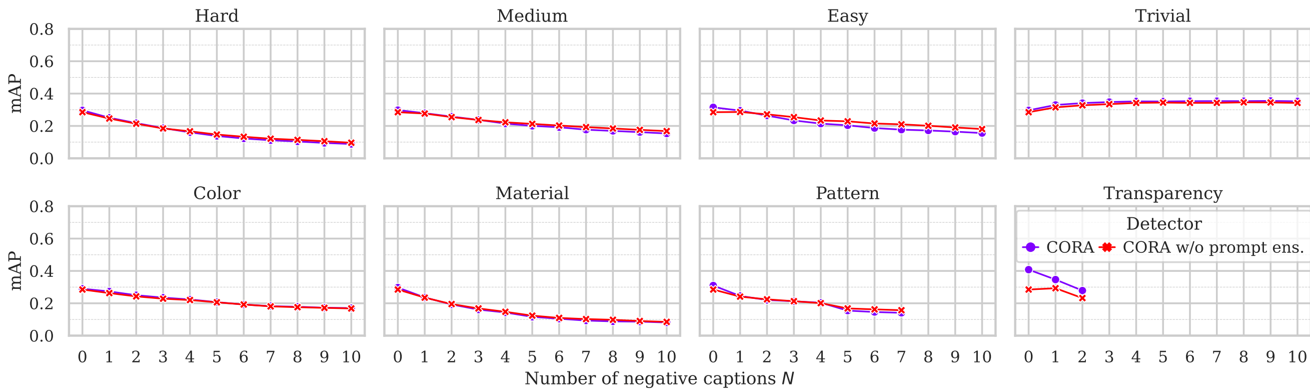


Figure 3. **Effect of the number of the prompt ensemble on CORA**

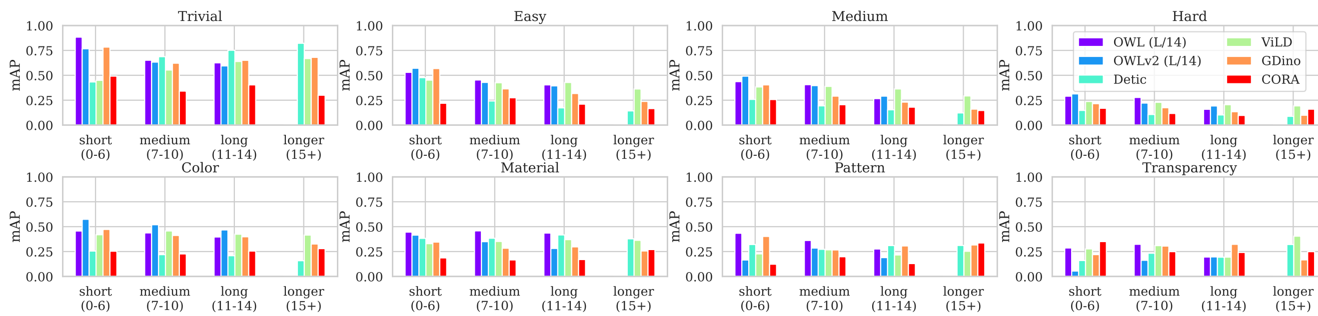


Figure 4. **Effect of the caption length**: We illustrate the mAP of detectors across Difficulty-based ($N$ = *5*) and Attribute-based ($N$ = *2*) benchmarks, with varying caption lengths for objects. Captions are categorized into four groups based on their average word count inside the corresponding vocabulary: *short* (6 or fewer words), *medium* (7-10 words), *long* (11-14 words), and *longer* (15 or more words). OWL-based detectors are excluded from the *longer* group due to their inability to process captions exceeding 16 tokens.

ered all generated predictions $d_j$ for the ground truth object $o_i$, without employing the class-agnostic NMS typically applied in our evaluation protocol. Instead, we set to zero the confidence score of predictions not overlapping with the ground truth, i.e., $h_j \leftarrow 0$ if $IoU(o_i, d_j) < 0.5$. The score $\mathbf{s}_i$ on which the rank is calculated is derived by considering, for all elements in the vocabulary of $o_i$, the prediction score associated with the caption having the highest confidence. Subsequently, the median rank is computed following the same procedure employed for the other detectors.

## E. Additional Results: mAP Small, Medium and Large

The data in Table 1 and Table 2 offers a comprehensive view of mAP scores across Difficulty-based and Attribute-based benchmarks, segmented by object sizes. As expected, the task complexity increases with smaller-area objects, as finer details become more discernible with larger objects.

## F. Additional Results: Role of caption lengths

Our generated captions have a significant variability in terms of length, which also depends on the particular benchmark employed. Therefore, there exists the possibility that the results reported in the paper are correlated with the sentence length. For this reason, we report in Figure 4 the mAP evaluated on subsets generated by grouping the sentences by length (intended as the average number of words in the annotation vocabulary).

These results show that, on average, the difficulty of the task increases slightly as the caption length increases, as indicated by the weak negative correlation (Pearson coefficient) between length and mAP:

| OWL (L/14) | OWLv2 (L/14) | Detic | ViLD | GDino | CORA |
|---|---|---|---|---|---|
| -0.34 | -0.22 | -0.02 | 0.05 | -0.31 | -0.08 |

It is interesting to note that, while OWL-based detectors and GroundingDINO show a moderate effect of caption length on mAP, ViLD, Detic, and CORA show greater resilience to such variations. In general, we observe that

