Investigating Compositional Challenges in Vision-Language Models for Visual Grounding

Supplementary Material

7. ARPGrounding dataset

We quantify the frequencies of the predominant attributes depicted in Figure 7. The frequencies of the most prevalent relations are depicted in Figure 8.

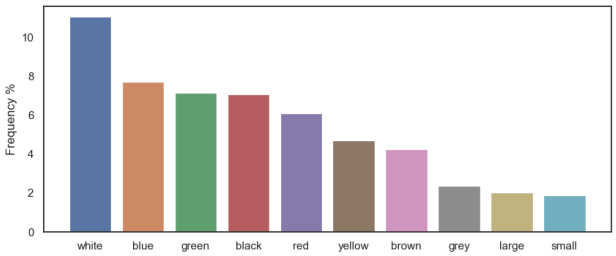


Figure 7. Attributes frequencies on the ARPGrounding dataset.

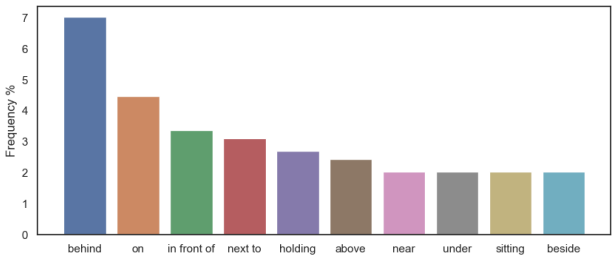


Figure 8. Relation frequencies on the ARPGrounding dataset.

Furthermore, we provide more examples of our ARPGrounding dataset. In Figure 9, we show examples of attribute compositionality. In Figure 10, we show examples of relation compositionality. In Figure 11, we show examples of priority compositionality.

8. Gradient-based Localization Details

In this section, we describe how we utilize explainability techniques to analyze the VLMs and achieve relatively good performance as discussed in Section 3.3. We individually analyze four models: CLIP, ALBEF, METER, and BLIP2. Due to the distinct architectures and training objectives of these models, our approach varies accordingly. We also conducted ablation experiments over the optimal layer as shown in Figure 12.

CLIP. For CLIP, we use ViT-B/32 variant released by OpenAI. CLIP belongs to two-stream architectures and uses separate encoders for image and text respectively. As for the intermediate attention map, we use the self-attention of the last layer of the image encoder. We use the similarity between the normalized image and text features as the objective function and compute its gradient with respect to the attention map. Since each such map is comprised of h heads, we average across heads after multiplying the attention map with its gradient. Then we get a heatmap of size 50×50, which is a symmetric matrix. We use the attention between 49 image patches and the CLS token and then reshape it to a size of 7×7.

ALBEF. We use the ALBEF variant that is fine-tuned with image-text retrieval task on COCO. ALBEF uses a 12-layer visual transformer ViT-B/16 as the image encoder and a 6-layer transformer for both the text encoder and the multimodal encoder. We use the cross-attention attention map of the third multimodel encoder layer. The gradient of the attention map is acquired by maximizing the image-text matching score. We use the gradient to weight the attention map as the same as CLIP. To adapt to different text lengths, we filter the attention map according to valid input text tokens and average across these tokens. Ultimately, we get a grounding heatmap of size 24×24.

METER. We use METER-CLIP16-RoBERTa fine-tuned on COCO IR/TR. There is one pre-trained image encoder and one pre-trained text encoder in the bottom part. On top of each encoder, there are six transformer encoding layers, with each consisting of one self-attention block, one cross-attention block, and one feed-forward block. The self-attention attention map of the fourth layer of the top image transformer encoder is utilized. The objective function is defined by the image-text matching score of the image-text pair, and we compute the gradient of the attention map. Similar to CLIP, we perform element-wise multiplication of the attention map with its gradient and subsequently average across heads. The input image is of size 384×384, with a patch size of 16×16, and the grounding heatmap is sized 24×24.

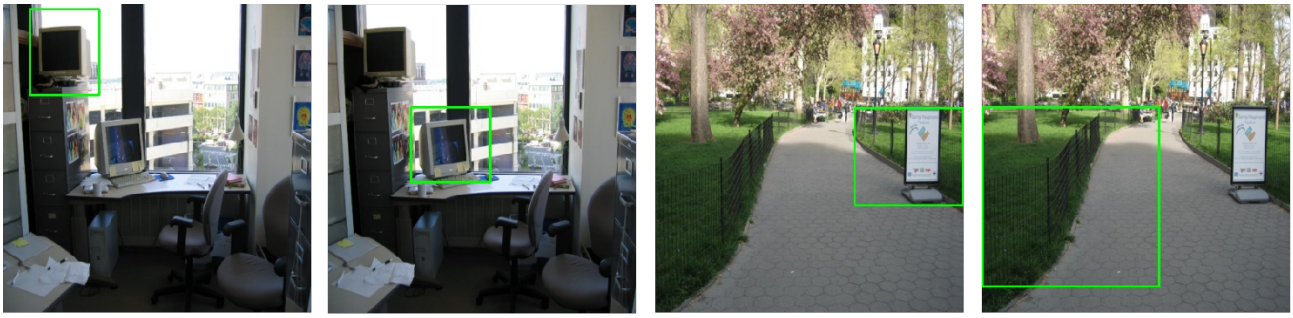
BLIP2. We use the ViT-G/14 variant of BLIP2 that is fine-tuned on COCO with image-text contrastive, image-text matching, and image-text generation objectives. BLIP2 comprises an image encoder and Q-Former which consists of two transformer submodules, one for visual feature extraction and another that can function as both a text encoder and a text decoder. The visual submodule encodes 32 queries and interacts with the frozen image encoder with cross-attention. We use the cross-attention attention map of the sixth layer of the visual submodule. We acquire the gradient of the attention map by maximizing the image-text

