COSC-5455 Term Project Presentation

Fine-tuning Strategies Based on DreamBooth

Yueran Cao

Zhengheng Li

Xizhong Xu

Zinian Wang

GEORGETOWN UNIVERSITY

Outline

- **Overview** by Yueran Cao
- **Methodology** by Xizhong Xu
- **Experiments** by Zhengheng Li
- **Evaluation** by Zinian Wang

Overview

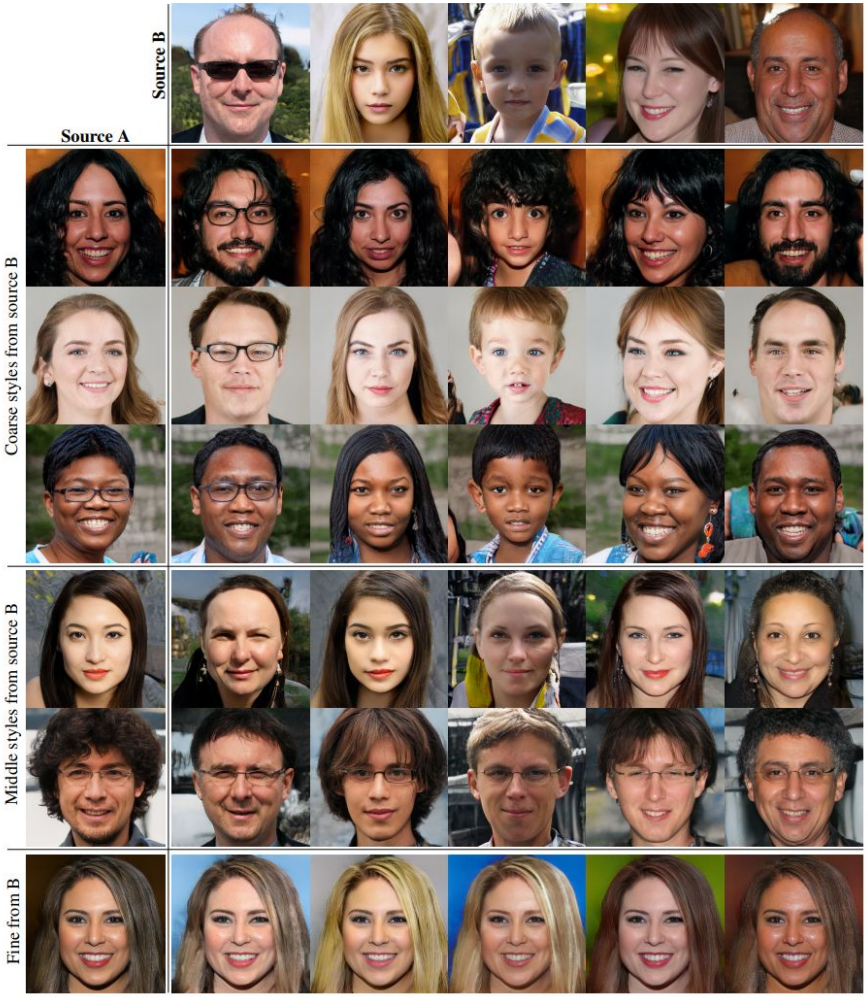
Overview: Text-to-Image Generation Models

- **Rapid Development**:
  - Significant progress in deep learning has enabled the generation of high-quality, semantically rich images from textual descriptions.

- **Applications**:
  - Advertising
  - Creative Contents
  - Virtual Reality

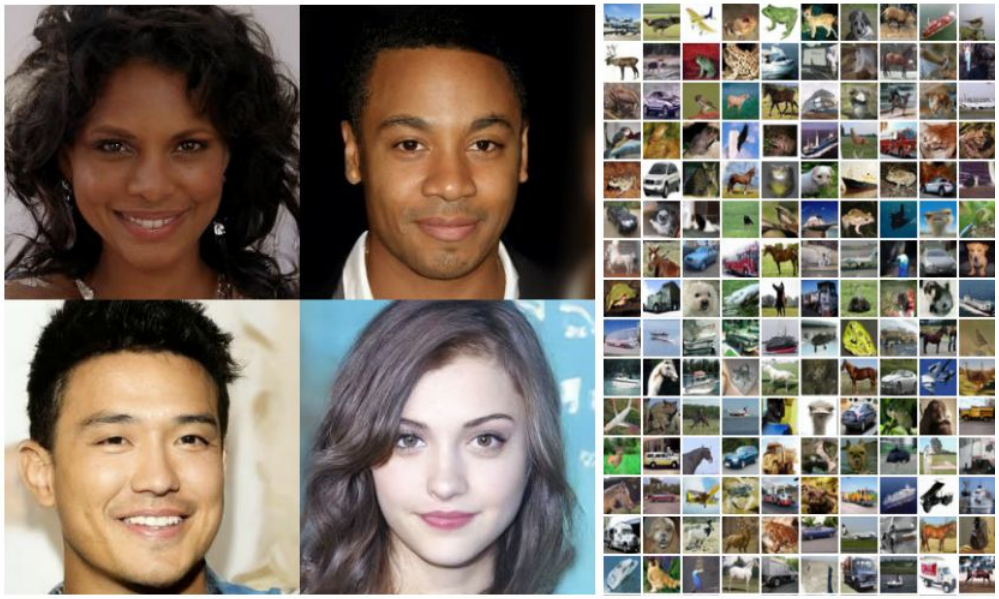
Advances in Core Technologies
Generative Adversarial Networks (GANs)

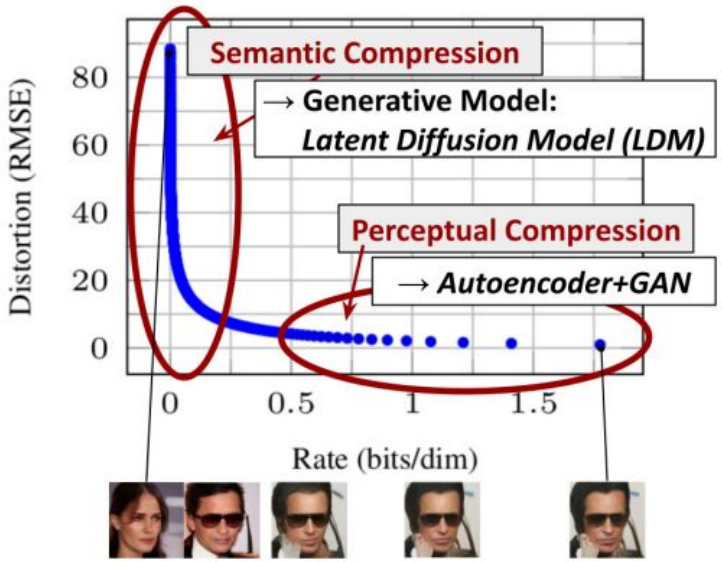
- **Representative Model**: StyleGAN [1]
- **Strengths:**
  - High-resolution image generation
- **Limitations:**
  - Struggles with semantic understanding, diversity, and consistency in complex scenarios.