Table 1. Mean Average Precision (mAP) of detectors on the Hard, Medium, Easy and Trivial sets of negatives (N=5), segmented by object sizes ($mAP_S$, $mAP_M$, $mAP_L$)

| | Hard | | | Medium | | | Easy | | | Trivial | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $mAP_S$ | $mAP_M$ | $mAP_L$ | $mAP_S$ | $mAP_M$ | $mAP_L$ | $mAP_S$ | $mAP_M$ | $mAP_L$ | $mAP_S$ | $mAP_M$ | $mAP_L$ |
| OWL (B/16) | 11.0 | 21.7 | **28.3** | 12.5 | 33.0 | 43.1 | 11.6 | 29.4 | 44.3 | 15.0 | 44.6 | 60.1 |
| OWL (L/14) | 13.8 | 22.7 | 27.0 | 18.3 | 34.4 | 39.9 | 19.3 | 39.3 | 42.9 | 33.0 | 56.0 | 67.9 |
| OWLv2 (B/16) | 14.8 | 21.9 | 26.4 | 19.5 | 34.2 | 38.8 | 15.9 | 33.4 | 44.0 | 26.3 | 48.0 | 54.6 |
| OWLv2 (L/14) | 14.0 | 21.9 | 26.2 | **24.2** | 34.6 | **43.9** | 13.8 | 34.3 | **48.6** | 32.0 | 55.8 | 65.8 |
| Detic | 10.5 | 11.5 | 12.2 | 11.6 | 20.2 | 18.4 | 19.9 | 19.3 | 19.1 | **39.4** | **71.1** | **75.0** |
| ViLD | **15.0** | **24.0** | 23.2 | 21.6 | **40.0** | 38.5 | **20.8** | **43.0** | 44.8 | 32.5 | 63.2 | 61.8 |
| GDino | 5.1 | 17.0 | 17.0 | 6.6 | 28.2 | 29.8 | 7.7 | 30.4 | 31.9 | 19.1 | 58.5 | 72.4 |
| CORA | 4.2 | 13.2 | 17.3 | 7.6 | 19.1 | 25.7 | 6.0 | 18.5 | 26.5 | 12.0 | 32.9 | 46.4 |

Table 2. Mean Average Precision (mAP) of detectors on the Color, Material, Transparency and Pattern sets of negatives (N=2), segmented by object sizes ($mAP_S$, $mAP_M$, $mAP_L$)

| | Color | | | Material | | | Transparency | | | Pattern | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $mAP_S$ | $mAP_M$ | $mAP_L$ | $mAP_S$ | $mAP_M$ | $mAP_L$ | $mAP_S$ | $mAP_M$ | $mAP_L$ | $mAP_S$ | $mAP_M$ | $mAP_L$ |
| OWL (B/16) | 15.6 | 39.1 | 48.2 | 11.2 | 29.7 | 42.5 | 17.0 | **31.3** | 37.5 | 2.7 | 24.8 | 28.3 |
| OWL (L/14) | 24.1 | 37.1 | 45.7 | **22.0** | 39.0 | **47.0** | 20.1 | 26.2 | 30.7 | 11.2 | **34.6** | **35.4** |
| OWLv2 (B/16) | 24.5 | 42.1 | 44.6 | 14.9 | 27.1 | 37.9 | **20.2** | 26.6 | 29.7 | 18.0 | 19.8 | 20.5 |
| OWLv2 (L/14) | **32.2** | 47.1 | **53.9** | 17.1 | 30.9 | 40.6 | 3.9 | 8.5 | 15.8 | 11.2 | 24.4 | 22.8 |
| Detic | 16.1 | 23.2 | 21.1 | 20.1 | **39.4** | 42.2 | 0.0 | 28.9 | **40.0** | **27.6** | 33.5 | 30.5 |
| ViLD | 26.3 | **49.6** | 44.7 | 16.7 | 37.6 | 39.9 | 13.9 | 30.9 | 34.0 | 13.0 | 28.9 | 24.0 |
| GDino | 12.1 | 39.4 | 45.9 | 7.0 | 28.0 | 34.4 | 19.7 | 22.8 | 27.6 | 7.6 | 26.9 | 34.7 |
| CORA | 9.4 | 24.8 | 32.5 | 7.7 | 17.0 | 26.8 | 10.8 | 25.8 | 34.9 | 4.2 | 21.9 | 28.2 |

the overall correlation remains quite bounded, meaning that caption length does not consistently affect the final results.

## G. Additional Benchmark Samples

We show additional samples from our benchmarks in Figures 8 (Color), 9 (Material), 10 (Pattern), 11 (Transparency), 12 (Trivial), 13 (Easy), 14 (Medium), and 15 (Hard).

## H. OWL Subset

Since OWL processes sentences not exceeding 16 words, we tried to re-run the experiments over all the detectors with captions longer than 16 words removed. We present the updated statistics of the proposed benchmarks filtered using this constraint in Table 3. The corresponding results for all models on this subset are detailed in Table 4, Table 5, and Figure 5. Notably, due to the limited number of removed annotations, the results exhibit minimal deviation from those of the complete benchmarks, with each detector following a consistent trend. This suggests that the important information is likely placed early in the caption, and the last part in longer sentences can be ignored.

## I. PACO-provided captions

The original PACO dataset offers a collection of 5,000 text entries corresponding to objects within the dataset. We run our evaluation protocol over an alternative benchmark built on these captions for completeness.

To create this benchmark, we randomly selected one caption from the ones linked to each object, considering that each object could be associated with more than one caption. Notice that we needed to run a meticulous manual revision process like the one implemented for our official benchmark. In this process, the caption associated with an object was systematically allocated to all objects within the same image categorized under the identical class. Subsequently, we manually assessed the objects, wherein objects discordant with the assigned caption were removed. We reported the statistics of the resulting benchmark in Table 8, and the results of the evaluated detectors in Table 6, Table 7 and Figure 6.

However, in our main analysis, we still preferred to employ the captions generated using LLMs as described in Appendix B. We followed such methodology for a series of reasons, explained in the following paragraphs.

