TRANSFER DENSE DIFFUSION ON SD 3.5

Yuehao Wu

December 24, 2024

1. INTRODUCTION

In recent years, the text-to-image generation technology has made significant progress with the development of diffusion models. As a representative work in this field, Stable Diffusion has attracted widespread attention due to its high generation quality, flexible architecture and open source. This report aims to adapt and implement the methods proposed in the paper "Dense Text-to-Image Generation with Attention Modulation" into the Stable Diffusion 3.5 (SD 3.5) model. The goal is to explore the potential role of attention modulation in diffusion models, explore and try to improve the accuracy and expressiveness of text-to-image generation tasks.

The core objectives of this work include the following three aspects:

1. Analyze the DenseDiffusion paper, understand and explain its methodological contributions and theoretical insights.

2. Convert these methods into an architecture based on the SD 3.5 model and make necessary component modifications.

3. Research findings, challenges encountered in the process, and possible next steps in the future

2. DENSE TEXT-TO-IMAGE GENERATION WITH ATTENTION MODULATION

2.1. Objectives

Existing text-to-image diffusion models often face challenges when processing dense description text, and it is difficult to generate high-quality images that match complex descriptions. Dense description text usually consists of multiple phrases, each of which corresponds to a specific area in the image. Existing models tend to ignore or confuse the detailed features of these areas. In addition, the method of generating images relying solely on text prompts lacks control over the scene layout, and it is difficult for users to accurately define the objects and layout relationships in the generated images. To address these issues, this paper aims to propose an efficient method without additional training, which enables pre-trained text-to-image models to better handle dense description text while providing control over scene layout. The goal is to improve the text fidelity and layout alignment of generated images while avoiding the increased computational cost and complexity due to model fine-tuning.

2.2. Methodology

By designing and adopting the following methods, this paper achieves a significant improvement in the fidelity of generated images to dense text descriptions and layout conditions without any fine-tuning of the pre-trained model, providing higher controllability and efficiency for text-to-image generation.

2.2.1. Dynamic Attention Modulation

The study found that the layout characteristics of the generated images are closely related to the attention mechanism of the model, especially the self-attention and cross-attention layers. In the early stages of the generation process, the attention map begins to determine the position and layout of the objects, while in the later stages, it gradually refines the detailed features of the objects, such as color and texture. Through quantitative analysis of the attention scores, the study shows that queries and key-value pairs belonging to the same object usually have higher attention scores. These results provide a theoretical basis for the proposed dynamic attention modulation method. Based on the above observations, this paper designs a layout-based dynamic attention modulation method, which enables the generated object features to be accurately mapped to the specified layout area by modifying the attention scores of the cross-attention and self-attention layers in real time. In the cross-attention layer, the method enhances the association between the image and text tags describing the same area by modulating the attention score, thereby ensuring the accuracy of the object in the layout condition. In the self-attention layer, the method avoids the confusion of object features by limiting the communication between different regions. The implementation of attention modulation includes two key operations: positive modulation of pairs belonging to the same region to enhance their scores, and negative modulation of pairs belonging to different regions to weaken their scores.

2.2.2. Adaptive mechanisms of value domain and mask region

In order to ensure that the modulation process does not destroy the generation ability of the pre-trained model, the

method further introduces multiple adaptive mechanisms. The first is "value domain adaptive attention modulation", which dynamically adjusts the modulation amplitude according to the maximum and minimum values of the original attention score to maintain the stability of the generation quality. The second is "mask region adaptive modulation", which adjusts the modulation intensity according to the size of the layout area, enhances the modulation for smaller areas, and weakens the modulation for larger areas to avoid visual inconsistency in the generated image. In addition, the modulation intensity will also change dynamically with the time step of the generation process, stronger in the early stage of generation and gradually weakened in the later stage to maintain the quality of the final image.

2.2.3. Optimization details

In the implementation process, this paper uses the initial denoising stage to modulate the attention map based on the stable diffusion model to avoid unnecessary interference to the refinement process in the later stage of generation. At the same time, in order to enhance the generation ability of multi-object text description, the method further decomposes the text features into multiple segments, and each segment is encoded separately to improve its independence.