Figure 9. Attribute examples from our ARPGrounding dataset.

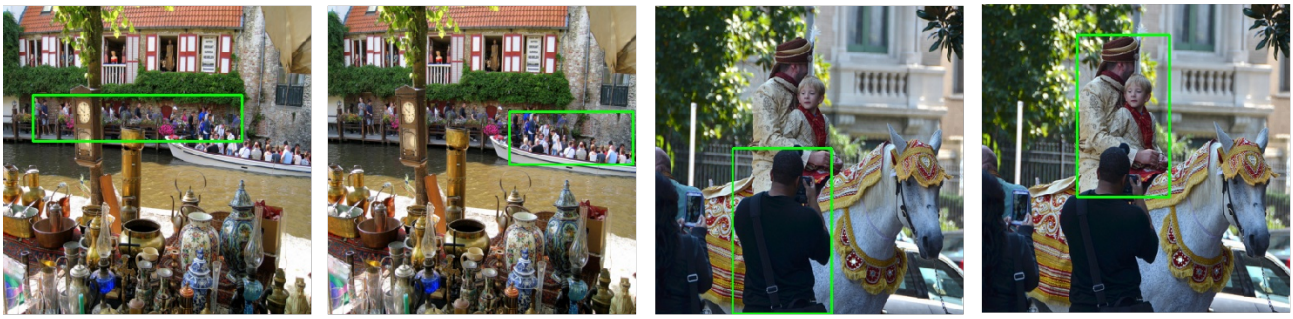
matching score. We use the gradient to weight the attention map and average across all attention heads and all query tokens. The input image is of size 364×364, with a patch size of 14×14, and the grounding heatmap is sized 26×26.

9. VLMs Grounding Visualization

In Figure 13, we present more visualizations of VLMs performing visual grounding of different attributes. It suggests that models perform better in distinguishing distractors of color but fail at reasoning about size. In Figure 14, we show visualizations to demonstrate the ability of models to understand relation and priority. It suggests that models encounter greater challenges in reasoning about relation and priority compositionality compared to attributes.



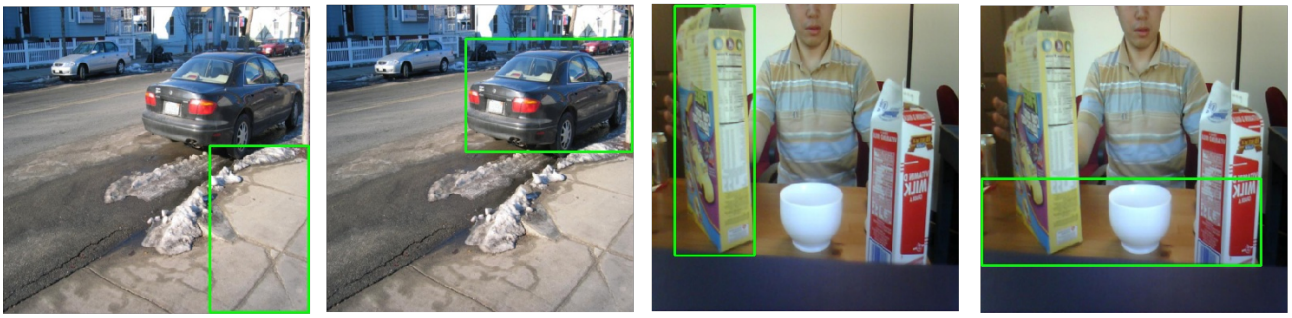
monitor above cabinet | monitor next to cabinet | fence to right of path | fence to left of path



people line to get boat | people getting on boat | man holding camera | man looking away from camera

Figure 10. Relation examples from our ARPGrounding dataset.

crosswalk in front of man | man at crosswalk | chair by table | table next to chair



sidewalk near car | car parked by sidewalk | cereal box on top of table | table has cereal box

Figure 11. Priority examples from our ARPGrounding dataset.

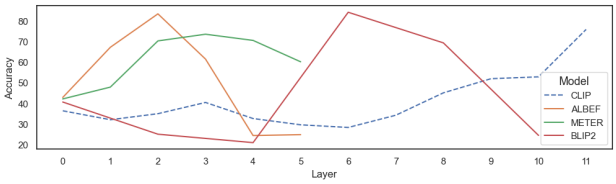


Figure 12. Grounding accuracy on the Flickr30k entities val.

