FEW-SHOT PERSONALIZATION OF TEXT-TO-IMAGE GENERATION MODELS

Gia-Nghia Tran[1,2]

[1] University of Information Technology, Ho Chi Minh City, Vietnam [2] Vietnam National University, Ho Chi Minh City, Vietnam

What ?

We introduce a framework for few-shot fine-tuning generative models in personalized image generation.

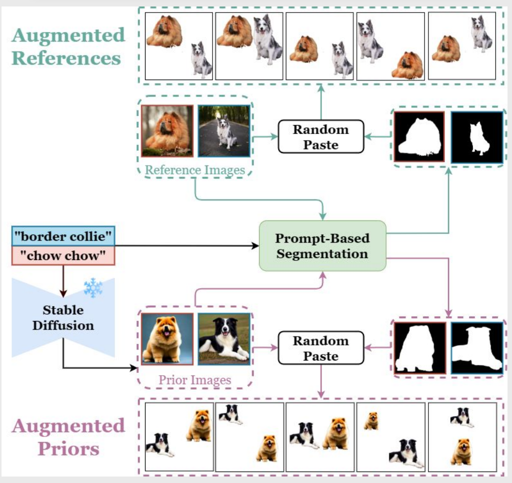
- Develop and integrate **augmentation techniques** to mitigate overfitting in low-data scenarios.
- Propose a **fine-tuning pipeline** to efficiently embed personalized concepts from a small reference set into the image generation process.
- **Implement and evaluate** the proposed approach for both single and multi-concept generation tasks.

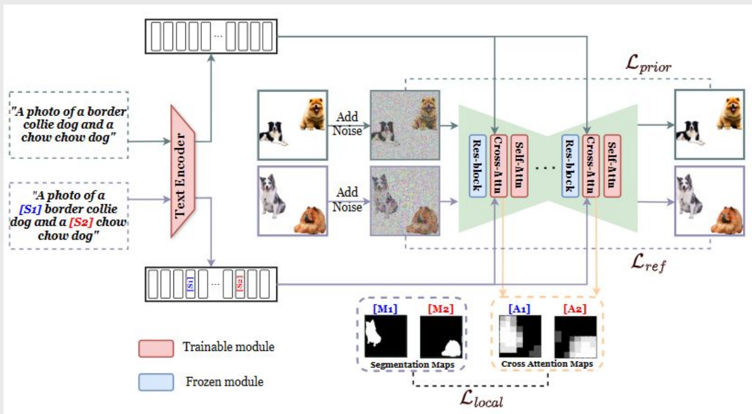
Why ?

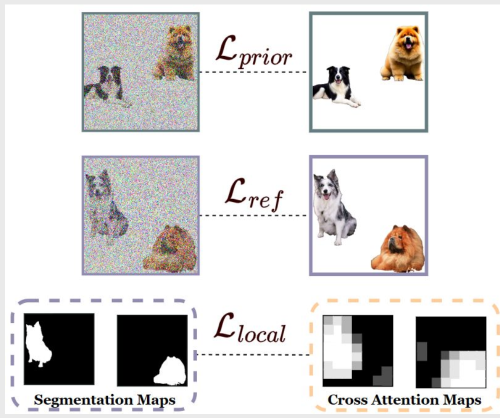
- While text-to-image diffusion models have made significant progress, they often **lack personalization** due to their training on large, generalized datasets. This leads to **poor adaptation when users have only a few reference images or a unique artistic style.**
- Existing methods face challenges such as **overfitting, loss of subject details, and reduced diversity in generated images.**

Overview

Data Augmentation → Fine-tuning Pipeline → Objective Functions







Description

1. Data Augmentation

- The Stable Diffusion model is employed to generate additional ***prior images*** based on the same fine-grained class as the reference images, thereby ***expanding the range of variations in pose, shape, and viewpoint.***