Table 3. **Benchmark based on OWL-compatible captions:** Statistics of the benchmarks based on **OWL-subset** benchmark for each different negative set comprising the number of images (Imgs), the number of annotated objects (Objs), objects-to-image ratio (Objs/Img), positive captions, positive captions per image, negative captions per positive caption, and objects per positive caption.

| Name | Negative Set Strategy | Imgs | Objs | Obj/Img | ✓Caps | ✓/Img | ✗/✓ | Objs/✓ |
|---|---|---|---|---|---|---|---|---|
| Hard | Random attribute subst. ($\times 1$) | 1390 | 2903 | 2.1 | 1816 | 1.3 | 9.9 | 1.6 |
| Normal | Random attribute subst. ($\times 2$) | 1187 | 2293 | 1.9 | 1483 | 1.2 | 10.0 | 1.5 |
| Easy | Random attribute subst. ($\times 3$) | 417 | 657 | 1.6 | 445 | 1.1 | 10.0 | 1.5 |
| Trivial | Random captions | 1389 | 2888 | 2.1 | 1810 | 1.3 | 10.0 | 1.6 |
| Color | Color attribute subst. | 1269 | 2485 | 2.0 | 1595 | 1.3 | 10.0 | 1.6 |
| Material | Material attribute subst. | 1277 | 2611 | 2.0 | 1639 | 1.3 | 10.0 | 1.6 |
| Transparency | Transparency attribute subst. | 177 | 323 | 1.8 | 180 | 1.0 | 2.0 | 1.8 |
| Pattern | Pattern attribute subst. | 188 | 294 | 1.6 | 193 | 1.0 | 7.2 | 1.5 |

| | Hard | Medium | Easy | Trivial |
|---|---|---|---|---|
| OWL (B/16) | 26.2 | 39.8 | 38.4 | 53.9 |
| OWL (L/14) | **26.5** | 39.3 | 44.0 | 65.1 |
| OWLv2 (B/16) | 25.3 | 38.5 | 40.0 | 52.9 |
| OWLv2 (L/14) | 25.4 | **41.2** | 42.8 | 63.2 |
| Detic | 12.3 (+0.8) | 20.9 (+2.3) | 22.3 (+3.7) | **68.1** (-1.6) |
| ViLD | 22.8 (+0.7) | 38.2 (+2.1) | 44.0 (+4.1) | 54.8 (-1.8) |
| GDino | 18.7 (+2.1) | 32.0 (+4.1) | 35.1 (+5.0) | 62.6 (-0.1) |
| CORA | 13.4 (-0.4) | 21.8 (+1.8) | 23.2 (+2.8) | 33.3 (-1.8) |

Table 4. **Benchmark based on OWL-compatible captions:** mAP on Difficulty-based benchmarks ($N = 5$).

| | Color | Material | Pattern | Transp. |
|---|---|---|---|---|
| OWL (B/16) | 45.3 | 37.3 | 26.6 | **34.1** |
| OWL (L/14) | 43.8 | **44.9** | **36.0** | 29.2 |
| OWLv2 (B/16) | 45.1 | 33.5 | 19.2 | 28.5 |
| OWLv2 (L/14) | **53.3** | 36.9 | 23.3 | 12.2 |
| Detic | 23.2 (+1.7) | 38.8 | 31.1 (+1.0) | 21.8 (-6.2) |
| ViLD | 43.9 (+0.7) | 34.7 (-0.2) | 25.6 (+1.1) | 27.6 (-2.5) |
| GDino | 43.2 (+2.2) | 30.9 (+0.7) | 31.1 (-0.1) | 26.9 (+1.5) |
| CORA | 24.3 (-0.7) | 17.3 (-2.0) | 16.9 (-5.1) | 28.4 (+0.5) |

Table 5. **Benchmark based on OWL-compatible captions:** mAP on Attribute-based benchmarks ($N = 2$).

**Limited Language Expressivity** PACO captions exhibit a tendency towards uniform syntactic structures, whereas we noticed that the utilization of an LLM introduces a welcomed variability. This variability facilitates the exploration of diverse natural language contexts, thereby enabling the evaluation of the detector in a broader array of scenarios. Furthermore, PACO captions occasionally manifest as linguistically unnatural — i.e., parts are always singular even if there are multiple instances of the same part, as shown in Figure 7.

**Multiple Shorter Captions** Despite PACO's high number of captions, a single PACO object may be associated with multiple less detailed captions. For instance, a towel with attributes like black color and fabric material may be found encoded in three different captions, such as *A black towel*, *A fabric towel*, and *A black fabric towel.* This can be considered a limit in our scenario, where our evaluation protocol works by ingesting ad-hoc crafted negatives obtained by modifying a single attribute in a possibly long, detailed sentence.

**Limited Diversity** The quantity of object groups within benchmarks derived from PACO captions is notably limited. The issue becomes apparent when examining the Transparency benchmark, which features a notably restricted number of object groups, as evidently shown in the Transparency row of Table 8. This inherent scarcity is further exacerbated by OWL-based detectors being confined to a subset of each benchmark. Consequently, even a single error can induce substantial fluctuations in the measured mAP.