Advances in Core Technologies: Diffusion Models

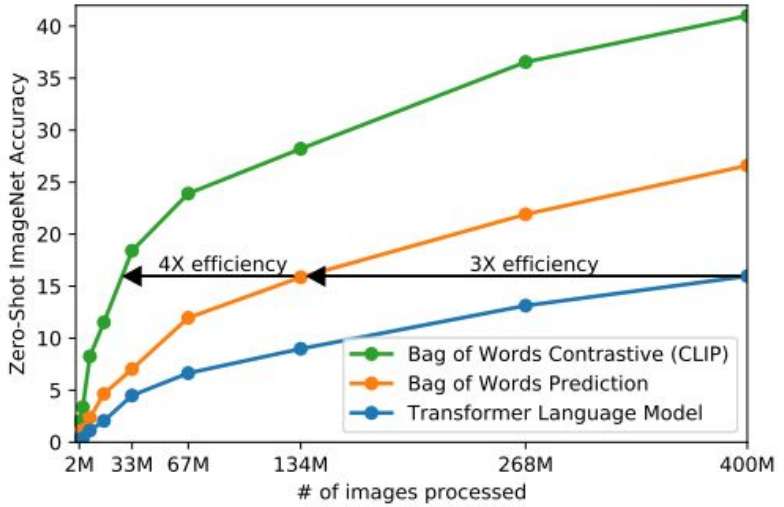
- **Representative Model**: DDPM [2] and Stable Diffusion [3]
- **Key Features**:
  - Stepwise denoising process generates high-quality images.
  - Seamlessly integrates semantic information for diverse, realistic outputs.





Advances in Core Technologies: Multimodal Models

- **CLIP [4]:**
  - Bridges text and image spaces for multimodal tasks.
- **Imagen [5]:**
  - Combines language models with diffusion models for semantically rich image generation.



A brain riding a rocketship heading towards the moon.

A dragon fruit wearing karate belt in the snow.

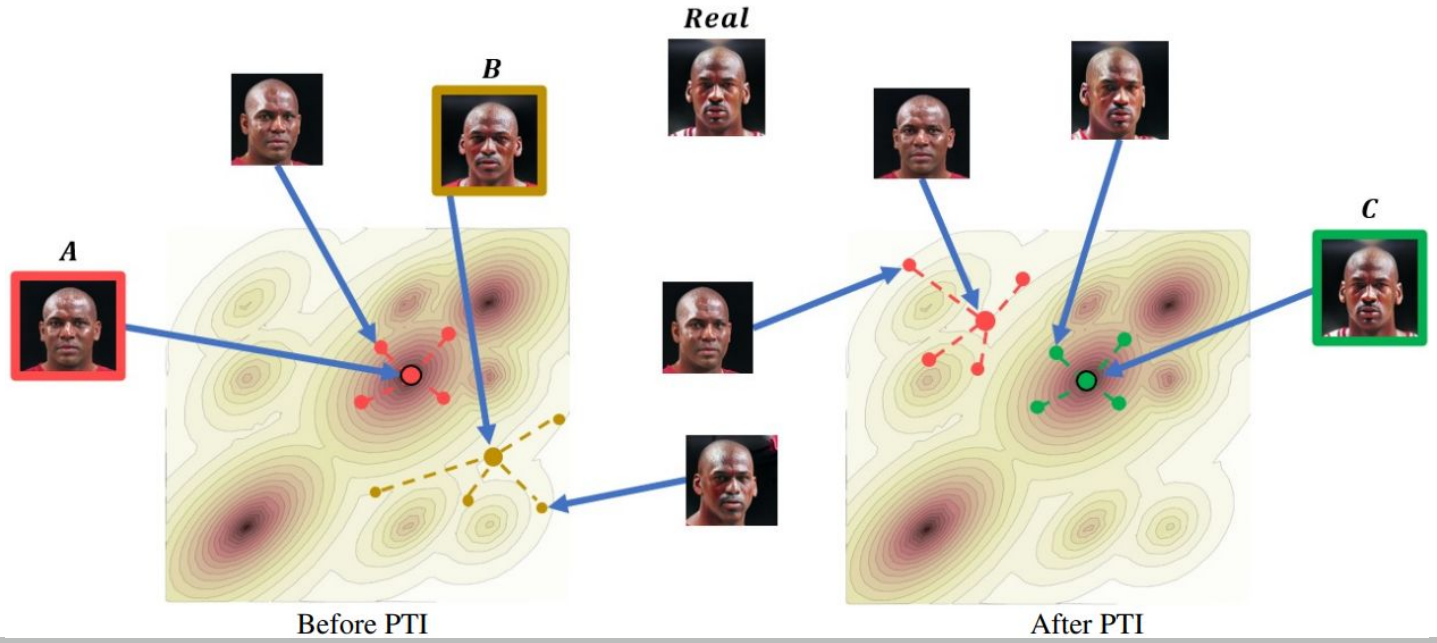
A strawberry mug filled with white sesame seeds. The mug is floating in a dark chocolate sea.

Limitations of Existing Methods: Subject Consistency Issues

- Models lack the ability to retain memory of specific subjects.
- **Example:** Fails to generate images identical to the reference subject, even with detailed text prompts.
- **Textual Inversion [6]:**
  - Attempts at personalization, but struggles with preserving subject fidelity.

Limitations of Existing Methods: Data Requirements and Generalization

- **Pivotal Tuning [7]:**
  - Performs well for face personalization but requires large datasets and is limited to specific domains (e.g., faces).
- GANs and diffusion models often overfit or experience mode collapse in few-shot (limited data) settings.



Why DreamBooth?

- **Designed for Personalized Generation**
  - Embeds specific subjects with only 3-5 images.
- **Builds on Pre-Trained Diffusion Models**
  - Leverages existing semantic knowledge for flexibility and generalization.
- **Preserves Subject and Diversity**
  - Maintains subject identity across contexts while avoiding overfitting.
- **Supports Research Goals**
  - Enables experiments on parameter effects

Project Directions

- **Goals:**
  - Use diverse portrait photos to train the model and analyze the impact of parameter variations on generated outputs.
- **Key Steps:**
  - **Model Training**
    - Fine-tune DreamBooth using different sets of portrait photos.
  - **Parameter Variation Analysis**
  - **Evaluation**

Methodology

What is DreamBooth?

- DreamBooth is a fine-tuning method
- It can generate a myriad of images of the subject in different contexts
- This model takes as input a few images (typically 3-5 images suffice) of a subject

Methodology

- **Key methods:**
  - Text-to-Image Diffusion Models and Personalization
  - Personalization of Text-to-Image Models
  - Class-specific Prior Preservation Loss
- **Key points:**
  - Fine-tuning text-to-image diffusion models.
  - Binding rare identifiers to specific subjects.
  - Prior-preservation loss to address language drift.

