Multiple Concept Personalization

Team 9 ACL
R13921010 電機所碩一 王靖睿
R13921001 電機所碩一 徐詩婷
R13522814 機械所碩一 張舜凱
R12942009 電信所碩二 廖珀毅

Abstract

Customizing text-to-image models for multi-concept scenarios remains challenging due to quality limitations, high training costs, and low success rates. In this project, we address these issues by leveraging the residual embedding method from Cones V2. This approach seamlessly combines arbitrary concepts without additional model tuning, delivering high-quality image generation on pair with advanced methods like ED-LoRA[2], while maintaining simplicity and efficiency.

Introduction

Text-to-image diffusion models have achieved remarkable success in generating realistic high-resolution images. Built on this foundation, techniques for personalization have also advanced, not only increasing the performance and accuracy, but also widen the application form single to multiple concept.

In this project, we inherited the main techniques and work from Cones V2[1], as we fine-tuned the text-encoder part of a pre-trained text-to-image diffusion model with images of a specific object, as we also calculate the difference between the tuned text-encoder with the original text-encoder, so that the generated object can be more similar to our customized subject than the original general category.

To effectively combine different subjects, we equipped layout guidance, a technique that not only is abstract and easy-to-obtain prior, also act as a spatial guidance to ensure the alignment of subject arrangement. Moreover, we also discuss with some common issues in multiple concept personalization, like attribute leakage, concept omission, or attention calibration loss etc.

There are eight subjects for training and four prompts which all contains two or more concepts for evaluation. It's trustworthy that our work can successfully deal with multiple concept personalization and preventing common breakdowns.

Related Works

	Method	Limitation
Textual Inversion	train new token embedding	Fail on multiple subjects
DreamBooth	fine-tunes all parameters	Memory consuming
Custom Diffusion[3]	New text embedding and K V matrices in CA layers	More than 3 Objects
Concept Conductor[2]	Attention Layers in LDM and ED LoRA	Computation Resources
Cones V2	Residual token embedding and layout guidance	Similar objects

Methodology

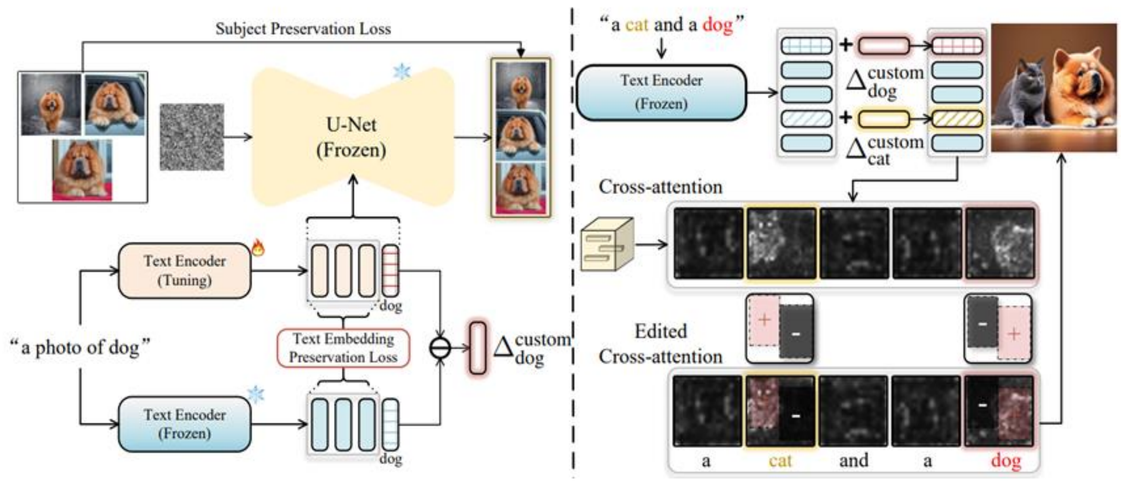


Figure 1. Structure Diagram of ConesV2

- **How we improve? (Version 1)**

We introduced higher intermediate resolution cross-attention weights and excluded lower intermediate resolution weights to enhance multi-scale representation. This adjustment is based on the consideration that performing cross-attention at very low resolutions can lead to pixel-level mixing during downsampling, such as blending features of similar concept (e.g., a cat and a dog). By focusing on higher resolutions, images can preserve spatial details and improve object separability.