## References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

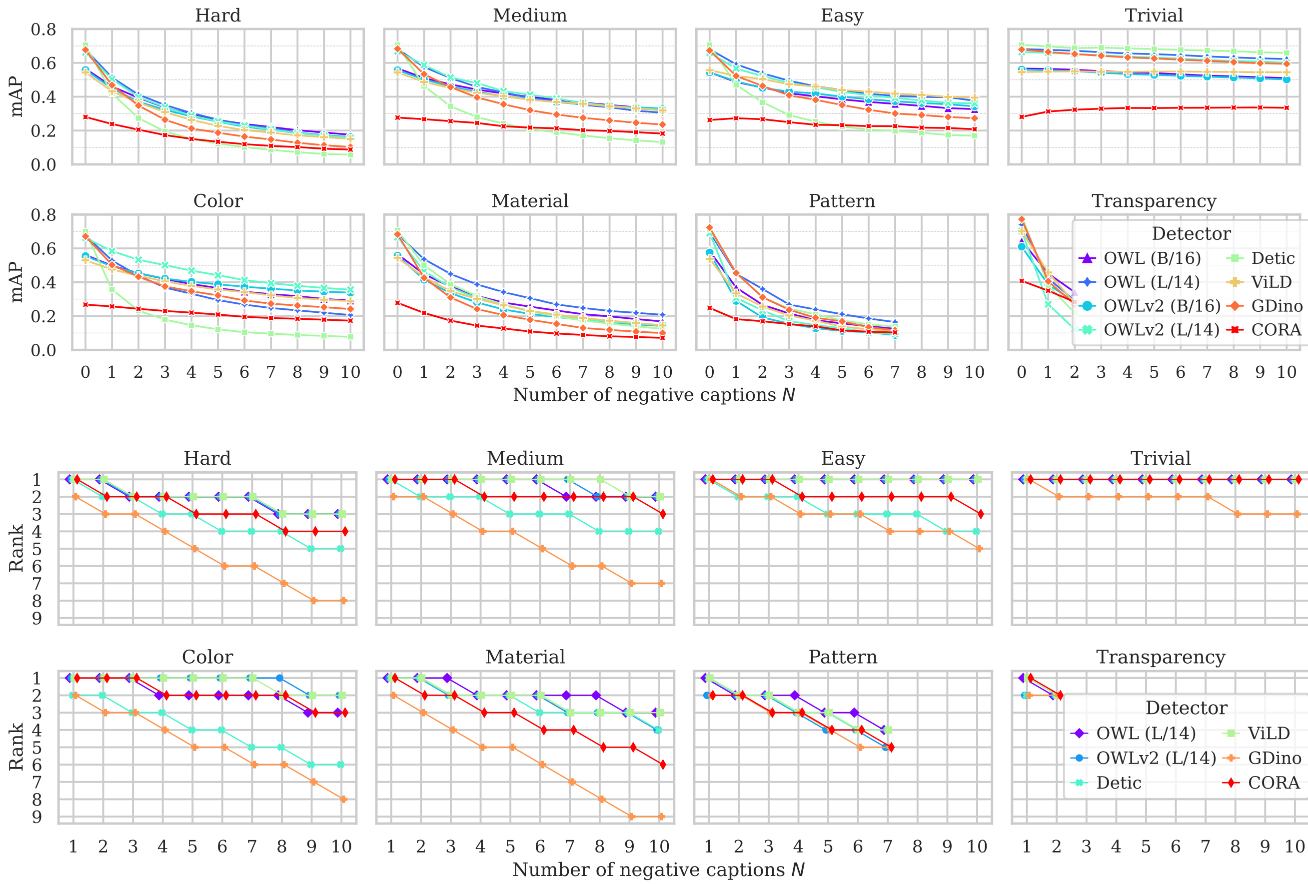


Figure 5. **Benchmark based on OWL-compatible captions:** Effect of the number of negative captions.

| | Hard | Medium | Easy | Trivial |
|---|---|---|---|---|
| OWL (B/16) | 29.5 | 37.2 | 42.8 | 57.6 |
| OWL (L/14) | 29.9 | 37.2 | **45.5** | 69.6 |
| OWLv2 (B/16) | 28.3 | 32.3 | 35.7 | 52.5 |
| OWLv2 (L/14) | **30.2** | 36.4 | 43.0 | 64.4 |
| Detic | 11.0 | 19.1 | 27.6 | **75.5** |
| ViLD | 26.6 | **40.5** | 41.1 | 60.2 |
| GDino | 27.8 | 38.1 | 44.2 | 71.9 |
| CORA | 22.0 | 28.8 | 32.4 | 44.3 |

Table 6. **Benchmark based on PACO-provided captions:** mAP on Difficulty-based benchmarks ($N = 5$)

| | Color | Material | Pattern | Transp. |
|---|---|---|---|---|
| OWL (B/16) | 48.2 | 37.6 | 28.6 | 24.7 |
| OWL (L/14) | 47.5 | 45.5 | **33.7** | 35.4 |
| OWLv2 (B/16) | 43.8 | 36.4 | 24.5 | 22.9 |
| OWLv2 (L/14) | **48.9** | 42.1 | 26.8 | 27.1 |
| Detic | 20.7 | **45.7** | 29.5 | **38.2** |
| ViLD | 47.4 | 36.3 | 25.8 | 28.1 |
| GDino | 48.2 | 37.6 | 31.4 | 28.8 |
| CORA | 32.5 | 29.2 | 24.0 | 32.5 |

Table 7. **Benchmark based on PACO-provided captions:** mAP on Attribute-based benchmarks ($N = 2$)

Table 8. **Benchmark based on PACO-provided captions:** Statistics for each different negative set comprising the number of images (Imgs), the number of annotated objects (Objs), objects-to-image ratio (Objs/Img), positive captions, positive captions per image, negative captions per positive caption, and objects per positive caption.

| Name | Negative Set Strategy | Imgs | Objs | Obj/Img | ✓Caps | ✓/Img | ✗/✓ | Objs/✓ |
|---|---|---|---|---|---|---|---|---|
| Hard | Random attribute subst. $(\times 1)$ | 1058 | 1326 | 1.3 | 1111 | 1.1 | 9.9 | 1.2 |
| Normal | Random attribute subst. $(\times 2)$ | 619 | 825 | 1.3 | 632 | 1.0 | 10.0 | 1.3 |
| Easy | Random attribute subst. $(\times 3)$ | 180 | 234 | 1.3 | 181 | 1.0 | 10.0 | 1.3 |
| Trivial | Random captions | 1058 | 1326 | 1.3 | 1111 | 1.1 | 10.0 | 1.2 |
| Color | Color attribute subst. | 901 | 1098 | 1.2 | 934 | 1.0 | 10.0 | 1.2 |
| Material | Material attribute subst. | 464 | 615 | 1.3 | 476 | 1.0 | 10.0 | 1.3 |
| Transparency | Transparency attribute subst. | 90 | 113 | 1.3 | 90 | 1.0 | 2.1 | 1.3 |
| Pattern | Pattern attribute subst. | 224 | 301 | 1.3 | 225 | 1.0 | 8.0 | 1.3 |

