SUPPLEMENTARY FOR MINIGPT-V2: LARGE LANGUAGE MODEL AS A UNIFIED INTERFACE FOR VISION-LANGUAGE MULTI-TASK LEARNING

Anonymous authors
Paper under double-blind review

In the supplementary, we provide more qualitative results that are generated from our model to demonstrate the vision-language multi-tasking capabilities.

1 INSTRUCTION TEMPLATE FOR VARIOUS VISION-LANGUAGE TASKS

RefCOCO/RefCOCO+/RefCOCOg: *[refer] give me the location of question*

VizWiz: *[vqa] Based on the image, respond to this question with a short answer: question and reply 'unanswerable' if you could not answer it*

Hateful Meme: *[vqa] This is an image with: question written on it. Is it hateful? Answer:*

VSR: *[vqa] Based on the image, is this statement true or false? question*

IconQA, GQA, OKVQA: *[vqa] Based on the image, respond to this question with a short answer: question*

2 ADDITIONAL QUALITATIVE RESULTS

To study how well our model is able to take visual input and answer questions based on task-oriented identifier, we use our model to perform multiple vision-language tasks including grounded image captioning in Fig. 1, Fig. 2, Fig. 3 and Fig. 4; Object parsing and grounding in Fig. 5, Fig. 6, Fig. 7 and Fig. 8; Referring expression comprehension in Fig. 9, Fig. 10, Fig. 11 and Fig. 12; Object identification in Fig. 13, Fig. 14, Fig. 15 and Fig. 16.

For each task, we share 4 examples for showing the vision-language capabilities of our model. The results in the demo provide direct evidence for the competing visual understanding capabilities of MiniGPT-v2 on multiple vision-language tasks. For example, in the cases of grounded caption, our model is able to give correct grounded image caption with detailed spatial locations of objects. In the cases of identify, the model also generates our expected object names. MiniGPT-v2 can understand the new scenes and follow the question identifier to respond. But we also need to note that our model still has some hallucination e.g., In Fig. 3, several persons are not grounded accurately, and in Fig. 4, there does not exist a vase in the image.

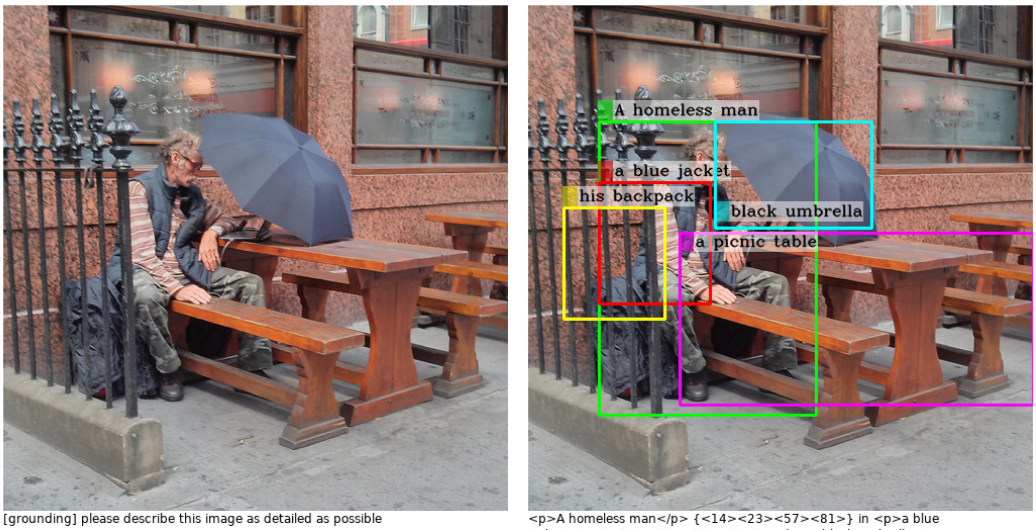

Figure 1: Detail grounded image caption example.

