Lab Assignment Report: Image-to-Prompt-to-Image Workflow

1. Workflow Creation

Did you create this workflow entirely on your own?

- **No**
- **Brief Description of the Process:**
    - I used a third-party workflow that was downloaded from GitHub https://github.com/pharmapsychotic/comfy-cliption/tree/main. This workflow allowed for the efficient transformation of images into text prompts and subsequently re-generating the images based on those prompts. The workflow begins by loading a checkpoint model used for image generation, along with a CLIP Vision model to analyze the visual features of the input image. The input image is then loaded into the system for processing. The CLIP Text Encode node is employed to generate both positive and negative text prompts based on the visual content of the image. These prompts are then transformed into a descriptive text that represents the image, which is displayed by the Show Text node. Using the generated text, the system creates a latent image representation through the EmptySD3LatentImage node. This latent image is fed into the KSampler node, which uses the latent data and text prompts to generate a new image that aligns with the described scene. Finally, the resulting image is previewed and saved. This process transforms the visual content into a textual prompt and uses that prompt to regenerate a re-imagined image.

2. Missing Custom Nodes

Did you encounter any missing custom nodes while using the workflow?

- **Yes**
- **List of Missing Custom Nodes:**
    - **CLIPiton Loader**:
        - **Download**: Download “comfy-cliption” in Custom Node Manager
    - **CLIPiton Beam Search**:
        - **Download**: Download “comfy-cliption” in Custom Node Manager
    - **EmptySD3LatentImage**:
        - **Download**: Download “comfy-cliption” in Custom Node Manager

3. Checkpoint/Model(s) Used in the Workflow

List of Checkpoint/Model(s):

- **dreamshaper_8: Base model:** SD 1.5, used for text-to-image generation. **https://civitai.com/models/4384/dreamshaper**
- **MOHAWK_v20:** Base model: SDXL 1.0, MOHAWK is a merge from various generation models so it can perform on many subjects. It is used for text-to-image generation. **https://civitai.com/models/144952/mohawk**

- **aamXLAnimeMix_v10:** Base model: SDXL 1.0, used for anime and stylized art image generation. **https://civitai.com/models/144952/mohawk**
- **absolutereality_v181:** Base model: SD 1.5, used for hyper-realistic image generation with lifelike details. **https://civitai.com/models/144952/mohawk**
- **westernAnimation_v1:** Base model: SD 1.5, used for generating western fantasy animation. **https://civitai.com/models/144952/mohawk**

4. Image-to-Prompt-to-Image Workflow Results

For each of the five given images, the following steps were followed:

- **Image 1:**
  - **Original Image:** **Re-generated Image:**

  - **Generated Text Prompt:** " a large group of people in white robes are standing in front of a temple with a sun in the background. the temple is adorned with intricate carvings and has a statue of a woman with outstretched arms. the sky is filled with clouds, and there are birds flying in the distance. the scene is illuminated by a warm, golden light, creating a serene and mystical atmosphere."

- **Image 2:**
  - **Original Image:** **Re-generated Image:**

- **Generated Text Prompt:** " a futuristic scene with a humanoid figure suspended in a transparent cylinder, surrounded by scientists in white lab coats, in a high - tech facility with advanced technology and a view of outer space through a large window. the room is filled with machinery and screens, and the overall atmosphere is one of advanced technology and advanced technology. the overall color palette is dominated by blues and whites,"

- **Image 3:**
  - **Original Image:** **Re-generated Image:**


  - **Generated Text Prompt:** " a stylized female figure with wings and a halo is depicted in a regal pose, wearing a white and red robe with golden patterns and a golden crown. she is surrounded by a halo - like aura and is accompanied by two white birds. the background is a deep blue, and the overall color palette is dominated by shades of red, orange, and gold."

- **Image 4:**
  - **Original Image:** **Re-generated Image:**


  - **Generated Text Prompt:** " a magical forest scene at night features twisted trees, a serene body of water, and a figure holding a staff with a glowing orb, set against a backdrop of a moonlit sky with a mix of blue, green, and yellow hues, with a sense of mystery and wonder. the overall atmosphere is mystical and otherworldliness, evoking a sense of mystery and otherworld"

- **Image 5:**
  - **Original Image:** **Re-generated Image:**

  - **Generated Text Prompt:** " a video game screenshot of a city with a man in a trench coat standing in front of a building with a neon sign that reads 'link '. the city is bustling with activity, and the sky is overcast. the characters are dressed in dark clothing, and there are various signs and advertisements scattered throughout the scene. the overall color palette is dominated by shades of blue and"

5. Observations/Self-Reflection

Brief Reflection:

The workflow I used effectively generates fine-grained and detailed text descriptions, which enhance the process of re-generating images by providing both broad context and details. However, I have identified several challenges. First, the generated text often exceeds the maximum length required by the CLIP text encoder, leading to truncation and loss of crucial text details (see image 2, 4 and 5). Second, while the CLIP Vision model excels at capturing objects and styles, it struggles with accurately describing positions and shapes (e.g., the shape of the moon), limiting its effectiveness in subsequent image generation.