We enhanced the static classifier free guidance scaling method by incorporating the noise sigma value into the calculation, implementing a CFG++ approach. By tying the adjustment to the noise level at each timestep, this CFG++ method allows for more adaptive control over conditional guidance, improving the balance between global coherence and fine details in the generated images.

- **How we improve? (Version 2)**

We modified the original tokenizer setting, making the tokenization process adaptive to the input prompt's actual length. This modification reduces unnecessary padding and prevents truncation of meaningful tokens, leading to more accurate text embeddings. As a result, the generated images are clearer and more detailed with this optimized padding approach, likely because the model can better concentrate on the relevant input information.

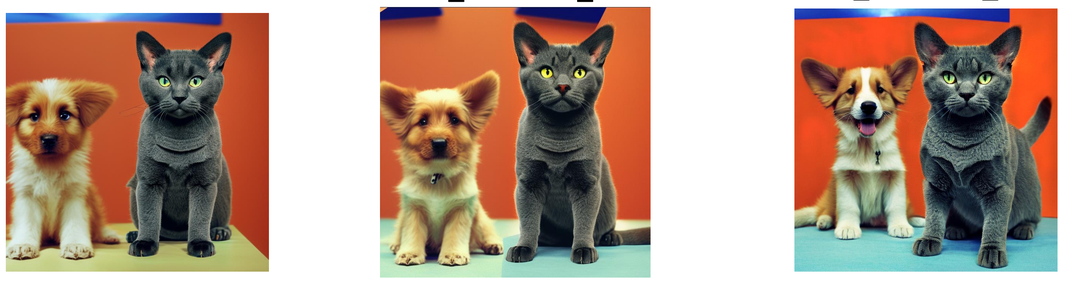
References

[1]Liu, Zhiheng, et al. "Customizable image synthesis with multiple subjects." Advances in Neural Information Processing Systems 36 (2024).
[2]Yao, Zebin, et al. "Concept Conductor: Orchestrating Multiple Personalized Concepts in Text-to-Image Synthesis." arXiv preprint arXiv:2408.03632 (2024).
[3]Gu, Yuchao, et al. "Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models." Advances in Neural Information Processing Systems 36 (2024).
[4]Kumari, Nupur, et al. "Multi-Concept Customization of Text-to-Image Diffusion." arXiv preprint arXiv:2212.04488 (2022).
[5] Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.
[6] ZhihengLiu,RuiliFeng,KaiZhu,YifeiZhang,KechengZheng,YuLiu,DeliZhao,JingrenZhou,and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. arXiv preprint arXiv:2303.05125, 2023.

Experiment Results

The figures highlight our second method, which refineds the token padding approach, surpassing the original ConesV2 and addressing the common challenge of similar concept generation encountered in both Custom diffusion and the ConesV2.

ConesV2 | Our_ConesV2_v1 | Our_ConesV2_v2



Fifgure2. Figures with prompt "A **cat** on the right and a **dog** on the left."

ConesV2 | Our_ConesV2_v1 | Our_ConesV2_v2

Figure3. Figures with prompt "A **flower** in a **vase**."

ConesV2 | Our_ConesV2_v1 | Our_ConesV2_v2

Figure 4. Figures with prompt "A **cat** wearing **wearable glasses** in a **watercolor** style."

ConesV2 | Our_ConesV2_v1 | Our_ConesV2_v2

Figure 5. Figures with prompt "A **dog** and a **flower** near a lake"

In the following figures, we achieved better quality with our own method comparing to the original ConesV2 and the results are closing to the Concept Conductor in cases where the concepts are quite similar.

ConesV2 | Our_ConesV2 | Concept Conductor

CLIP-I: 76.75
CLIP-T: 29.41

CLIP-I: **77.43**
CLIP-T: **30.22**

CLIP-I: 77.81
CLIP-T: 30.85

Figure 6. Figures with prompt "A **cat** on the right and a **dog** on the left."

Conclusions & Future Works

In this work, we identified challenges in generating from similar categories. Through our approach, we successfully achieved state of the art performance.

To sum up, we opted for Cones due to its balanced between the performance and memory efficiency compared to Concept Conductor. By addressing edge cases and the generation process, we delivers higher quality with reduced memory usage and smaller computation resources than using concept conductor. With only 5KB required in each residual embedding whereas each trained LoRA demands 4.6MB.

Inspired by the power of the new token embeddings, we aim to use less memory to finetune the model to reach SOTA for memory efficient image personalization.