2.3. Contributions

The main contribution of this paper is to propose an improved attention modulation method that enables the pre-trained diffusion model to more efficiently generate images that meet complex text description and layout requirements. This method significantly improves the performance of the model in processing densely described text, while allowing users to precisely control objects and scenes in the image by specifying layout conditions. Compared with other existing methods that require training or fine-tuning, this method significantly reduces the computational cost. Through a large number of experiments, this paper verifies the advantages of this method in terms of text fidelity and layout alignment. Quantitative results show that DenseDiffusion surpasses existing training-independent methods in multiple evaluation metrics, while the generation quality is comparable to models trained specifically for layout control. Overall, this paper provides an efficient and flexible solution for the study of text-to-image generation, and points out valuable directions for future related work.

3. IMPLEMENTATION

3.1. Environment

The code implementation and experimental validation were performed in Google Colaboratory, an online Jupyter Notebook environment integrated with Python libraries that suitable for machine learning and deep learning.

Category	Parameter/Configuration
Computing Platform	Google Colab Pro
GPU	NVIDIA Tesla T4
System Memory (RAM)	51 GB
Disk Storage	235 GB
CUDA Version	12.2
Python Version	Python 3.10
Diffusers Version	0.31.0
Transformers Version	4.47.1
Gradio Version	3.43.2
Bitsandbytes Version	0.45.0

Table 1. System Configuration

3.2. Stable Diffusion 3.5

Compared with previous generations and other diffusion models, Stable Diffusion 3.5 significantly improves text description complexity, layout control accuracy, and generation efficiency. Through multimodal feature processing, dynamic attention modulation, adaptive modulation mechanism, and optimized generation strategy, the model performs particularly well in complex multi-object scenes. At the same time, its training-independent characteristics make it more flexible and provide more efficient solutions for a variety of application scenarios.

The core of Stable Diffusion 3.5 is multiple MM-DiT Blocks. Each MM-DiT Block contains a multi-layer attention mechanism, a feedforward network, and a normalization module, supports the interaction and fusion of multimodal features, and introduces a value range adaptation mechanism when dynamically adjusting the attention value.

3.3. Transfer

3.3.1. Improve layout controls

To improve the layout control in Stable Diffusion, segmentation masks can be used to explicitly constrain the generated image features based on the Dense Diffusion method. In the initial stage of the diffusion model (the stage with a large number of time steps), mask constraints are applied to each segment of the noise image to ensure that the features are distributed in the specified area. And according to the segment size and target complexity, the attention weight is dynamically adjusted so that smaller segments receive higher attention values.

$$
\begin{aligned}
latent &= latent \cdot segment\_mask \\
&\quad + noise \cdot (1 - segment\_mask) \\
segment\_attention\_weights &= compute\_attention\_weights \\
&\quad (segment\_size, target\_complexity) \\
attn\_weights &= attn\_weights \cdot segment\_attention\_weights
\end{aligned} \tag{1}
$$

3.3.2. *Dynamic attention modulation mechanism*

The core of Dense Diffusion is dynamic attention modulation, which dynamically adjusts the attention distribution according to the textual cues and segmented regions. To integrate this feature into Stable Diffusion, dynamic modulation can be achieved by modifying the cross-attention and self-attention modules. For the Cross-Attention layer, the segment weight factor is added to the Query-Key weight calculation of the Cross-Attention, and the attention range is limited by the mask. The mask of each segment can be used as an additional input, combined with the text features, to guide the model to dynamically distribute the attention. In the Self-Attention layer, the segment mask is introduced to suppress the feature communication between different segments, enhance local consistency, and limit the attention to the same segment area.

$$
\begin{aligned}
attn\_weights &= softmax\left(\frac{query \cdot key^{\top}}{\sqrt{d_{model}}}\right) \\
attn\_weights &= attn\_weights \cdot segment\_mask
\end{aligned} \tag{2}
$$

$$
\begin{aligned}
attn\_weights &= softmax\left(\frac{query \cdot key^{\top}}{\sqrt{d_{model}}}\right) \\
attn\_weights &= attn\_weights \cdot intra\_segment\_mask
\end{aligned} \tag{3}
$$

3.3.3. *Segmented control capability*

Dense Diffusion supports binding independent text prompts to each segment. To achieve this in Stable Diffusion, segmentation conditions and segment-level encoding can be introduced at the input stage. The input text prompt is divided into segments, and each segment generates an independent feature vector. Then, the tokenized segment information (such as [SEGMENT_1] ... [SEGMENT_2]) is used to clarify the text range corresponding to each segment.

