SUPPLEMENTARY FOR MINIGPT-V2: LARGE LAN-GUAGE 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.





A homeless man {<14><23><57><81>} in a blue jacket {<14><35><36><59>} and black umbrella {<37><23><68><44>} sits at a picnic table {<30<<45><100<79>} while his backpack {<7><40><27><62>} is beside him .

Figure 1: Detail grounded image caption example.



[grounding] please describe this image as detailed as possible



A young African-American business man $\{<8><1><64><89>\}$ in a blue blazer $\{<7><31><52><89>\}$ looks up at the camera with a pleased smile $\{>4<3><43>$ while working at his desk, using a laptop computer

Figure 2: Detail grounded image caption example



 $\label{eq:cp} People {<13><63><18><74>}<delim>{<57><69><66><75 \\><delim>{<2><66><8><74>}<delim>{<65><69><70><75>}<delim>{<2><66><8><74>}<delim>{<65><69><70><75>}<delim>{<1><60><3<<73>}<delim>{<49><68><51><71>} sit and lay$ $on beach chairs {<55><64><97><92><delim>{<24<<7$ $1><55><88>}<delim>{<2<<92><90>} while \end{tabular}$

Figure 3: Detail grounded image caption example



[grounding] please describe this image as detailed as possible

A crepe {<38><51><90><78>} sits on a white plate {<29><45><100<<83>} near a cup of orange juice {<16><20><37><63>} near a cup of orange (<15><11><37><52>} with a stem in the background .

Figure 4: Detail grounded image caption example







clocks {<40><68><58><91>}<delim>{<32><27><57><42
>}<delim>{<13><43><32><63>}<delim>{<65><49><79><65>}<de
lim>{<63><48><78><65>}

Figure 5: Object parsing and grounding example



[detection] A woman in a colorful dress is sitting in a plane and typing on a calculator

< woman</p> {<9><10><100><100>} a colorful dress {<12><44><100><100>} a plane {<0><0><100><98>} a calculator {<7><74><53><87>}





sofas {<28><43><41><72>}<delim>{<69><50><98><73> }<delim>{<47><56><59><70>}

Figure 7: Object parsing and grounding example



[detection] camera

camera {<17><29><57><89>}

Figure 8: Object parsing and grounding example



[refer] 45th president of United States



{<45><36><98><99>}

Figure 9: Referring expression comprehension example



Figure 10: Referring expression comprehension example



Figure 11: Referring expression comprehension example



Figure 12: Referring expression comprehension example



[identify] this {<80><45><100><75>} is

a bicycle



Figure 13: object identification example



[identify] this {<25><35><35><70>} is

Figure 14: object identification example



saxophone

Figure 15: object identification example



Figure 16: object identification example