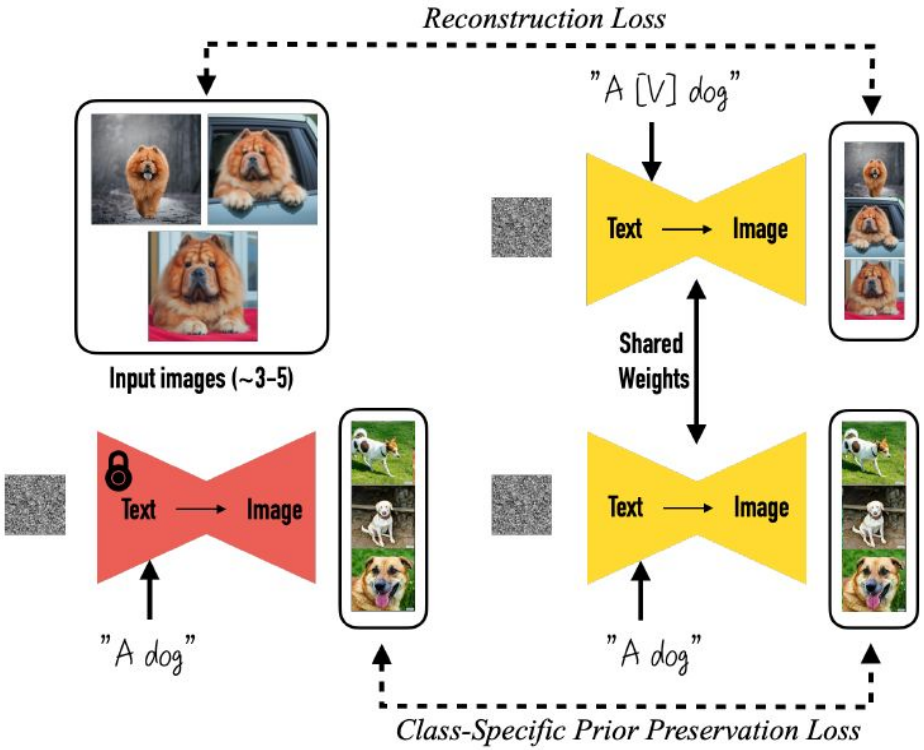
Methodology:
Text-to-Image Diffusion Models and Personalization

- Diffusion models are probabilistic generative models that are trained to learn a data distribution by the gradual denoising of a variable sampled from a Gaussian distribution.
- **Training Goal:**
  - Minimize the denoising error to improve fidelity.
- **Input and Output:**
  - Input: Initial noise image and text prompt
  - Output: Generated image

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},t}\left[w_t \|\hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}, \mathbf{c}) - \mathbf{x}\|_2^2\right]$$

Methodology:
Personalization Text-to-Image Diffusion Models

- **Natural Idea:**
  - Fine-tune the model with a few subject images and a text prompt containing a rare identifier.
- **Author's method:**
  - Allows the model to retain some prior knowledge and then overfit to this small dataset.
  - e.g., [V] dog represents a unique identifier for a specific dog
- **Key tech:**
  - Leveraging Prior Knowledge
    - e.g., "A dog" for A [V] dog
  - Rare Token Selection
    - Avoid conflicts with pre-existing semantics



Methodology:
Class-specific Prior Preservation Loss

- Outputs are overly similar to training images without Prior-Preservation Loss.

- **Training Goal:**
  - Overcome language drift and reduced output diversity

- **Solution**:
  - Use generated samples from the pre-trained model as supervision to retain class prior.
  - Include a prior-preservation term in the loss function to balance subject fidelity and diversity.

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},\boldsymbol{\epsilon}',t}\big[w_t \|\hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}, \mathbf{c}) - \mathbf{x}\|_2^2 + \lambda w_{t'} \|\hat{\mathbf{x}}_\theta(\alpha_{t'} \mathbf{x}_{\text{pr}} + \sigma_{t'} \boldsymbol{\epsilon}', \mathbf{c}_{\text{pr}}) - \mathbf{x}_{\text{pr}}\|_2^2\big]$$

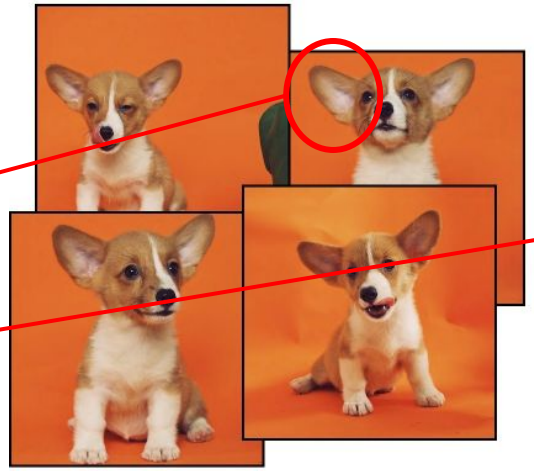
Experiments

Experiments

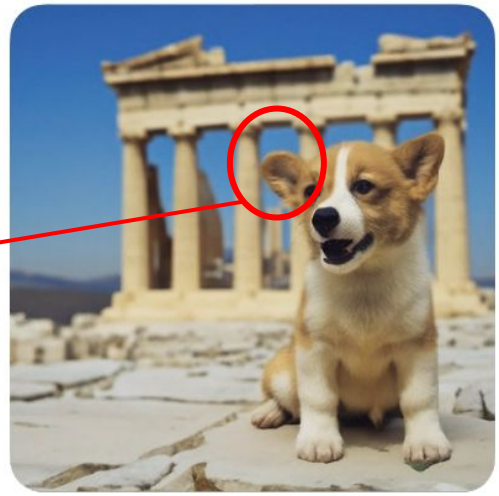
- Fine-tuning faces on Stable Diffusion v1.5 with DreamBooth.
- Original work fine-tuned with objects while ours fine-tuned with faces:
  - Original: 4-6 images per subject, non-human.
  - Ours: requires more images per subject and more training steps to generated acceptable results.

MUCH LONGER!!!

