Project Proposal: Dynamic Image Generation via Multi-Modal Weight Assignment and ControlNet Interaction

Jialiang Yuan Jingyun Ma

Abstract

This project introduces a novel method for dynamic image generation and manipulation using multi-modal inputs. Using a combination of ControlNet and multi-modal weight assignment, users can make real-time adjustments to visual features such as object prominence, position, and style. Additionally, users can manipulate the image via a dialogue-like interface, allowing them to interact with the model naturally and intuitively to generate desired images.

The evaluation of this model will compare the result with the same prompt generated with diffusion models, and the same prompt generated with ControlNet plus diffusion models. The overall performance will be judged based on accuracy, efficiency, and usability. The expected result of this dynamic image generation system would be a highly interactive system, enabling the user to make accurate and flexible changes compared to existing models.

1. Project Overview

Currently, The ControlNet architecture has proven to be an effective solution for multi-modal inputs, as demonstrated by Zhang et al. [1]. The field lacks intuitive, real-time control image generation models incorporating multi-modal inputs. Although ControlNet offers great results for image generation and control, it requires multiple iterations of prompt editing and lacks dynamic, multi-modal input handling in real-time settings [1]. Also, ControlNet struggles with multi-condition generation and becomes inefficient as the number of conditions grows. [2]

This project aims to develop a dynamic image generation system that combines multimodal inputs with a large language model (LLM), enabling users to make real-time adjustments to visual features. The system integrates segmentation models, image-to-text models, and depth maps. The LLM will help decide the dynamic weights for each input modality and form a dialogue-like interface that lets users manipulate images easily and efficiently. Users can adjust the prominence, position, color, and other features of objects by providing natural language commands.

The desired outcome of this project would be a highly interactive system, enabling the user to make accurate and flexible changes compared to existing models. The minimum goals are to enable user control and incorporate the image-to-text model into the system.

2. Team Member Roles/Tasks

2.1. Jialiang Yuan

1. Responsible for integration of ControlNet with multi-modal input models.
2. Oversee the development of the user interface for real-time adjustments.
3. Coordinate dynamic weight adjustment algorithm using LLM for interpreting user input.

2.2. Jingyun Ma

1. Implement segmentation and depth map models for accurate image decomposition.
2. Develop image-to-text labeling for object recognition.
3. Collaborate on optimizing LLM dialogue interpretation for user commands.

3. Resources

- **Hardware**: Access to GPU clusters for training ControlNet, Depth Map, Segmentation model, Image-to-Text model, and Large-Language models. We have a 24G 3090 and 16G 4070 Ti Super for model training.
- **Software**: TensorFlow or PyTorch for model development.
- **Pre-trained Models**: Utilize pre-trained models for ControlNet[1] and generated models , use Stable Diffusion 3[2]. For segmentation models, use Segment Anything[3]. For Image-to-text, use vit-gpt2-image-captioning[4]. For depth map, use Depth Anything[5].

[1] ControlNet: https://huggingface.co/lllyasviel/ControlNet
[2] sd3: https://huggingface.co/stabilityai/stable-diffusion-3-medium
[3] https://github.com/facebookresearch/segment-anything
[4] https://huggingface.co/nlpconnect/vit-gpt2-image-captioning
[5] https://huggingface.co/depth-anything/Depth-Anything-V2-Large

4. Reservations

- Scalability: Real-time performance of the system might face challenges if the computational cost of multi-modal input processing is high.
- User Interface: Ensuring that the dialogue-based interface is intuitive and responds accurately to user inputs may require extensive user testing and iteration.
- Integration of multimodals: Combine different models, dynamically assign weight to different components. There might be challenges in combining these components.
- Performance: Utilizing multi-modals with ControlNet might result in unexpected performance. Ensure the accuracy and efficiency of the outcome.

5. Relationship to Background

Our team has experience in computer vision and natural language processing. We previously worked on image segmentation tasks[6] and have implemented large language models to generate a prompt for image generation. This project builds on our experience by combining these skills to create an interactive real-time system for dynamic image generation.

References

[1] Jing Zhang, Qiang Liu, Peng Wang, et al. Controlnet: Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[2] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 1

[6] Using yolov8, stable diffusion and Flux