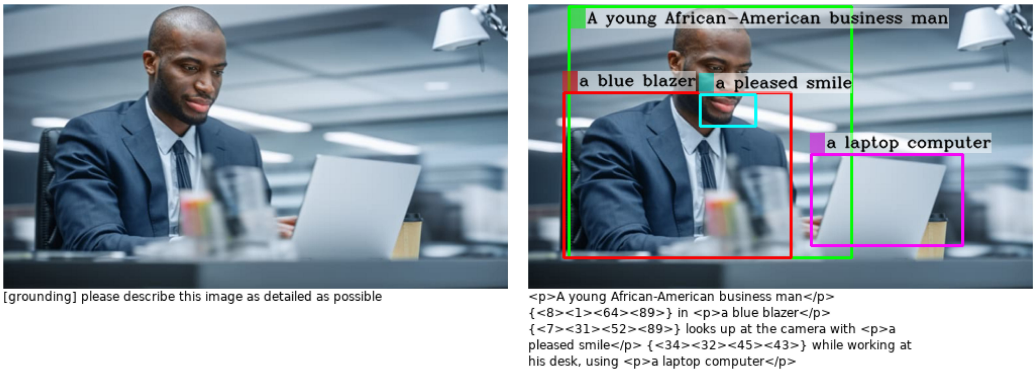

Figure 2: Detail grounded image caption example

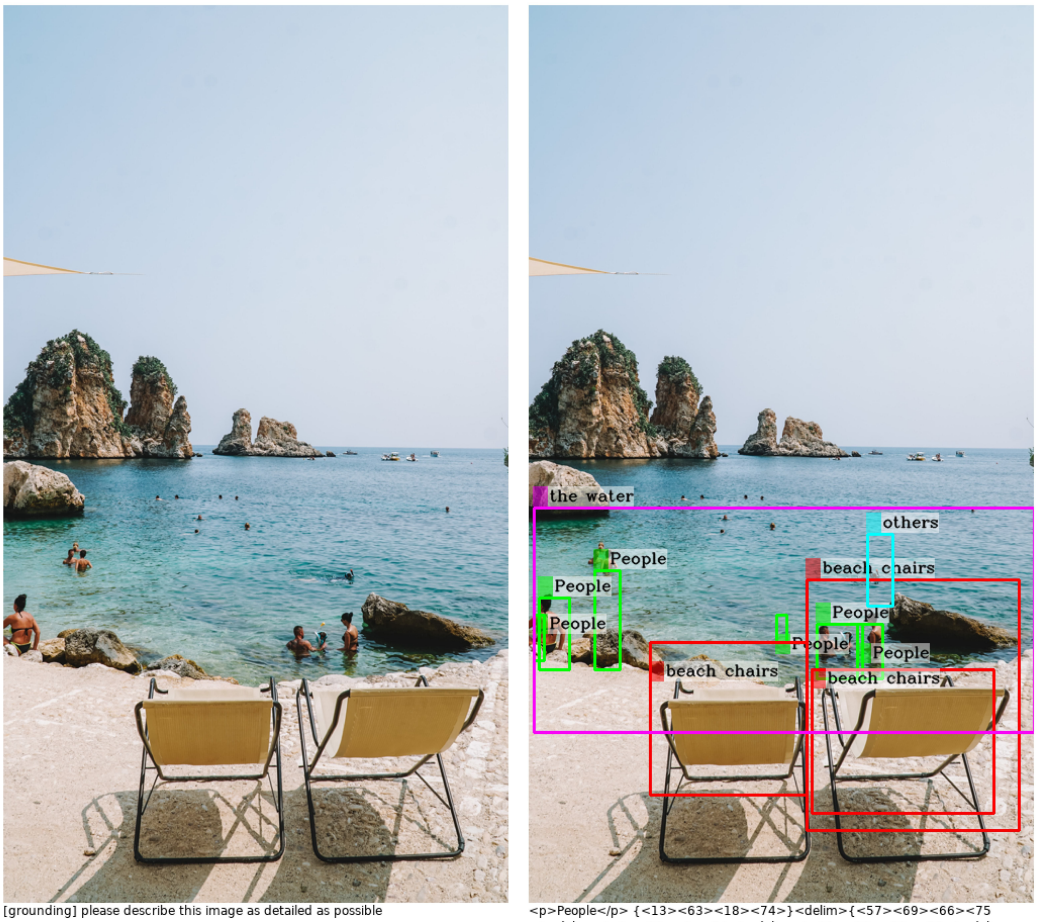

Figure 3: Detail grounded image caption example

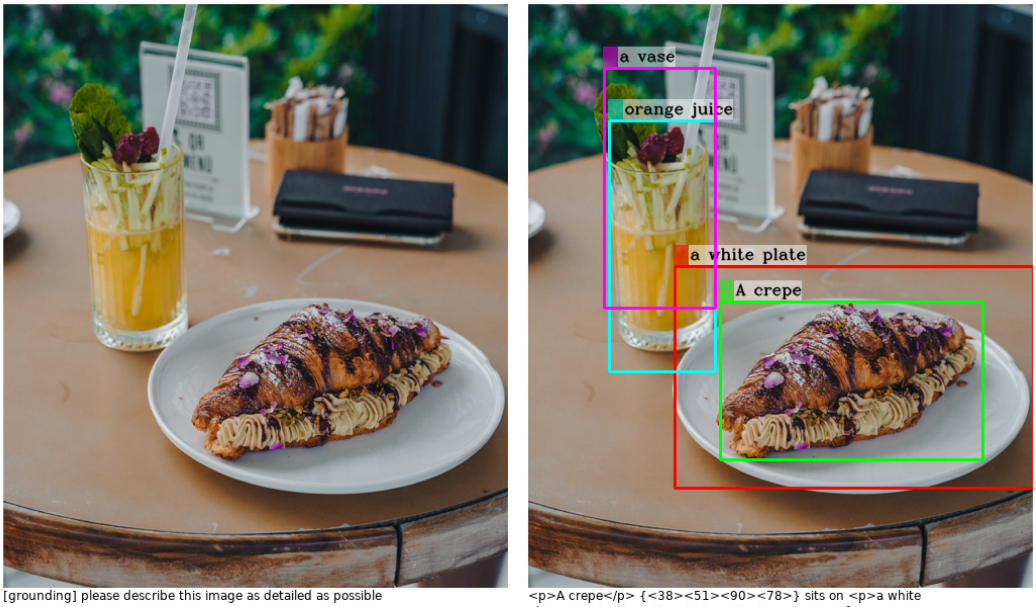

Figure 4: Detail grounded image caption example