Corgi: that's not me



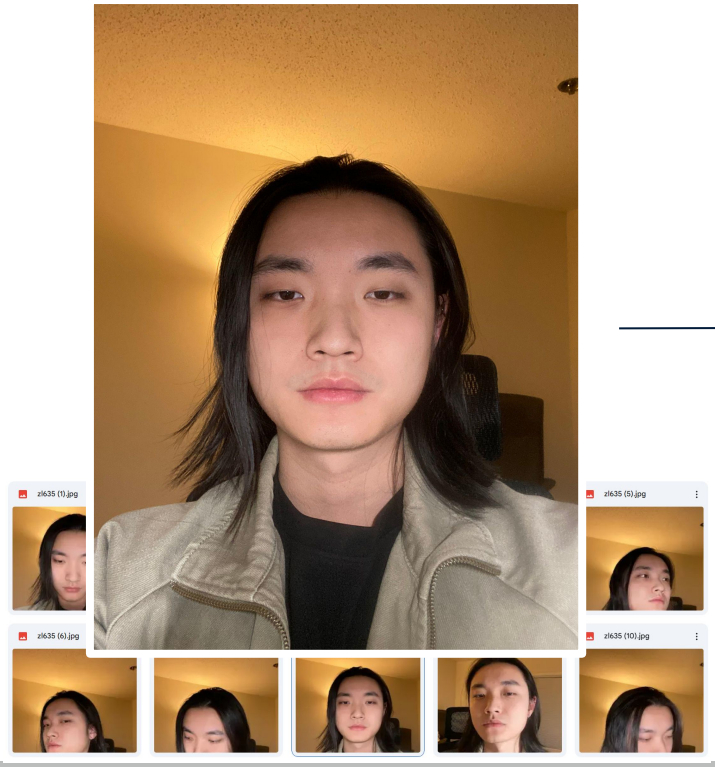
Input images



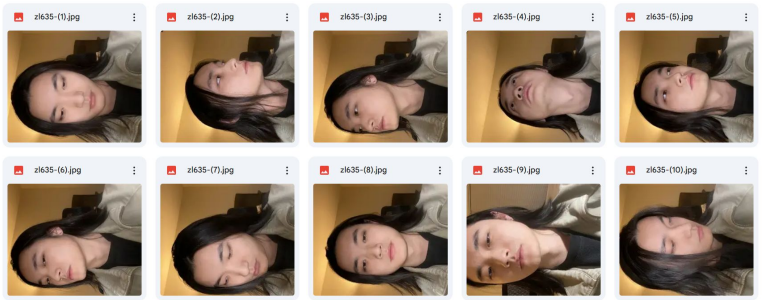
in the Acropolis

Experiments

- 1st attempt, naming: zl635 (GU ID)
- Unprocessed training images, used Smartcrop to crop images to 512*512.
- 10 instances, 100 training steps per instance, 1000 total steps, LR = 2e-6.



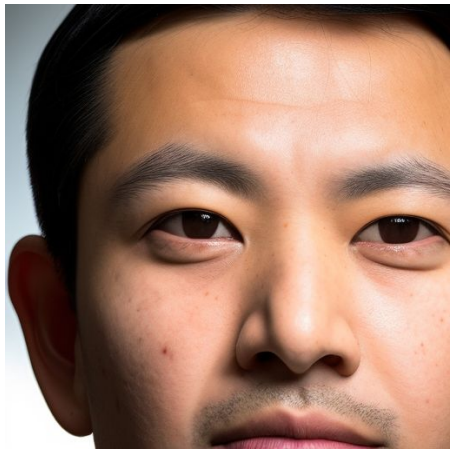
Smartcrop

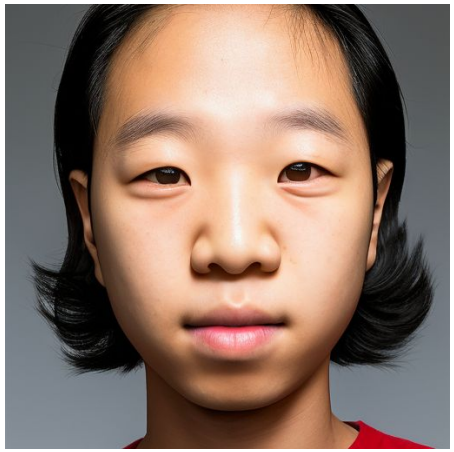
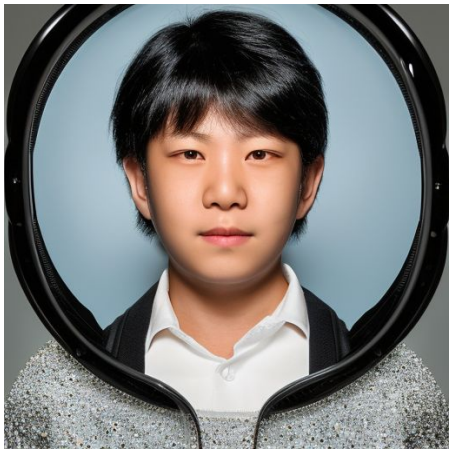
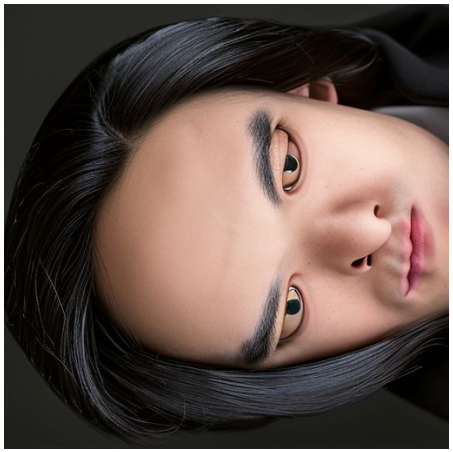


For some reason, images were rotated after applying Smartcrop.

Experiments

- Horrible results!

