Abstract

Current image generation methods involve users inputting textual information as prompt, which serves as a text-based conditional control input to guide the image generation process. This text-based approach may have difficulty meet the user's requirements and initial design intentions in the generated image. In common cases, the users shall either regenerate the image from scratch or manually modify the generated image using their drawing skills. These design processes are normally time-consuming and labor-intensive, and it requires a certain level of drawing skills and experiences from the users. Recently, deep learning-based image generation models have demonstrated remarkable success in generating high quality images. Generating coherent image sequences is essential for applications such as manga, animation, and other visual storytelling mediums, where maintaining consistency across multiple frames is critical to preserving narrative flow and visual continuity. However, the state-of-the-art image generation current models still face challenges in generating coherent image sequences, such as insufficient consistency in image features, including character appearance, background elements, and object positions across frames. Generating high-quality and consistent image sequences is crucial for storytelling and visual effects in video game development, virtual reality experiences, and film production. Traditional image generation methods often fail to meet the requirements of maintaining feature consistency across sequential images, necessitating the development of new methods to enhance image generation quality and consistency.

This thesis proposes a two-stage image generation model designed to generate multiple sequential images based on a series of coherent text prompts. In the first stage, the model generates a set of sketches based on the user's textual input. Users can modify these sketches and arrange them according to their design intentions. In the second stage, the model utilizes the revised sketches from first stage to generate a series of images with consistent image features. In this work, we propose the Multi-Aggregation Attention (MAGA) mechanism in the second stage, which can significantly enhance feature aggregation and refinement, thereby improving image consistency. The MAGA mechanism incorporates an external memory component during the image generation process, storing conditioning information from previous network layers and establishing connections between the current text, sketches, and spatial features, and all historical inputs including prior

text and sketch information, thereby effectively increasing consistency across sequential images.

To enhance user flexibility, we proposed a sketch selection workflow, offering two more options. First, users can manually input their own sketches, which allows for highly customized and precise visual elements that reflect specific user intentions. Alternatively, users can select sketches from a high-quality, pre-constructed database designed to support the image generation process. This sketch database, called SketchXL, includes about 1,500 coherent storylines of continuous sketches, thus addressing the lack of sequential consistency found in existing datasets.

Through a series of quantitative and qualitative experiments, we verified that our model can demonstrate outstanding performance in image quality, consistency, and fidelity of the generated images from the input sketches and text prompts. Firstly, the proposed two-stage model demonstrated an advantage over single-stage generation models by allowing users to adjust the final output by editing the sketches generated in Stage 1, rather than the final generated images. This approach can not only simplify the drawing process but also enhance the quality of the generated images. In our experiments, we compared the proposed model against several state-of-the-art models under the same input conditions. We evaluated the generated images for consistency and detail preservation to the input sketches and text prompts. The experimental results show that the proposed model has significant advantages in generating continuous image sequences, maintaining consistency in character features and background images while preserving details. Additionally, we quantified image quality using the NIMA score, with the proposed model achieving an overall image quality score of 6.04, outperforming other state-of-the-art models (5.84 for Stable Diffusion v1.0, 5.99 for Stable Diffusion XL, 4.95 for ControlNet). Overall, our method provides an efficient and high-quality image generation solution for the applications of animation and comic creations, with broad potential applications.

Keywords: Latent Diffusion Model, Attention Mechanism, Consistent Images Generation.