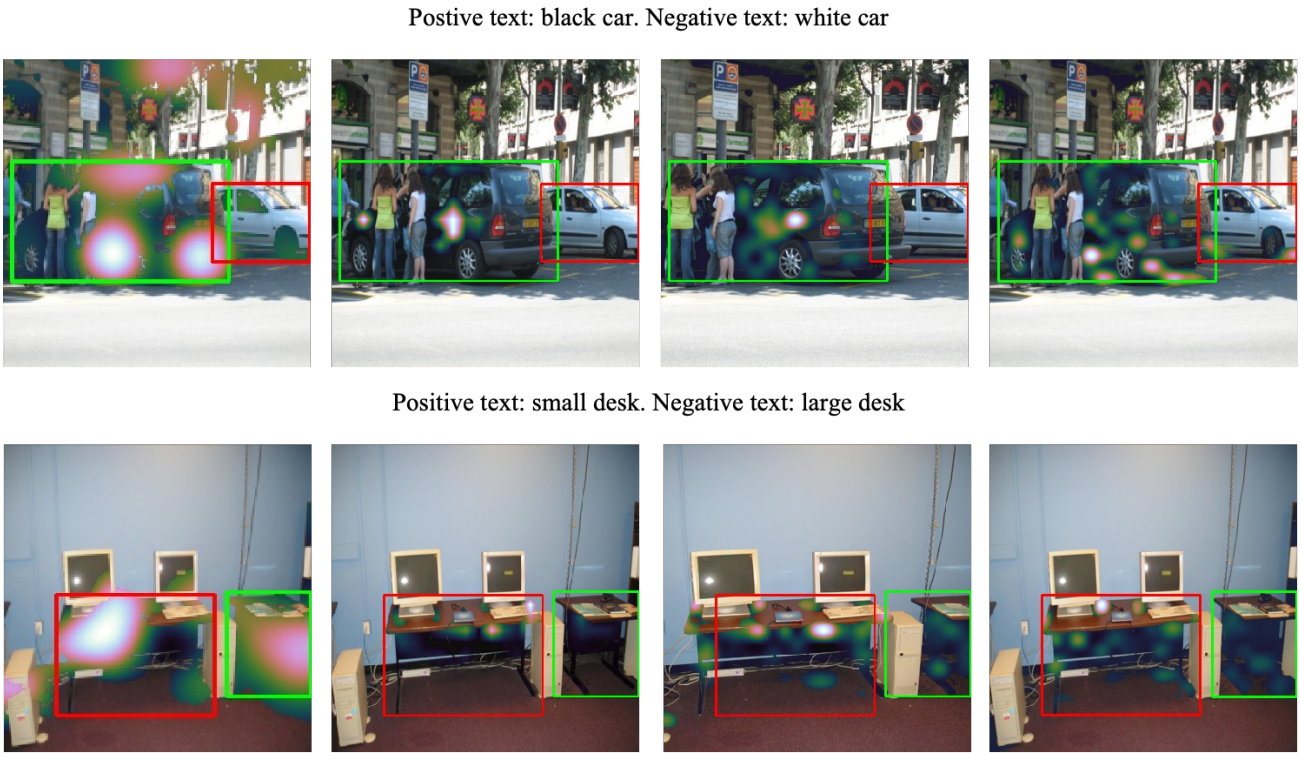
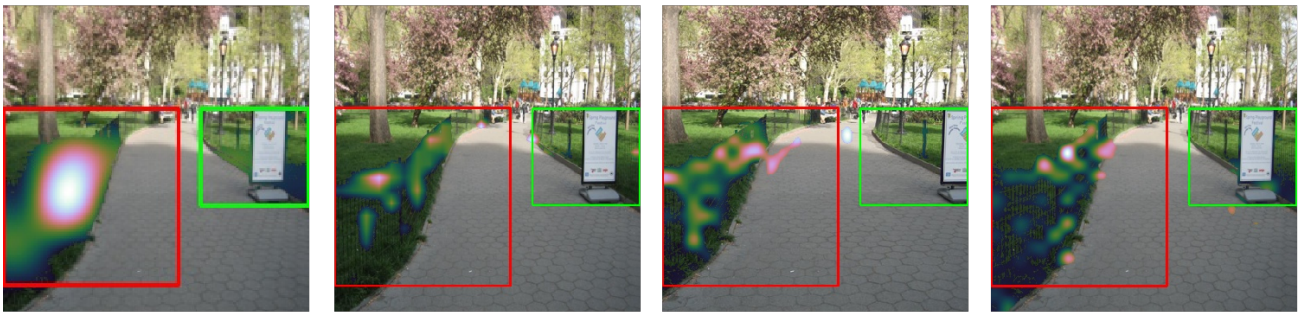


Figure 13. Visualization of grounding heatmap of VLMs on attribute samples.

Positive text: fence to right of path. Negative text: fence to left of path



Positive text: crosswalk in front of man. Negative text: man at crosswalk

CLIP ALBEF METER BLIP2

Figure 14. Visualization of grounding heatmap of VLMs on relation and priority samples.