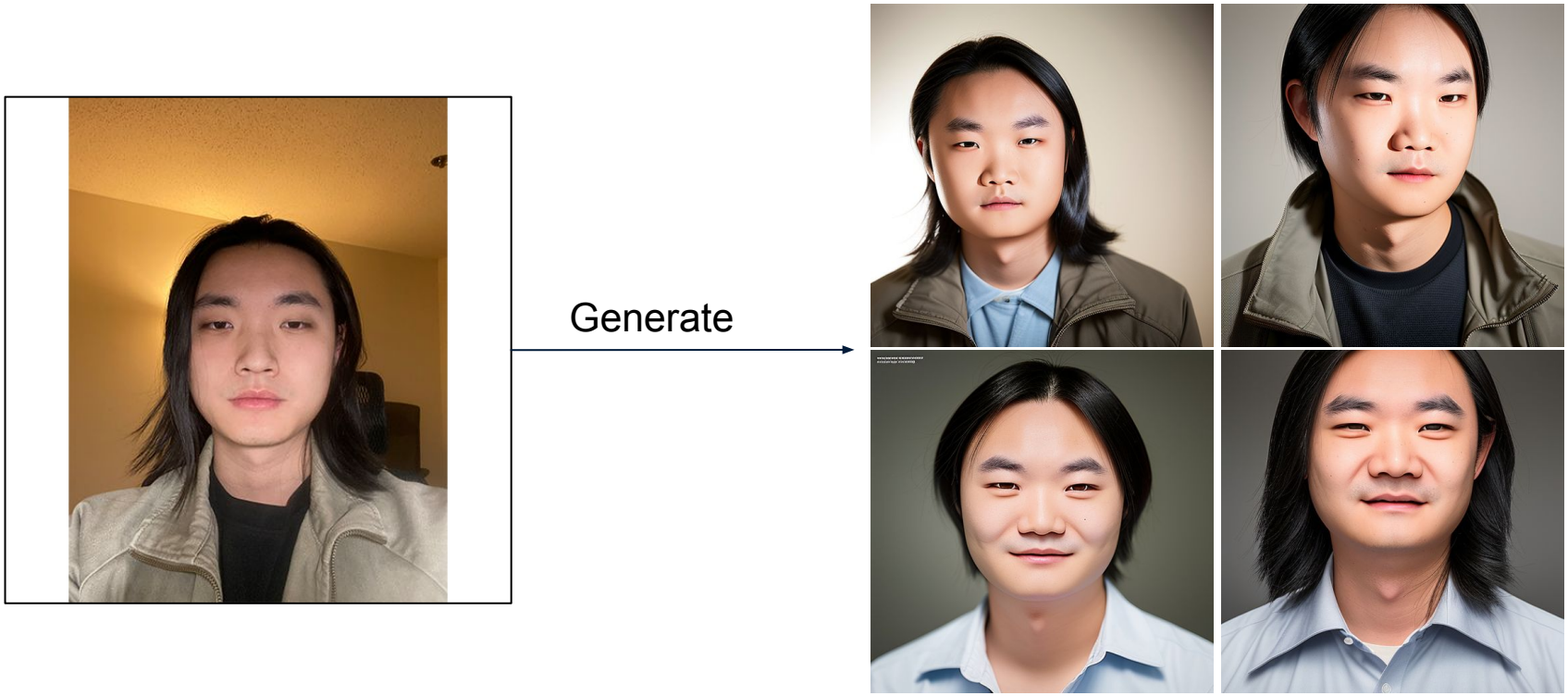


???

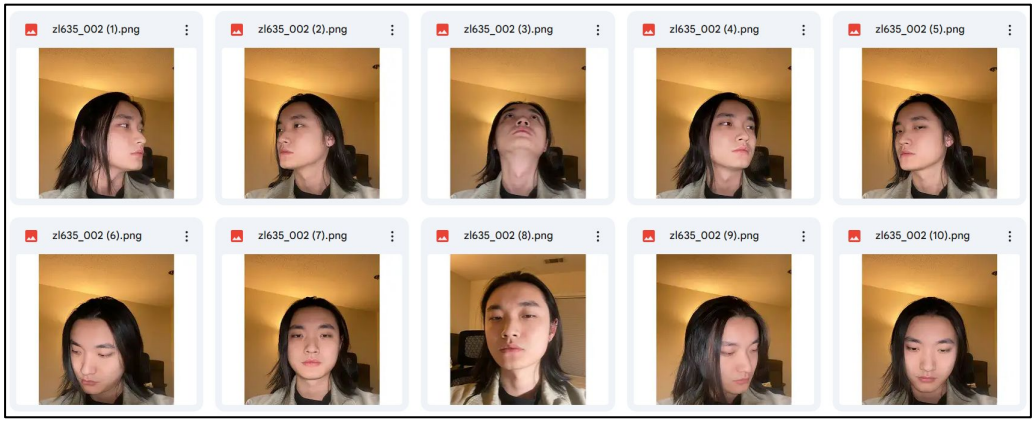
Experiments

- Second attempt: zl635_002
- Preprocessed data by adding white borders and manually set to 512*512, not rotated this time
- Increased steps per instance to 150 (from 100).

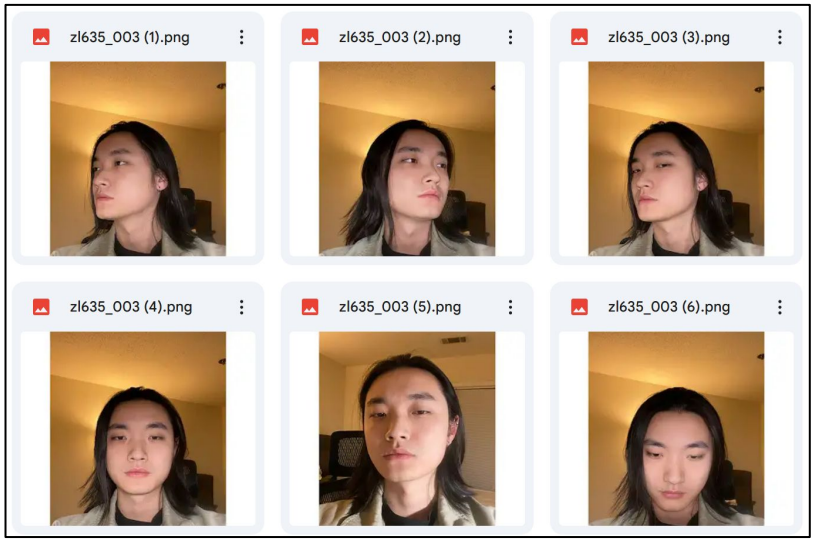


Experiments

- zl635_003 and zl635_004
- Removed images where face is turned too much. 10 instances to 6 instances.



