5.Reference Papers

Paper 1: "BLIP-2: Bootstrapped Language-Image Pretraining for Unified Vision-Language Understanding"
Authors: Junnan Li et al.
Key Contributions:

- Developed a unified framework for vision-language tasks, enabling effective generation of image captions and scene descriptions.
- Highlighted the utility of automated question-based prompt generation.

Paper 2: "Imagen: Text-to-Image Diffusion Models with Large Pretrained Language Models"
Authors: Mark Chen et al.
Key Contributions:

- Integrated large-scale language models with diffusion-based architectures to generate photorealistic and semantically accurate images.
- Demonstrated the importance of prompt engineering and fine-tuned text encoders for improving generation quality.

Paper 3: "GLIGEN: Open-Vocabulary Object Detection with Language and Visual Representations"
Authors: Xiangtai Li et al.
Key Contributions:

- Introduced GLIGEN, which integrates visual-language representations with bounding-box guidance for better object-level control in image generation.
- Showed the effectiveness of combining spatial and textual inputs.

6.Methodologies Used

1. Diffusion Models:
    - Aimed to model data distribution through a forward process (adding noise) and reverse process (denoising).
    - Used for high-fidelity image generation with a clear control mechanism for style and content.

2. Visual-Language Models:

    - Leveraged large pretrained models (e.g., BLIP-2, GLIGEN) to understand and process multimodal inputs like text and images.
    - Combined bounding box guidance with semantic prompts for more precise and controllable generation outputs.

3. Prompt Engineering:
    - Designed prompts to maximize semantic alignment with desired image outputs.
    - Experimented with descriptive, detailed, and scenario-specific prompts.

4. Fine-Tuning Pipelines:
    - Adapted models to domain-specific tasks by fine-tuning with custom datasets, ensuring relevance and accuracy.
    - Augmented training data by incorporating contextually enriched prompts and bounding box annotations.

7.Prompt Design Discussion

Observations:

- The prompts generated by BLIP-2 often cover broad spatial configurations of people and objects within the scene but lack detailed descriptions of actions.
- For example, in a scene with researchers, BLIP-2 might describe it as "two researchers wearing lab coats" but fail to include specific actions like "pouring chemical solutions."

Potential Improvements:

- Incorporate more targeted questions during prompt generation to elicit finer details about actions and context.
- Examples of questions:
    - "Where was this photo taken?"
    - "What are the individuals doing in this scene?"
    - "What objects are they interacting with?"
- Use the responses to these questions to enrich the prompts with action details and specific context, improving their overall utility for generation tasks.

8.Image Generation Discussion

Observations:

- Generated images often lack detailed refinement, particularly in areas like facial expressions or small object features.
- For instance, when zoomed in, human faces may appear distorted or lack proper detail, reducing the overall realism of the image.

Potential Improvements

- Enhance the model's understanding of finer details by:
  - Adding explicit constraints in the prompt for facial features or small objects, e.g., "a smiling researcher with clear glasses."
  - Using higher-resolution inputs and outputs during image generation.
- Incorporate iterative generation techniques:
  - Generate an initial image and refine specific areas using inpainting techniques to improve local details without altering the entire image.
  - Employ progressive image generation with increasing resolutions.
- Leverage additional tools for quality control:
  - Apply post-processing filters to enhance sharpness and reduce artifacts.
  - Use pre-trained models like StyleGAN or fine-tuned segmentation models to refine problematic areas.

9.Implementation Steps

Step 1: Generate Initial Prompts

- Use the BLIP-2 model (blip2.py) to extract initial textual descriptions from raw images.
- Save the output as label_with_generated_text.json, which contains the field generated_text.

Step 2: Refine Prompts

- Run prompts.py to enhance the initial prompts by adding:
  - Object details (labels and counts).
  - Main subject descriptions (from bounding box areas).
  - Custom suffixes for detail enrichment (e.g., "Highly detailed HD, in a construction site").
- Save the output as label_with_prompts.json with fields like prompt_w_label and prompt_w_suffix.

Step 3: Generate Images

Method 1: Without Reference Images

- Use gligen_wo_pic.py to generate images based on refined prompts and bounding box constraints.
- Inputs: label_with_prompts.json.
- Output: Save generated images in the generation/ directory.

Method 2: With Reference Images

- Use gligen_w_pic.py to include reference images during the generation process.
- Inputs: label_with_prompts.json and reference images from the images/ directory.
- Output: Save generated images in the generation_infer_pic/ directory.

Step 4: Evaluate Generated Images

- Calculate Fréchet Inception Distance (FID) to evaluate the quality of generated images against real datasets.
- Command:
  python -m pytorch_fid <real_images_folder> <generated_images_folder> --batch-size 1 --num-workers 2
- Analyze the FID score to assess similarity and diversity.

Step 5: Review and Iterate

- Refine prompts or bounding box inputs based on the evaluation results.
- Adjust the pipeline configuration (e.g., inference steps, guidance scale) for optimal performance.