R13944045_CVPDL_HW3

Select model:

In the stage of selecting the prompt word model, I compared about a quarter of the images generated by **BLIP2-OPT-2.7B** and **BLIP2-OPT-6.7B,** respectively. By observing the prompts and images generated. I found that the performance of 6.7b is indeed excellent, but although the 6.7b model performs slightly better in terms of descriptive accuracy and semantic understanding, the difference is not significant. After considering the required computing resources and generation speed, I chose to use the 2.7b model for generation to strike a balance between performance and computational efficiency.

Prompt design:

I used three different prompting strategies. The first method focuses on generating basic scene descriptions using the 2.7b model. The second strategy integrates tag enhanced description generation, integrating structured tag information from JSON metadata to improve semantic accuracy. The third approach utilizes positive and negative prompting techniques based on the second approach to better control the specificity of generated content.

However, regarding the third type of prompt words, as I did not consider the limited length of the prompt words, most of the positive and negative prompt words added later were truncated during generation, which also resulted in little difference in performance compared to the second method.

Image generation:

In this stage, I attempted to generate a portion of the images using both Realistic_Vision-V2.0 (RV2) and GLIGEN models. Although GLIGEN (around 45) usually achieves good results on FID, I believe RV2 (around 55) is noticeably more natural and harmonious in terms of visual perception than GLIGEN.

Future work:

The generated images demonstrate several limitations, including blurry facial features and objects not properly aligned within their bounding boxes. To address these detail rendering and spatial control issues, future work could explore the integration of ControlNet, which shows promise in enhancing both visual fidelity and layout accuracy of the generated results.

Table:

	Text grounding			
Prompt	generated_text	prompt_w_label	prompt_w_suffix	layout
FID	54.8	56.78	56.33	45.50