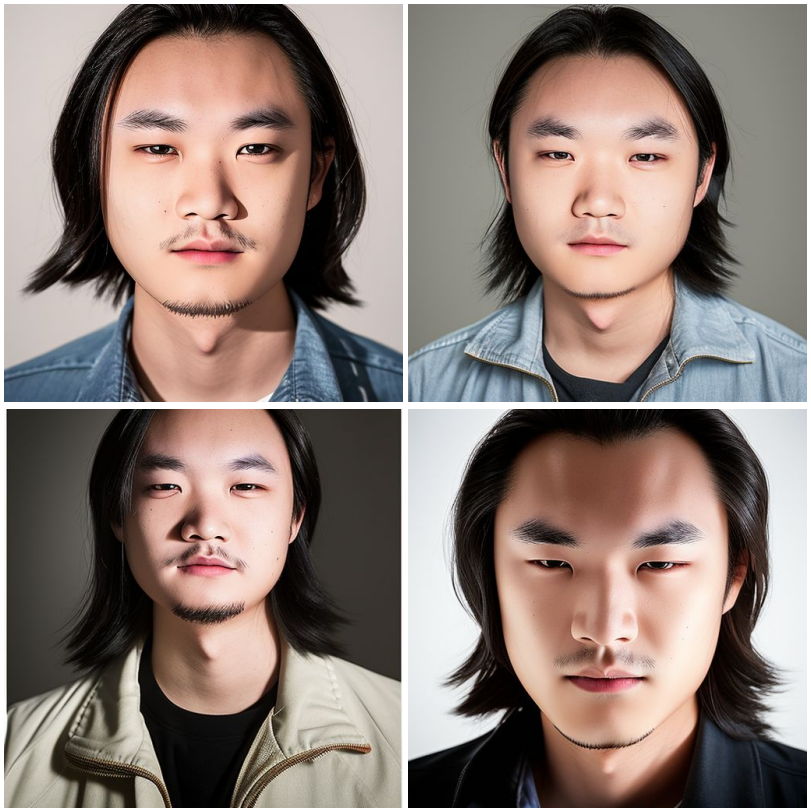
Before



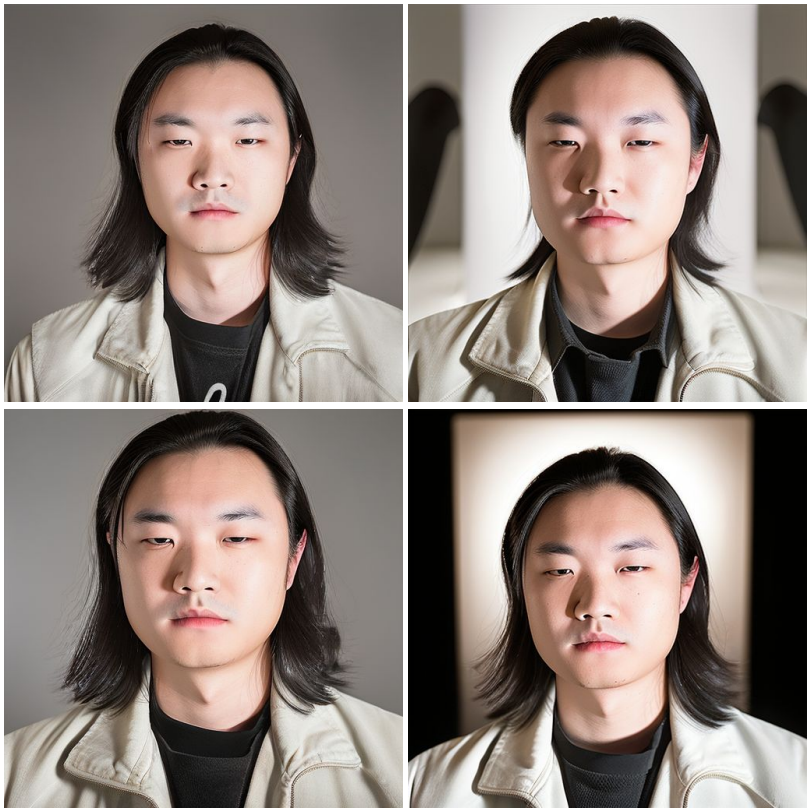
After

Experiments

zl635_003: 150 steps per instance

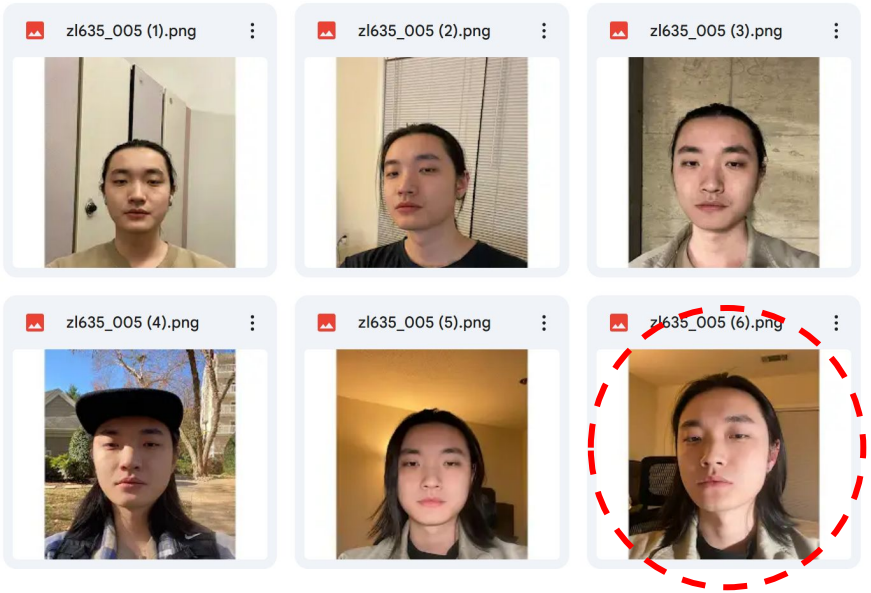


zl635_004: 500 steps per instance



Experiments

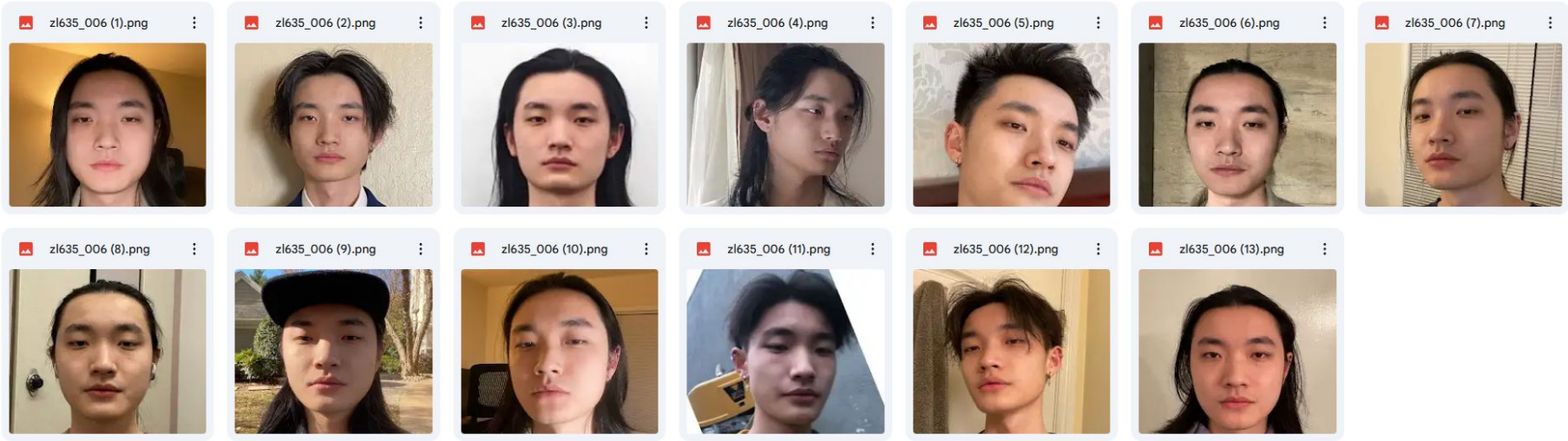
- zl635_005
- Replaced some images to increase variance.
- 1000 steps per instance.



Overfit?