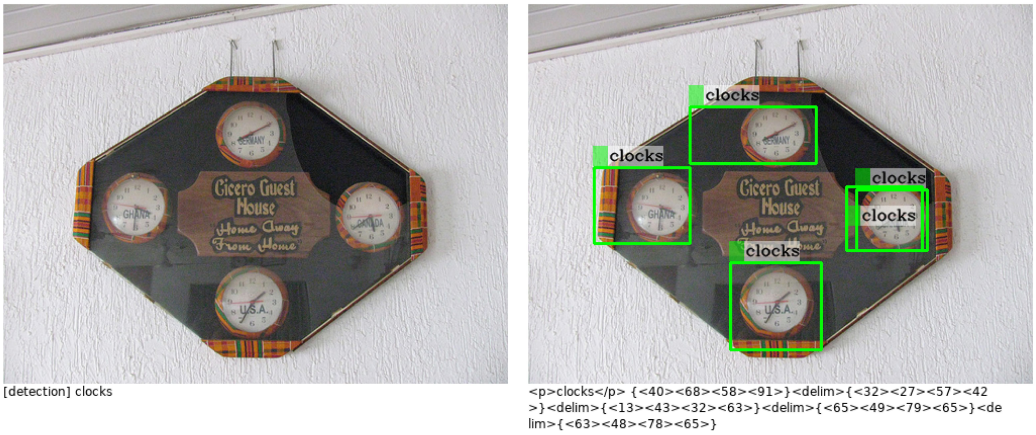

Figure 5: Object parsing and grounding example

Figure 6: Object parsing and grounding example

Figure 7: Object parsing and grounding example

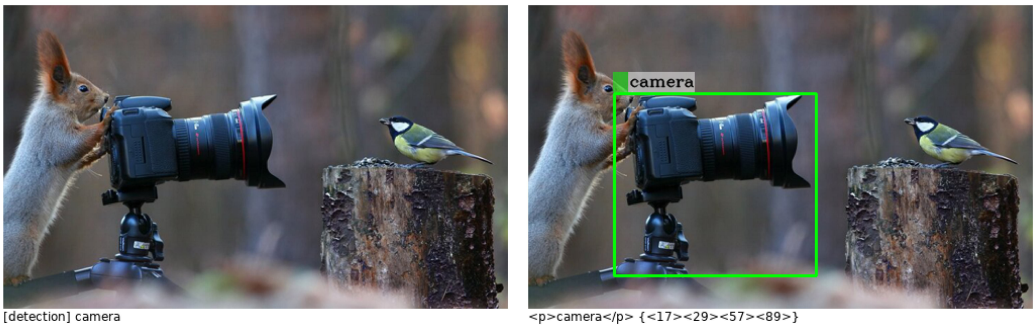

Figure 8: Object parsing and grounding example

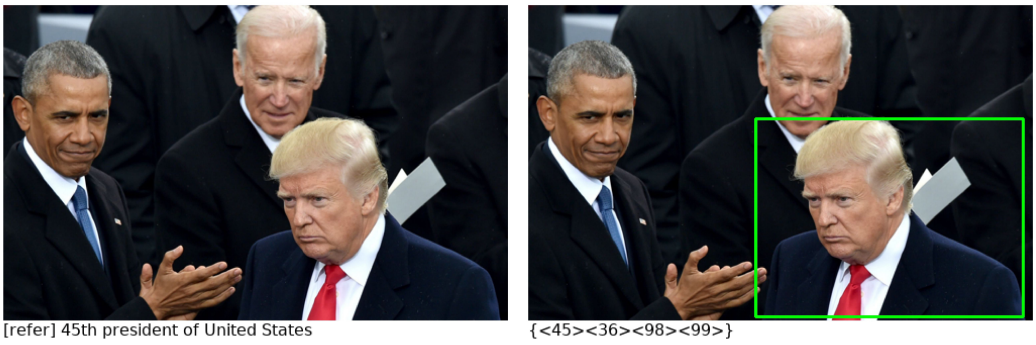