$$
\begin{aligned}
segment\_embedding &= segment\_mask \cdot learnable\_embedding \\
text\_features &= text\_features + segment\_embedding
\end{aligned} \tag{4}
$$

Combine the segmentation condition (mask) with the corresponding text features and add segmentation information through an additional embedding layer

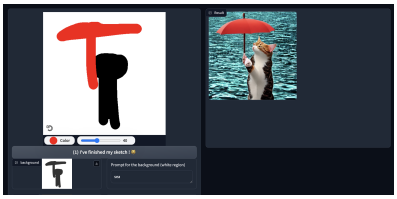

Fig. 1. Dense Diffusion - 1

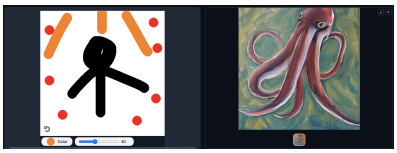

Fig. 2. Dense Diffusion - 2

4. RESULT ANALYSIS

We use *A magnificent squid, shimmering in shades of deep indigo and vibrant teal, glides gracefully through an enchanting rose garden where blossoms of crimson, blush, and gold radiate under the soft, dappled sunlight* as the prompt for diffusion generation of each model.

4.1. Dense Diffusion

First, we generate the image using "cat with red umbrella" as the cue word. Dense Diffusion performs well in simple scenes. However, when the cue word is more complex, we can observe that the model tends to ignore some details when processing complex scenes, and these details fail to meet the requirements of the cue word well.

4.2. Stable Diffusion 3.5

Under the same prompt word conditions, the images generated by Stable Diffusion 3.5-large show significant improvements in detail precision and generation accuracy, and all provided keywords are more comprehensively covered. The images generated by Stable Diffusion 3.5-medium show that their overall quality is still high, but there are some unrea-

Fig. 3. SD 3.5 - Medium

Fig. 4. SD 3.5 - Large

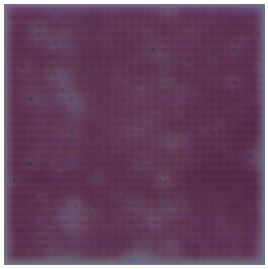

Fig. 5. Transfer - 1

sonable aspects, such as the suspension and breakage of the octopus tentacles.

4.3. Transfer

It was found that the segment mask had a limited impact on the generated image. It only guided the generated result to a certain extent, but did not significantly improve the quality or structural accuracy of the image. In addition, when the model regenerated the image based on the prompt, it was unable to get rid of the interference of noise. That is, the consistency of the generated result in the denoising process was low and the semantic fit of the prompt was low. This shows that the current adaptation model is likely to have defects in the attention mechanism, which affects the final output result.

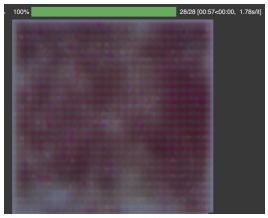

Fig. 6. Transfer - 2

5. FUTURE WORK

5.1. Package version compatibility

Dense Diffusion uses diffusers 0.20.2, which does not support Stable Diffusion 3.5. This version difference will lead to functional incompatibility and hinder the integration of Dense Diffusion dynamic functions. Therefore, we still need to explore and test the performance of Dense Diffusion modules in the updated version of diffusers to determine whether functional docking can be achieved without changing the core logic, and to further improve performance by taking advantage of the optimization features of the new version of the library.

5.2. Optimization and adaptation of dynamic attention logic

Although Stable Diffusion has achieved alignment of text and image features through cross-attention and self-attention, its mechanism is relatively static and lacks the dynamic modulation capability of Dense Diffusion. We need to correctly introduce dynamic modulation function in the cross-attention layer of stable diffusion, adjust the attention weight in real time according to the complexity of the input text (such as multiple objects or dense descriptions), and modify the self-attention layer to dynamically adjust the receptive field of attention according to the segmented area, limit the interference across segments, and strengthen the consistency within the segment.

5.3. Alignment of segmentation mask and latent space

The representation of the segmentation mask of Dense Diffusion in the latent space must be consistent with the generation process. If the mask representation is not used properly in the latent space, the generation of the segmented content will not meet expectations and the overall image quality will be reduced. We still need to introduce a learnable embedding for the segmentation mask and embed it into the latent space of the diffusion process. This embedding needs to be throughout the entire generation process and ensure that the segmentation mask remains consistent at different resolutions in the latent space.

6. REFERENCES