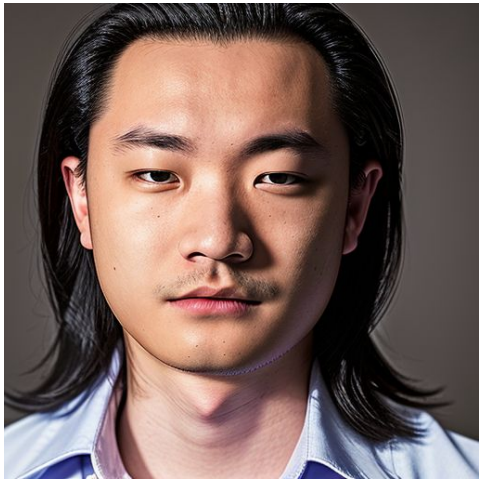
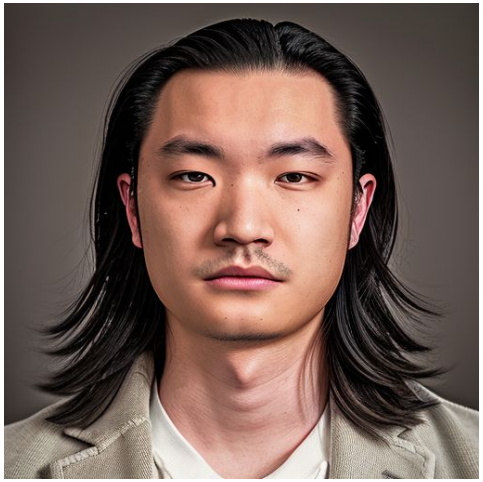
Experiments

- zl635_006
- 13 instances. More variance.
- Cropped to only head and neck.
- 300 steps per instance.



Experiments

- Starting to generate some better results.

Experiments

- zl635_007
- Changed LR from 2e-6 to 2e-7. 900 steps per instance.
- Results were not desirable.

Experiments

- zl635_008
- LR changed back to 2e-6.
- 500 steps per instance.

Experiments

- zl635_009
- 700 steps per instance.
- Not as good as zl635_008, maybe overfitting?

Experiments

- **Best model: zl635_008**
- 13 instances with some variance.
- LR = 2e-6 and 500 steps per instance.

- **More** training data (13) than original work (4-6) and the right LR and steps.

Experiments

With the success of the 1st subject, we moved on to fine-tune a 2nd model to see if we can replicate the success, and to evaluate the effects of training data and hyperparameters.

Evaluation

Evaluation

- Perceptual Similarity (LPIPS): measure the perceptual similarity between two images. (**the smaller, the better**)

- FID (Frechet Inception Distance): evaluate the quality of generated images by comparing their feature distributions with those of real images. (**the smaller, the better**)

- use 18 original images and 20 generated images each model for evaluation.

Evaluation

	CYR_001	CYR_002	CYR_003	CYR_004	CYR_005	CYR_006	CYR_007	CYR_mix
LR	2e-6	2e-6	2e-5	5e-6	2e-6	2e-6	2e-6	2e-6
Steps	100	150	100	100	75	250	500	100
LPIPS	0.6780	0.6830	0.6044	0.6526	0.6891	0.6662	0.6590	0.6601
FID	181.207	183.073	152.9508	180.524	184.895	185.094	204.161	189.179

Red: Best Result
Orange: Second Best Result
Yellow: Thirst Best Result

Evaluation

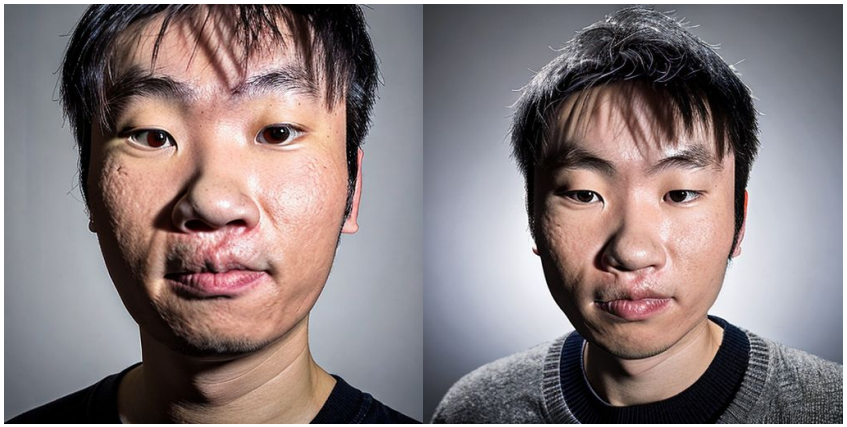
- CYR_001
- Learning Rate: 2e-6 100 steps per instance

LPIPS: 0.6780

FID: 181.2075

Evaluation

- CYR_002
- Learning Rate: 2e-6 150 steps per instance



LPIPS: 0.6830

FID: 183.0726

Evaluation

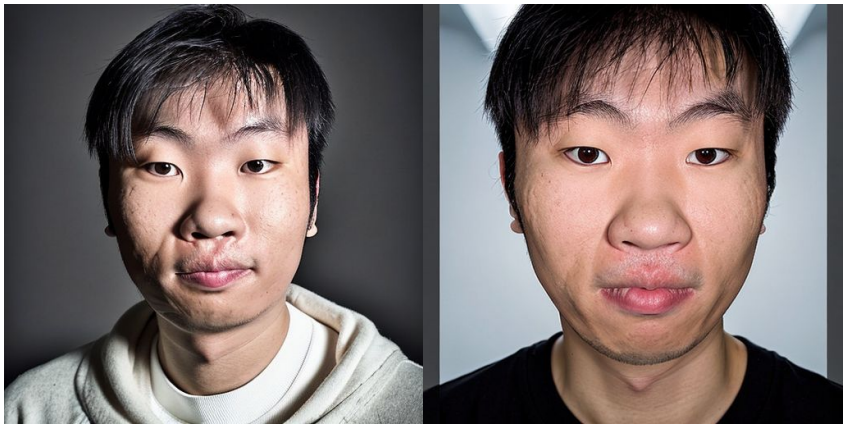
- CYR_003
- Learning Rate: 2e-5 100 steps per instance

LPIPS: 0.6044

FID: 152.9508

Evaluation

- CYR_004
- Learning Rate: 5e-6 100 steps per instance



LPIPS: 0.6526

FID: 180.5239

Evaluation

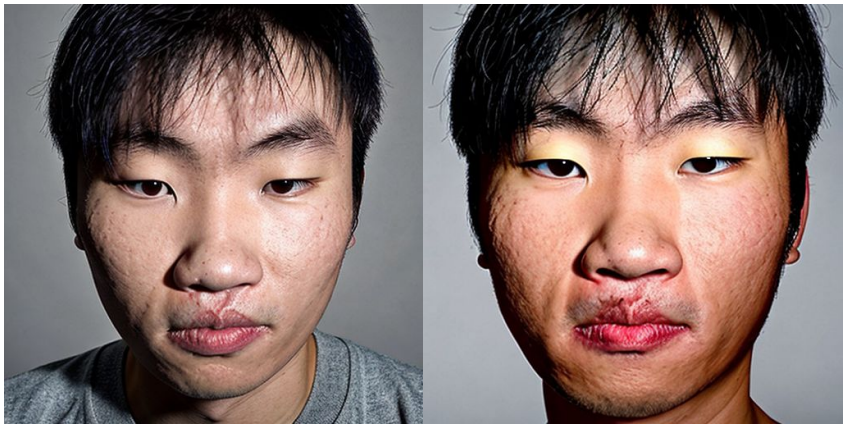
- CYR_005
- Learning Rate: 2e-6 75 steps per instance