Figure 9: Referring expression comprehension example

Figure 10: Referring expression comprehension example

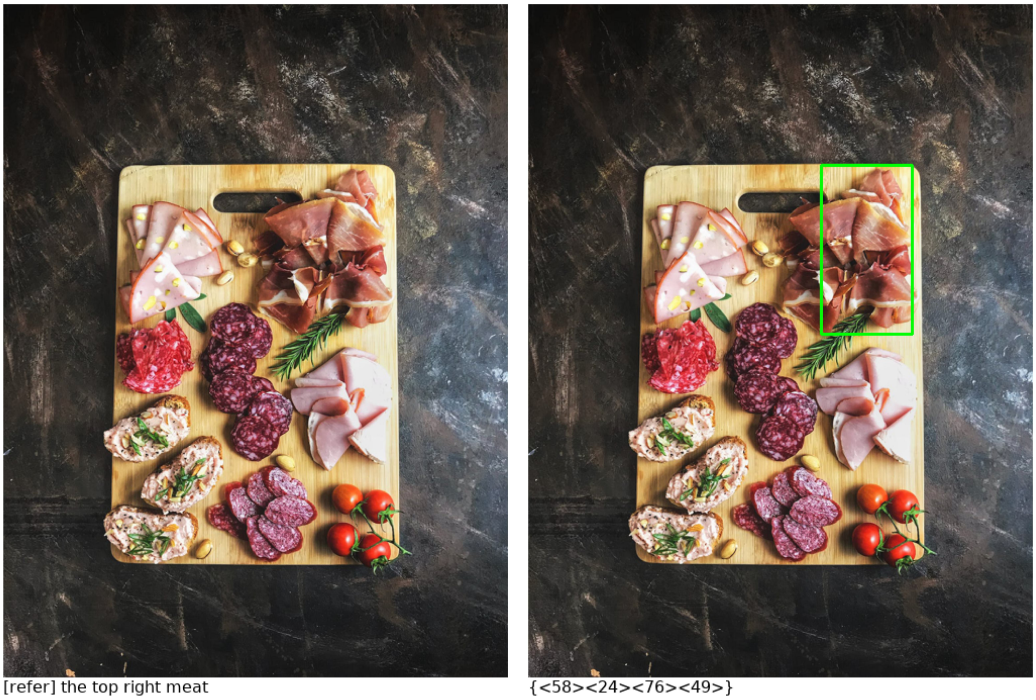

Figure 11: Referring expression comprehension example

Figure 12: Referring expression comprehension example