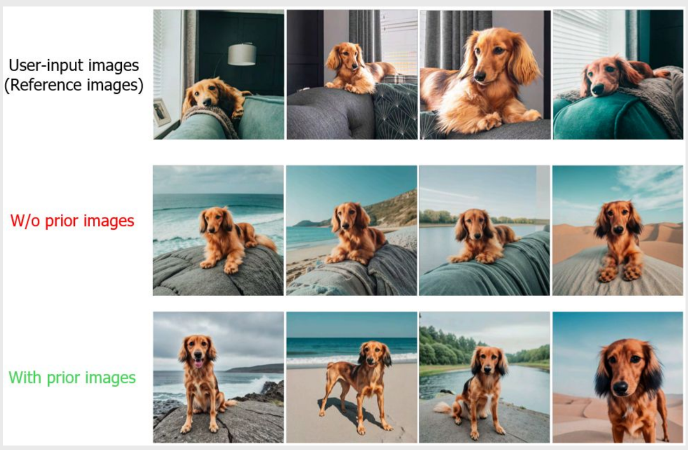


Figure 1. Encouraging diversity with prior images.

- The ***segmented subjects are randomly translated and resized onto a simple background***—potentially overlapping each other—to further ***diversify the training data*** for both reference and prior images.

2. Fine-tuning Pipeline

- Tuning ***Cross-attention layers*** improves the alignment between textual prompts and generated visual features.

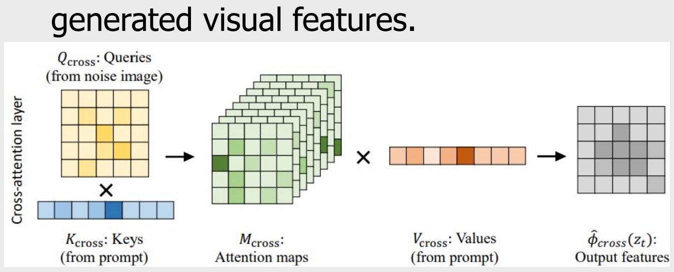


Figure 2. Cross-attention layers in Stable Diffusion.

- Adjusting ***Self-attention layers*** enhances the model's ability to capture complex spatial relationships and fine details.

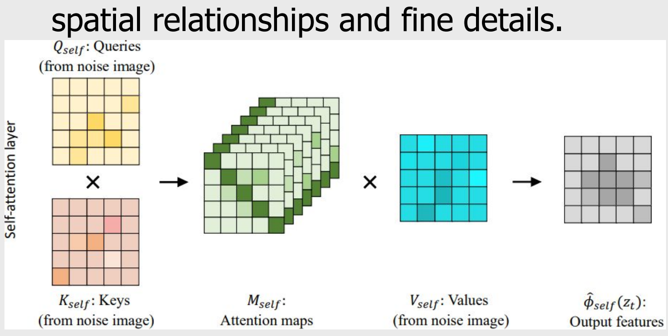


Figure 3. Self-attention layers in Stable Diffusion.

- Refining the ***Text Encoder*** allows for a more precise representation of the semantic space.

- At each training step, the model's task is to ***predict the amount of noise*** that has been added to the original image.

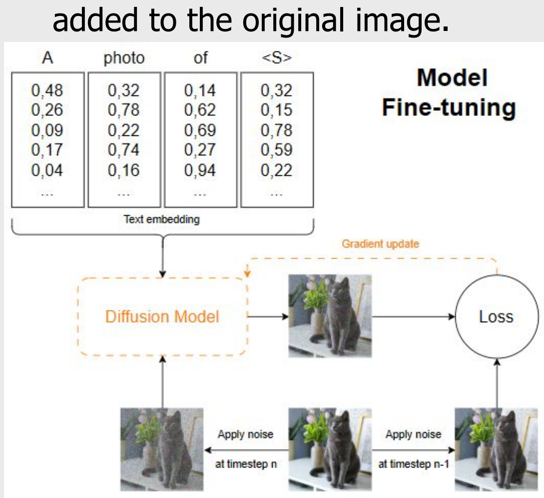


Figure 4. Model Fine-tuning process.

3. Objective Functions

- *L_ref* and *L_prior* are used to optimize subject reconstruction and are based on the *Mean Squared Error* (MSE) function.
- *L_local* gives precise supervision whether a pixel belongs to the segmented region and based on the *Binary Cross Entropy* (BCE) function.

NII

Gia-Nghia Tran – University of Information Technology
TEL : 0983932250 Email : nghiatg.19@grad.uit.edu.vn