LPIPS: 0.6891

FID: 184.8948

Evaluation

- CYR_006
- Learning Rate: 2e-6 250 steps per instance 9 instance



LPIPS: 0.6662

FID: 185.0944

Evaluation

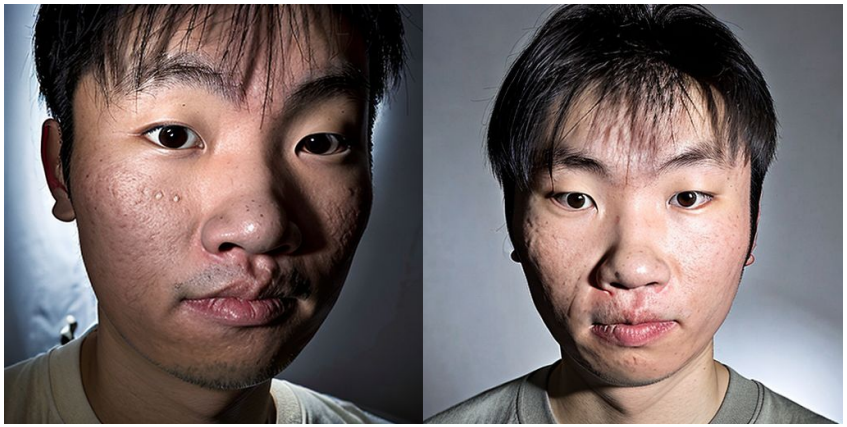
- CYR_007
- Learning Rate: 2e-6 500 steps per instance

LPIPS: 0.6590

FID: 204.1613

Evaluation

- CYR_mix
- Learning Rate: 2e-6 100 steps per instance 10 instance each face



Trained with two different face datasets

LPIPS: 0.6601

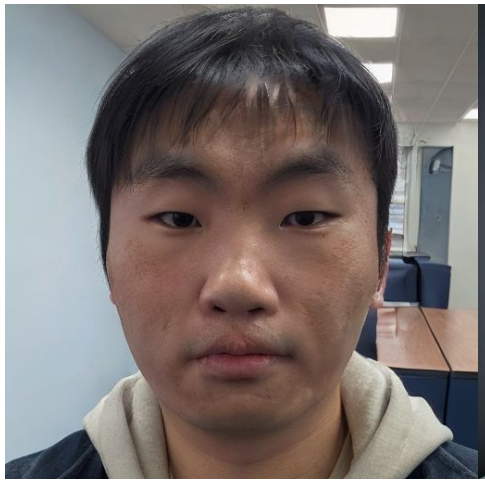
FID: 189.1789

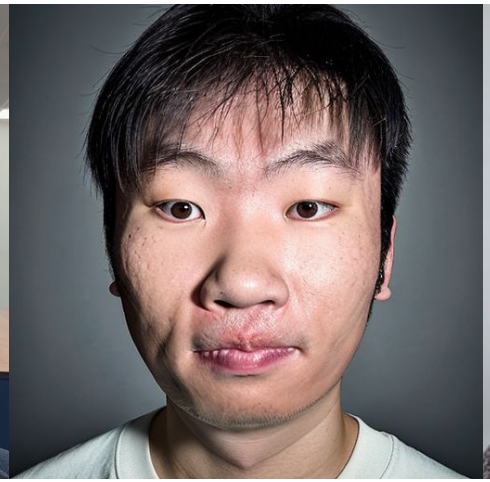
Evaluation

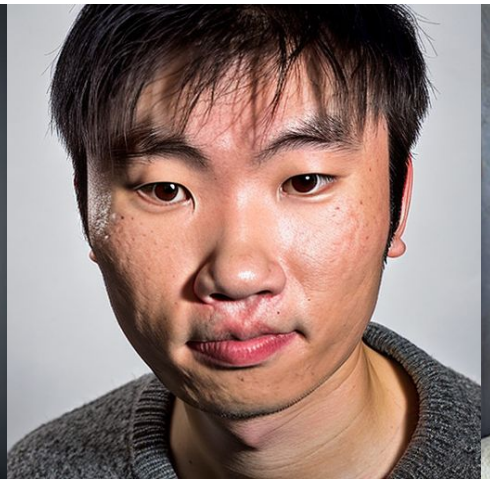
	CYR_001	CYR_002	CYR_003	CYR_004	CYR_005	CYR_006	CYR_007	CYR_mix
LR	2e-6	2e-6	2e-5	5e-6	2e-6	2e-6	2e-6	2e-6
Steps	100	150	100	100	75	250	500	100
LPIPS	0.6780	0.6830	0.6044	0.6526	0.6891	0.6662	0.6590	0.6601
FID	181.207	183.073	152.9508	180.524	184.895	185.094	204.161	189.179

Red: Best Result
Orange: Second Best Result
Yellow: Thirst Best Result

Evaluation







CYR_003	CYR_004	CYR_001	CYR_007
LPIPS: 0.6044	LPIPS: 0.6526	LPIPS: 0.6780	LPIPS: 0.6590
FID: 152.9508	FID: 180.5239	FID: 181.2075	FID: 204.1613

References

[1]. T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," *arXiv preprint arXiv:1812.04948*, Dec. 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1812.04948

[2]. J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," *arXiv preprint arXiv:2006.11239*, Jun. 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2006.11239

[3]. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *arXiv preprint arXiv:2112.10752*, Dec. 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2112.10752

[4]. A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," *arXiv preprint arXiv:2103.00020*, Mar. 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2103.00020

[5]. C. Saharia *et al.*, "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," *arXiv preprint arXiv:2205.11487*, May 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2205.11487

[6]. R. Gal, A. Alaluf, Y. Atzmon, O. Patashnik, A. Bermano, G. Chechik, and D. Cohen-Or, "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion," *arXiv preprint arXiv:2208.01618*, Aug. 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2208.01618

[7]. R. Roich, A. Mokady, A. Bermano, D. Cohen-Or, and D. Lischinski, "Pivotal Tuning for Latent-based Editing of Real Images," *arXiv preprint arXiv:2106.05744*, Jun. 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2106.05744

[8]. N. Ruiz, Y. Li, V. Jampani, D. P. W. Ellis, and A. Veit, "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," *arXiv preprint arXiv:2208.12242*, Aug. 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2208.12242


Thank You!