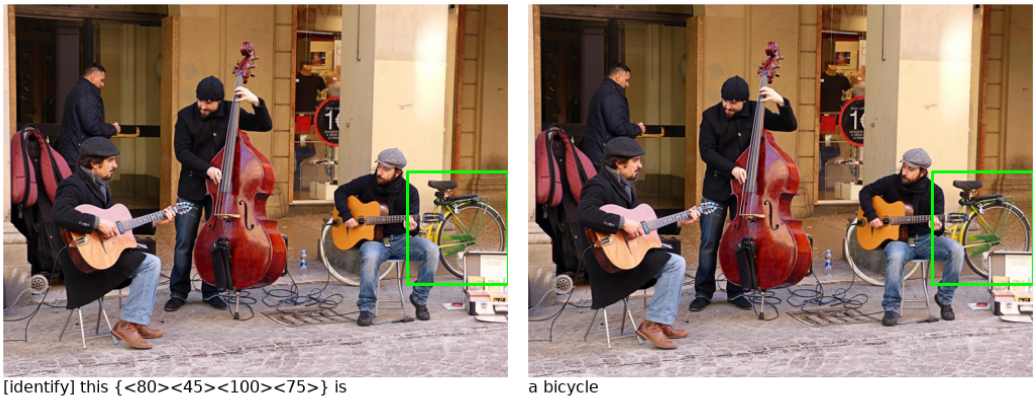

Figure 13: object identification example

Figure 14: object identification example

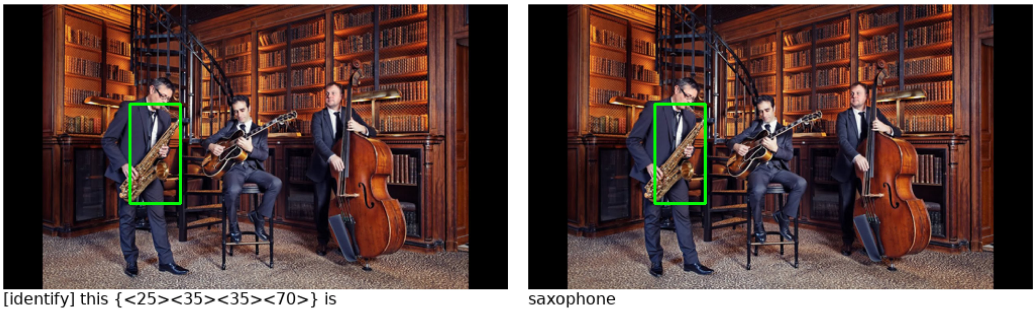

Figure 15: object identification example

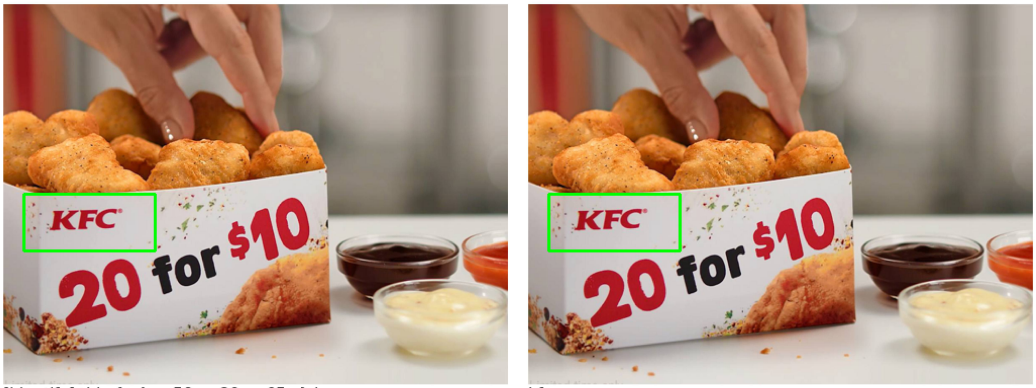

Figure 16: object identification example