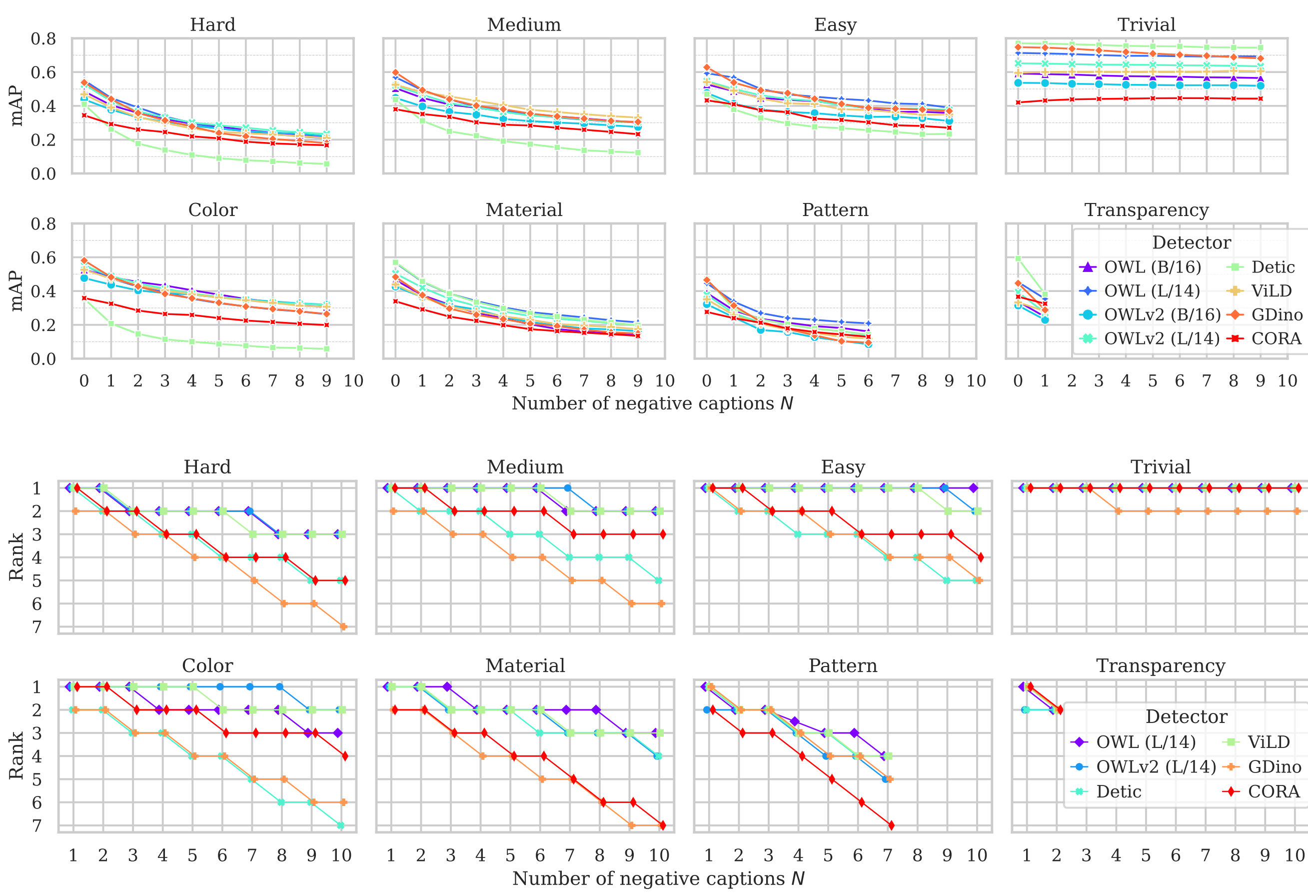


Figure 6. **Benchmark based on PACO-provided captions:** Effect of the number of negative captions.

