1. Image Captioning and Prompt Design

1.1 Comparing Pre-trained Models

Selected Models for Comparison

1. **Salesforce/blip2-opt-2.7b**
    - **Advantages**: Good balance between VRAM requirements and caption quality.
    - **Limitations**: Requires moderate VRAM (~12GB) for optimal performance.
2. **Salesforce/blip2-flan-t5-xl**
    - **Advantages**: Lower VRAM requirement (~8GB), faster inference speed.
    - **Limitations**: Produces less detailed and contextually accurate captions.

Evaluation Metric

- Captions generated by both models were evaluated based on:
    1. **Semantic Accuracy**: Does the caption correctly describe the main objects and scene context?
    2. **Detail Level**: Does the caption include detailed object descriptions and relationships?
    3. **Readability**: Is the caption coherent and easy to understand?

Results

Model	Semantic Accuracy	Detail Level	Readability	VRAM Usage
BLIP-2-opt-2.7b	Good	Moderate	Good	~12GB
BLIP-2-flan-t5-xl	Moderate	Low	Moderate	~8GB

Example Captions

Model	Generated Caption
BLIP-2-opt-2.7b	"A worker wearing safety gear at a construction site."
BLIP-2-flan-t5-xl	"A person standing outdoors."

Conclusion

- **BLIP-2-opt-2.7b** was selected for further tasks due to its superior performance in generating contextually accurate captions.

1.2 Prompt Design

Prompt Templates

Two templates were created to incorporate `label` information, including their counts and the main subject's position:

1. **Prompt Template 1: `prompt_w_label`**
    - **Structure**: Combines the generated caption with object labels, their counts.
2. **Prompt Template 2: `prompt_w_suffix`**
    - **Structure**: Extends `prompt_w_label` by adding 9-grid layout positions for the main subject and a descriptive suffix for more detail and context.

Generated Prompts

`prompt_w_label` Example	`prompt_w_suffix` Example
"Construction workers on a construction site. This photo contains 2 Heads, 1 Face, 2 Ears, 1 Hands, 2 Gloves, 1 Safety-vest, 4 Shoes, 2 Helmets, 1 Safety-suit, and 2 Persons."	"Construction workers on a construction site. The main subject appears in position 5 of the 9-grid layout. This photo contains 2 Heads, 1 Face, 2 Ears, 1 Hands, 2 Gloves, 1 Safety-vest, 4 Shoes, 2 Helmets, 1 Safety-suit, and 2 Persons. Highly detailed HD, in a construction site."

2. Text-to-Image Generation

2.1 Text-Only Generation

Three types of generated prompts were used:

1. **generated_text**: Text-only description with object details.
2. **prompt_w_label**: Text-only description with object details.
3. **prompt_w_suffix**: Text-only description with additional suffixes.

Results

Prompt Type	FID Score

generated_text	69.01
prompt_w_label	72.99
prompt_w_suffix	74.91

Analysis

- **generated_text:**
    - The simplest prompt type, containing only the basic caption generated by the BLIP-2 model.
    - FID is the lowest among the three, indicating that simpler prompts result in images that are more aligned with the real image distribution.
- **prompt_w_label:**
    - Adds detailed information about the labels and their counts to the generated_text.
    - The increase in FID suggests that the additional label details may have introduced challenges for the model in balancing visual fidelity and textual constraints.
- **prompt_w_suffix:**
    - Extends prompt_w_label by appending high-detail suffixes (e.g., "Highly detailed HD, in a construction site").
    - The highest FID score indicates that the increased complexity and specificity in the prompt may have led to overfitting or difficulty in generating realistic images.

Conclusion

- Generated Text performs the best in terms of FID, likely due to its simplicity and close alignment with the model's training data.
- Adding labels and high-detail suffixes increases the FID score, suggesting that more complex prompts may lead to divergence from the real image distribution.

2.2 Text Generation Referring to Image

Using the prompt from 2.1 Text-Only Generation, we refer to the image.

Prompt Type	FID Score
generated_text	31.32
prompt_w_label	31.23

prompt_w_suffix	31.19

Analysis

- **generated_text:**
    - The simplest prompt type, containing only the basic captions generated by BLIP-2.
    - Achieves a slightly higher FID score compared to other prompts, likely due to its lack of specificity.
- **prompt_w_label:**
    - Adds detailed information about the labels and their counts to the generated_text.
    - Results in a small improvement in FID, as the additional label information helps align the generated images more closely with the real data.
- **prompt_w_suffix:**
    - Extends prompt_w_label by appending high-detail suffixes (e.g., "Highly detailed HD, in a construction site").
    - Achieves the lowest FID score, indicating that the added context slightly improves the alignment between the generated and real images.

Conclusion

- The FID scores for all three prompt types are very close, indicating that the model performs similarly regardless of prompt complexity.
- prompt_w_suffix provides the best results, albeit with a marginal improvement, and may be preferred for tasks requiring the most realistic images.

3. FID Evaluation

Evaluation Process

- **Tool Used**: pytorch-fid
- **Command**: python -m pytorch_fid images generation_suffix --batch-size 1 --num-workers 2

Methods Used

1. **Text + Layout** (gligen_wo_pic.py)
    - Generated images using bounding box annotations and textual prompts.

2. **Text + Layout + Reference Image** (gligen_w_pic.py)
    - Included reference images to guide the generation process.

Results

Method	prompt	FID Score
Text + Layout	prompt_w_label	72.99
	prompt_w_suffix	74.91
Text + Layout + Reference Image	prompt_w_label	31.23
	prompt_w_suffix	31.19

Analysis

- **Text + Layout** generated accurate object placement but sometimes missed finer details.
- **Text + Layout + Reference Image** improved context consistency and produced more visually appealing images.

Conclusion

Including layout and reference images significantly improved FID scores, demonstrating better alignment with real-world data.

4. Conclusion

Key Findings

1. **Best Captioning Model**: blip2-opt-2.7b was the most effective for generating detailed and contextually accurate captions compared to blip2-flan-t5-xl.
2. **Best Prompt**: prompt_w_suffix produced visually detailed images when combined with reference images, as it leverages both the detailed textual descriptions and the guidance from reference images to produce visually detailed and semantically consistent images.
3. **Best Method**: Using **Text + Layout + Reference Image** yielded the best results in terms of image quality and context consistency.

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.