CVPDL HW3

B09901142 EE4 呂睿超

December 15, 2024

Stage 1 - Image Captioning

1. Adopted method - as required, I used BLIP-2 to do image to text captioning.
2. Model selection
   (a) Because of VRAM limit: I only chose to try these two models: blip2-opt-2.7b and blip2-flan-t5-xl
   (b) I only try to visualize some samples of these two results, and they are similar.
   (c) stage 1 model final selection: blip2-opt2.7b
3. Prompt Template: I followed the guidance of the slides and attempted the same template.
   (a) generated_text: direct output
   (b) prompt_w_label: direct output + all the ground truth labels
   (c) prompt_w_suffix: prompt_w_label + HD suffix

Stage 2 - GLIGEN Text+Box to Image

1. For my last result, I simply selected the (text+layout)-to-image GLIGEN method.
2. various tried methods
   (a) I tried to use guidance image of every object type for each box. But it turns out to be too limited that the model tries to squeeze the exact guidance image to each of the corresponding gligen box.
   (b) I tried to use simply the original image as the guidance image for reference, but it ruins the generation.
3. Prompt Selection : generated_text - I tried to use prompt_w_label but it seems to be conflicting with the gligen_phrases that I fed into the model

FID Result

	Text grounding	**Text grounding**	**Layout to image**
prompt	generated_text	prompt_w_label	generated_text
FID	63	67	58

Table 1: FID Results