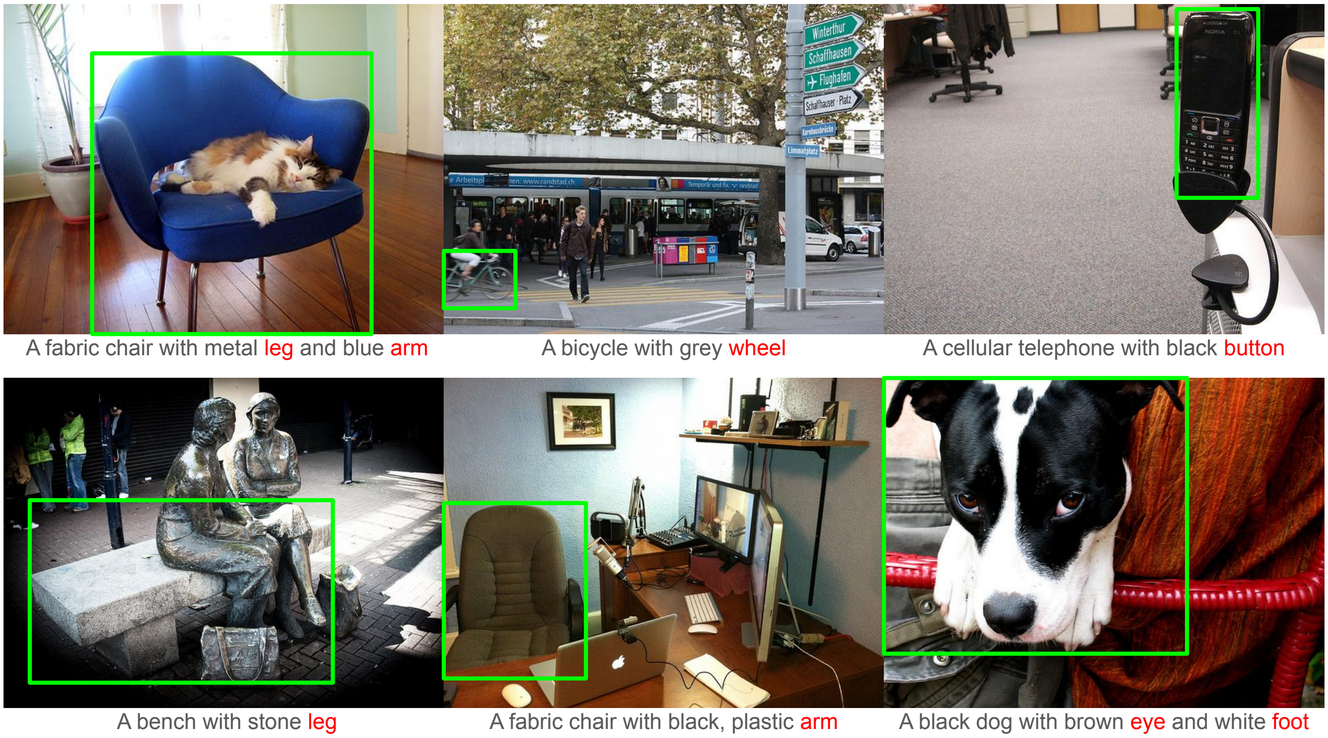


Figure 7. **Benchmark based on PACO-provided captions:** Samples where PACO captions do not include the plural form for parts. With its integrated common sense, a Large Language Model effectively addresses this by intuitively determining when pluralization is needed, resulting in sentences that feel more naturally structured.

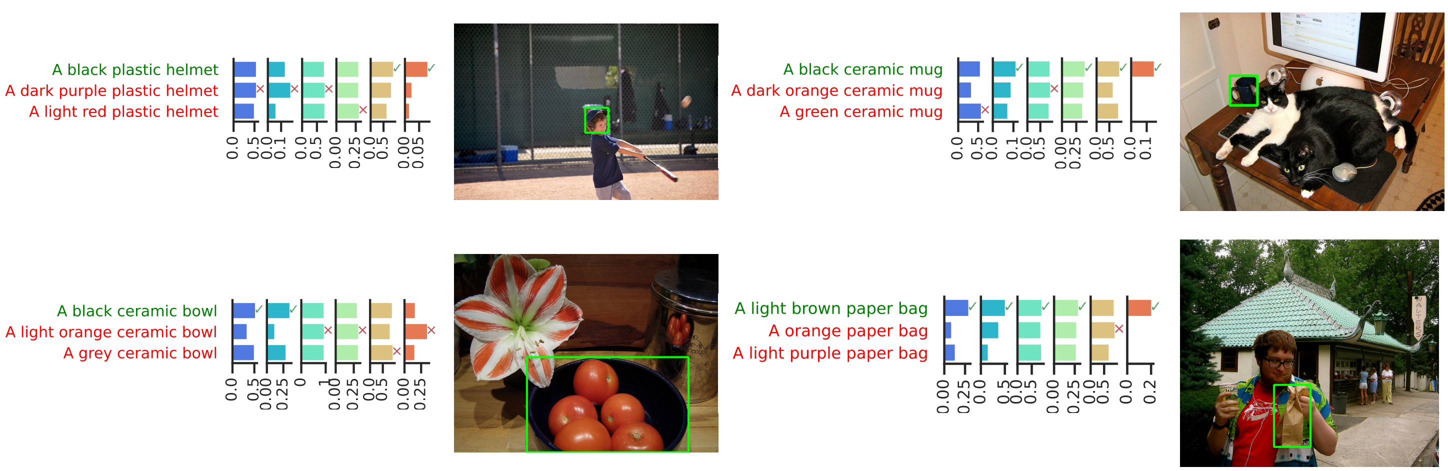


Figure 8. More samples from the Color benchmark. Legend: OWL (L/14), OWLv2 (L/14), Detic, ViLD, GDino, Cora

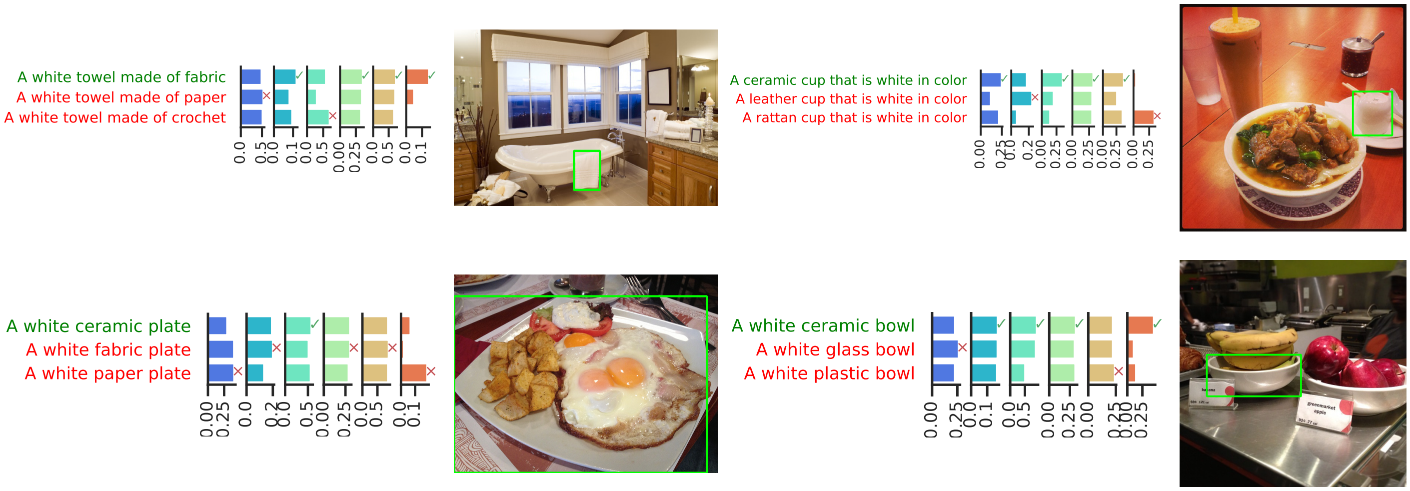


Figure 9. More samples from the Material benchmark. Legend: OWL (L/14), OWLv2 (L/14), Detic, ViLD, GDino, Cora

