Text-to-Image Generation using CLIP and Diffusion Models - SSY340 - Project Group 24

Daniel González Muela Sotiris Koutsoftas

Chalmers University of Technology
Electrical Engineering Department

Introduction

Diffusion models provide a powerful approach for image generation, outperforming traditional generative models like GANs and VAEs by leveraging a stable, iterative denoising process. This project explores diffusion models, particularly DDPM and DDIM, and integrates CLIP-based guidance for controlled text-to-image generation.

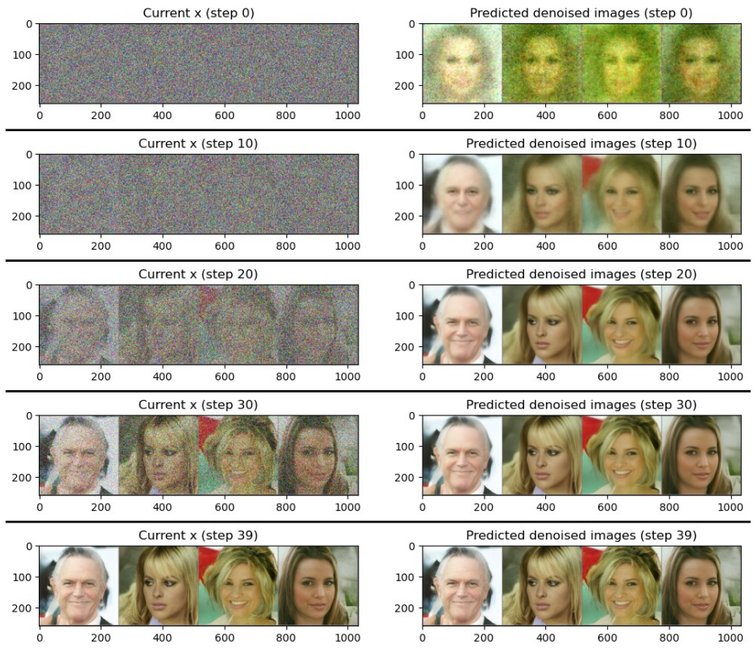


Figure 1: Denoising stages in diffusion process on a celebrity face dataset.

DDPM & DDIM: Noise Generation and Denoising

Denoising Diffusion Probabilistic Models (DDPM) and Denoising Diffusion Implicit Models (DDIM) are core frameworks for diffusion-based image generation. DDPM iteratively adds noise to data and learns to reverse this process, while DDIM reduces the steps required, making generation faster.

DDPM DDIM

Fine-tuning is applied to adapt these models to new datasets. After 10 epochs of fine-tuning on a butterfly dataset, distinct butterfly features emerge.

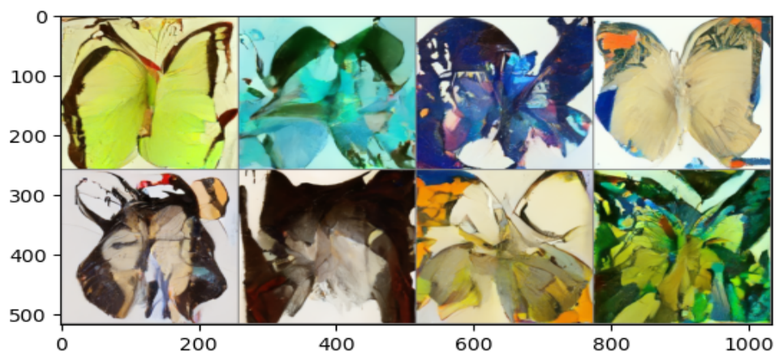


Figure 2: Generated butterflies after 10 epochs of fine-tuning.

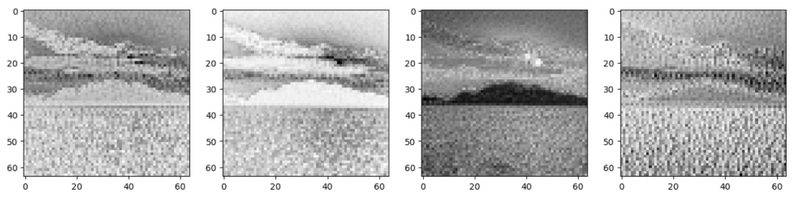
Guidance with CLIP enables text-driven refinement by using CLIP embeddings to steer the diffusion process toward specific descriptions.

Figure 3: CLIP Guidance applied to WikiArt Dataset.

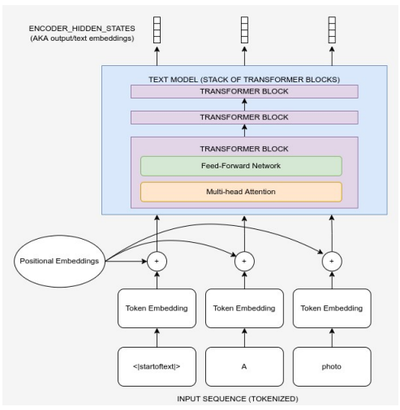
Figure 4: CLIP Guidance applied to Butterfly Dataset.

Stable Diffusion & CLIP Guidance

Stable Diffusion, an efficient text-to-image model that operates in a compressed latent space, enabling faster generation while maintaining high image quality. By incorporating CLIP embeddings as a conditioning mechanism, we achieve improved alignment between the generated images and specific textual descriptions, allowing for fine-grained control over the visual output.



Latent Diffusion Process



Text Conditioning with CLIP

Key Points:

- Faster generation in a compressed latent space.
- Improved control over generated content through CLIP guidance.
- Adaptable to various datasets, including WikiArt and butterflies.

Guided Image Generation Results

By leveraging advanced guidance techniques using CLIP and BERT, we achieved variable levels of text-aligned image generation across structured and abstract datasets, showcasing the potential and limitations of different guidance mechanisms.

Figure 5: Image generated using the prompt: "A dancer made of smoke and light, performing on a stage of shattered glass"

Key Observations:

- **CLIP Guidance:** Effective for aligning images to text prompts across diverse styles.
- **BERT Guidance:** Failed in terms of image generations w.r.t the input promt, images remain abstract, bad quality, highlighting the importance of multimodal training.
- **Latent Diffusion Efficiency:** Allows high-resolution outputs with reduced computational overhead.

- Optimal CLIP guidance scales (0.8 out of 1) yield the best quality images, balancing prompt adherence and realism.
- BERT-guided results reveal its limitations, with both high and low guidance strengths resulting in chaotic images due to its text-only focus (lacks visual understanding).
- Latent diffusion and VAE compression enable efficient high-fidelity generation.

Figure 6: Comparative BERT-guided images generated with increasing guidance scales for the prompt "A futuristic cityscape under a neon sky."

References

[1] High-Resolution Image Synthesis with Latent Diffusion Models.
https://arxiv.org/pdf/2112.10752

[2] Stable Diffusion Tutorial.
https://huggingface.co/learn/diffusion-course/en/unit3/1

[3] Fine-Tuning and Guidance.
https://huggingface.co/learn/diffusion-course/en/unit2/2#what-you-will-learn

sotirisk@student.chalmers.se muela@student.chalmers.se