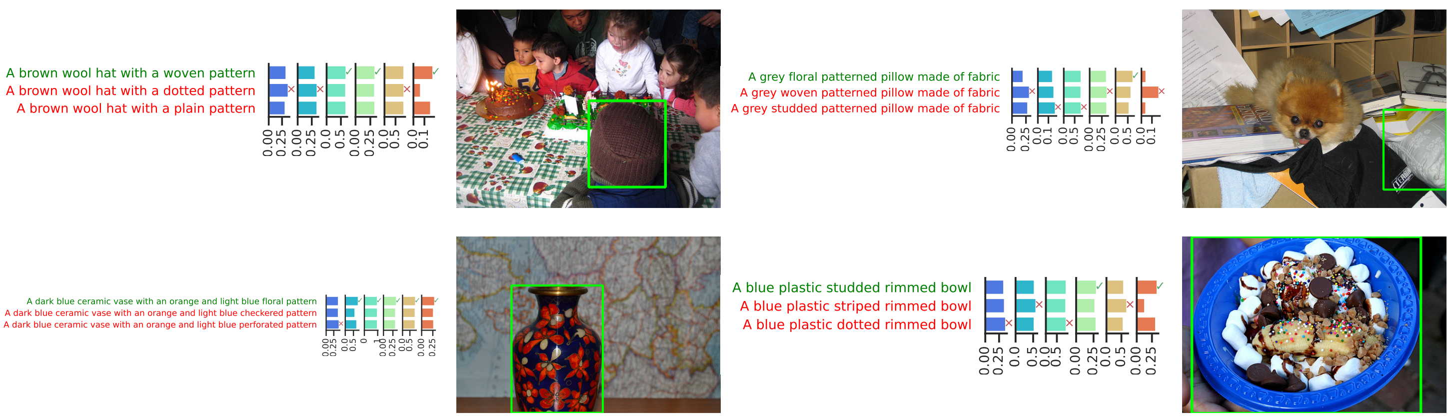


Figure 10. More samples from the Pattern benchmark. Legend: OWL (L/14), OWLv2 (L/14), Detic, ViLD, GDino, Cora

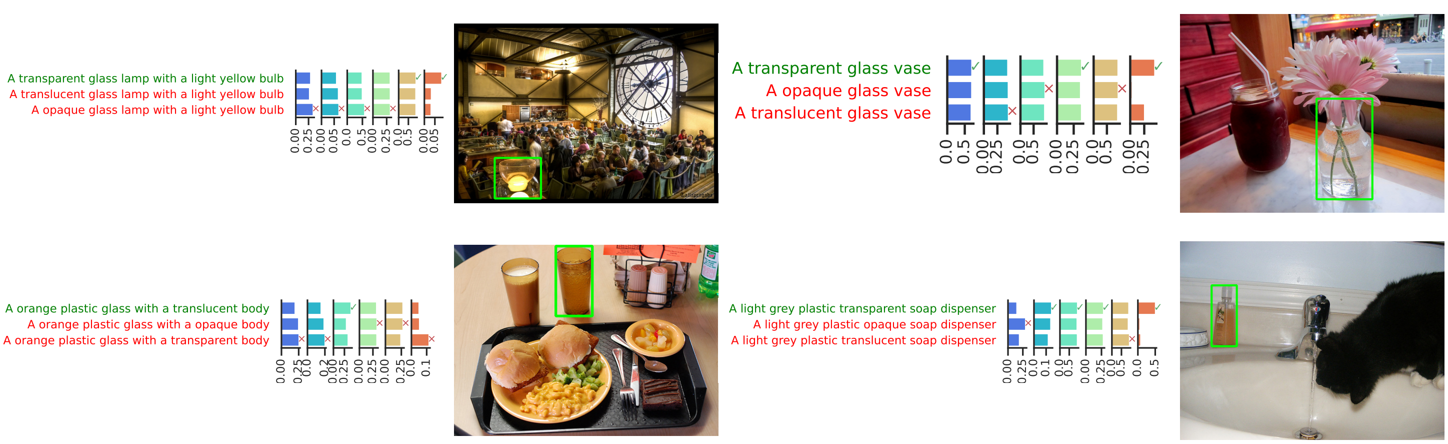


Figure 11. More samples from the Transparency benchmark. Legend: OWL (L/14), OWLv2 (L/14), Detic, ViLD, GDino, Cora


Figure 12. More samples from the Trivial benchmark. Legend: OWL (L/14), OWLv2 (L/14), Detic, ViLD, GDino, Cora

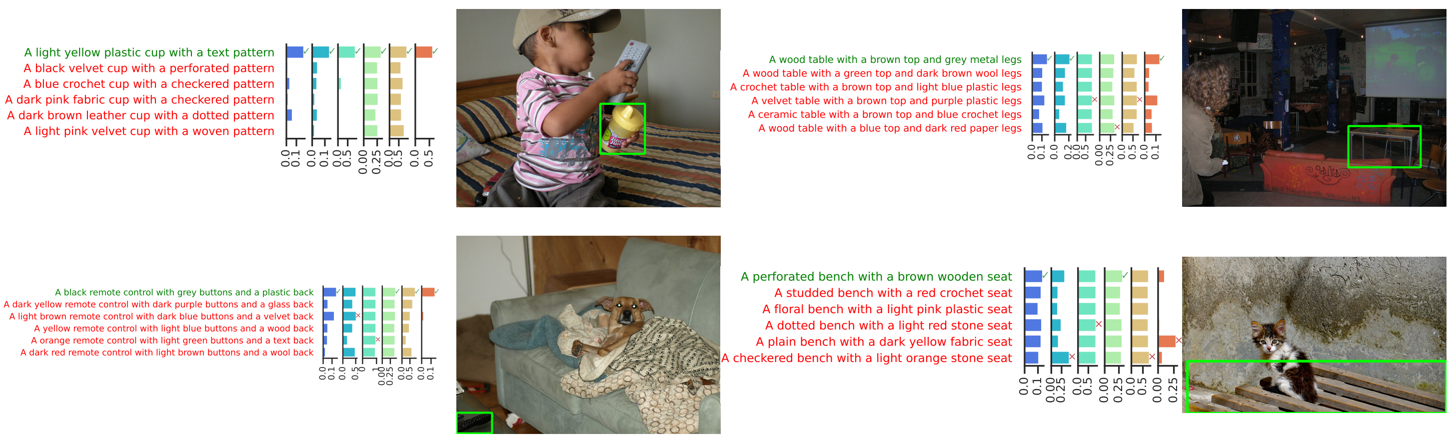


Figure 13. More samples from the Easy benchmark. Legend: OWL (L/14), OWLv2 (L/14), Detic, ViLD, GDino, Cora

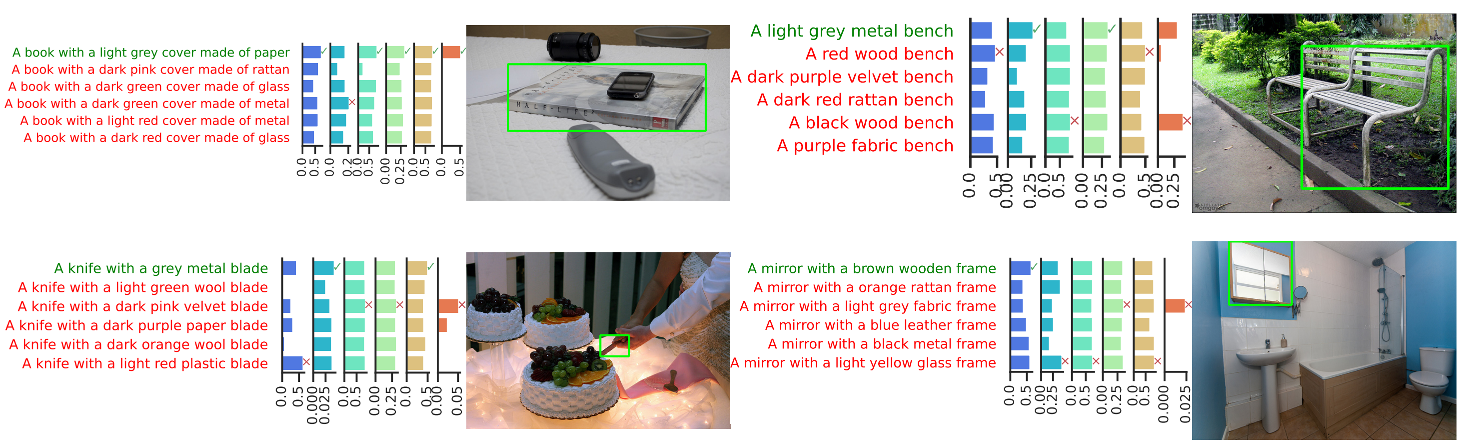


Figure 14. More samples from the Medium benchmark. Legend: OWL (L/14), OWLv2 (L/14), Detic, ViLD, GDino, Cora

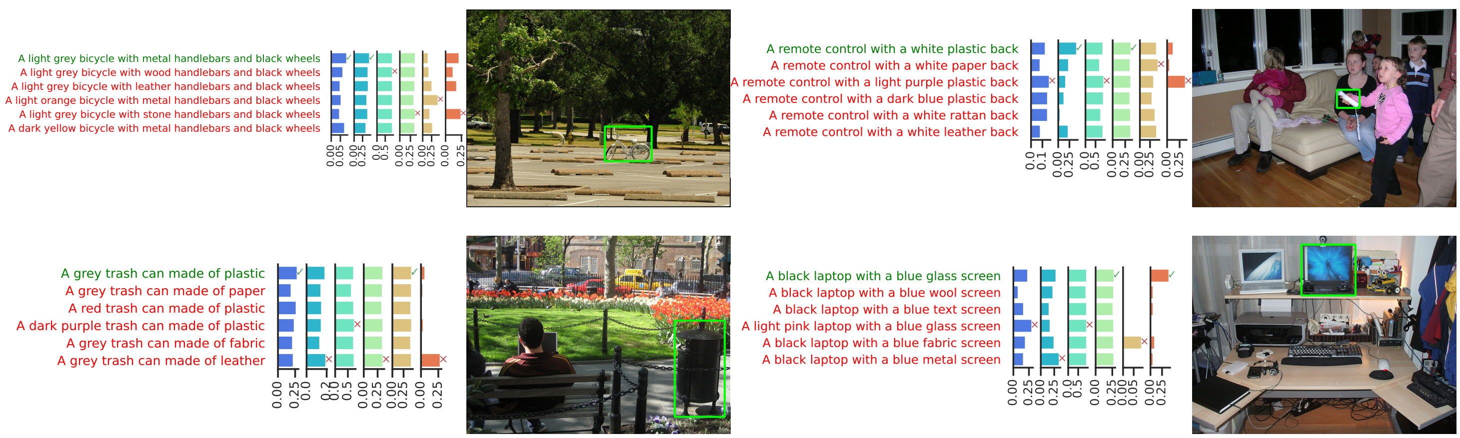


Figure 15. More samples from the Hard benchmark. Legend: OWL (L/14), OWLv2 (L/14), Detic, ViLD, GDino, Cora