CILDE - Controlled Image Generation Application via LoRA-Enhanced Diffusion Models

Qiang Wang*
National University of Singapore
e1327918@u.nus.edu

Weiqi Huang*
National University of Singapore
e1327928@u.nus.edu

Abstract

With the advent of generative AI, diffusion models have emerged as a forefront technology for high-fidelity image generation, offering unprecedented realism and versatility. However, their application in creating personalized images of individuals in various scenes and styles, while maintaining accurate facial details and body postures, presents significant challenges. This paper introduces a novel solution that leverages the synergy of Low-Rank Adaptation (LoRA) fine-tuning with two specialized ControlNet models to address these challenges effectively. LoRA fine-tuning is adeptly applied to enable Stable Diffusion models to recognize and accurately render the faces of specific individuals, such as generating images of a person like Donald Trump with high fidelity. Concurrently, our two distinct ControlNet models are meticulously designed for independent yet complementary control over the image's style and the subject's posture, offering a broad spectrum of customization from professional ID photos to relaxed portraits, and precise manipulation of body language and gestures. Through our experiments, we demonstrate the effectiveness of LoRA in personalizing facial recognition within diffusion models and highlight the versatility and precision of our ControlNet models in styling and posture adjustment. Our work paves the way for a new era of personalized photography, providing a cost-effective and innovative alternative to traditional photo studios, and broadening the horizons for personalized image generation with AI.

1 Introduction

The quest for personalized image generation [Ho et al., 2020] has led to significant interest in leveraging generative AI technologies. Among these, diffusion models [Sohl-Dickstein et al., 2015] have shown promise due to their ability to generate images of remarkable quality and realism. A critical challenge, however, is achieving precise control over the generated images, particularly in rendering specific individuals with accurate facial details and customizable styles and postures. This project introduces an innovative solution to this challenge, combining the strengths of Low-Rank Adaptation (LoRA) and ControlNet [Zhang et al., 2023] models within a unified framework.

LoRA fine-tuning is applied to Stable Diffusion models Rombach et al. [2022], enabling them to recognize and accurately render the faces of designated individuals. This capability is crucial for personalized image generation, where the goal is to produce images of a specific person under various scenarios. Meanwhile, to address the need for versatile style and posture customization, we develop two separate ControlNet models. One ControlNet model is dedicated to style control, allowing for the generation of images in a range of styles from formal ID photos to casual portraits and beyond. The other ControlNet model focuses on posture control, ensuring that the generated images accurately reflect the desired body positions and movements. These models can operate both independently and in tandem, offering unparalleled flexibility in image generation.

Preprint. Under review.

Our report delves into a comparative analysis of fine-tuning methods for Stable Diffusion, highlighting the unique advantages of LoRA for personalized facial rendering. Additionally, it explores the capabilities and innovations of our style and posture control methods. By offering a scalable and efficient alternative to traditional photography services, this project has the potential to revolutionize how personalized images are generated and utilized across various domains.

2 Related work

2.1 Diffusion models

Diffusion models represent a class of advanced generative models that have made significant strides in image generation tasks. These models operate by initializing with random Gaussian noise and iteratively refining this noise into coherent images through a reversed Markov process. Originally, this process involves a forward phase that gradually converts an image to noise, which is then reversed in the generative phase.

To enhance efficiency and reduce computational demands, the Latent Diffusion Model [Rombach et al., 2022] utilizes a pre-trained autoencoder [Kingma and Welling, 2013, Van Den Oord et al., 2017] to encode images into a lower-resolution latent space, where the denoising process then takes place. This approach has been exemplified by the large-scale implementation known as Stable Diffusion, which underpins various applications including ControlNet , Editanything [Gao et al., 2023], and others.

2.2 LoRA Fine Tuning

Low-Rank Adaptation (LoRA) [Hu et al., 2021] is a technique designed to efficiently fine-tune large pre-trained machine learning models. In the realm of artificial intelligence, especially in areas involving large models such as transformers used in natural language processing, the conventional training or fine-tuning process can be resource-intensive. LoRA addresses this challenge by introducing low-rank matrices that modify the behavior of the existing weight matrices within a model.

This approach is notable for its efficiency, as it allows for the selective updating of model parameters rather than retraining the entire network. By focusing on a small subset of adaptable parameters, LoRA [Hu et al., 2021] significantly reduces the computational cost and memory usage during fine-tuning. This method is particularly advantageous when adapting large models to specific tasks or smaller datasets where overfitting or computational constraints are concerns. By leveraging LoRA [Hu et al., 2021], practitioners can achieve high levels of customization and performance without the overhead typically associated with large-scale model training.

2.3 Conditional Control of Image Generation Process

ControlNet [Zhang et al., 2023] is an advanced neural network architecture designed for precise control over specific attributes in generative tasks. It is particularly useful in scenarios where fine-grained manipulation of image characteristics, such as pose, expression, and lighting, is required. ControlNet integrates a control mechanism that allows users to specify and adjust these attributes directly, enabling more targeted and customizable outputs. This capability makes ControlNet ideal for applications in computer graphics, animation, and augmented reality, where detailed and controlled image synthesis is crucial.

2.4 Other Similar Techniques/Products

2.4.1 Dreambooth

DreamBooth [Ruiz et al., 2023] is a training technique within the Stable Diffusion framework that fine-tunes the entire diffusion model using just a few images of a specific subject or style. It operates by associating a unique prompt token with the example images, allowing the model to generate new images that maintain the stylistic and structural characteristics of the training samples. While DreamBooth excels in creating personalized content from minimal input, we chose to implement LoRA for our project due to concerns about DreamBooth's higher computational requirements and longer training times. Additionally, DreamBooth's susceptibility to overfitting, especially with limited

and unevenly distributed datasets, posed a challenge to our goal of efficiently generating controlled images from a small number of personal photos.

2.4.2 Roop

Commonly available face swap applications, such as Roop, typically perform direct face replacements on existing images, lacking the flexibility to create new images of a person in various poses. These applications replace faces on pre-existing images and do not support the generation of new images with arbitrary poses of specific individuals. Developing a model that requires only a few input photos to generate versatile images of a person holds significant practical value, addressing the limitations of current face swap technologies.

3 Methodology

3.1 LoRA for Person Recognition

To enable the Stable Diffusion model to recognize specific individuals, we adopted the Low-Rank Adaptation (LoRA) fine-tuning approach. This section details how we utilized LoRA to fine-tune the Stable Diffusion model for the purpose of recognizing distinct persons.

3.1.1 Methodology of LoRA

The main idea of Lora is shown in the Figure 1.

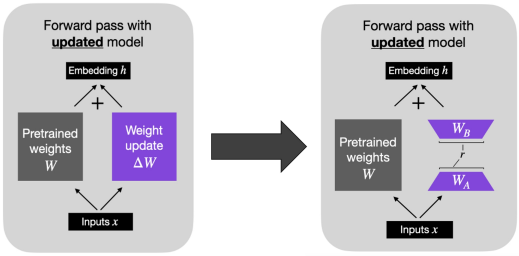


Figure 1: LoRA's structure diagram

The original weight of the model is W. We freeze the original weight of the model W in the process of updating the model, and use ΔW as the update weight to update the model. The size of ΔW matrix is the same as the original weight W, assuming that it is a very large matrix of d times k size, directly updating ΔW will cause a lot of trouble, and can not reduce the difficulty of training. Therefore, LoRA introduced an innovative method to update the model by splitting ΔW into a low-rank matrix W_A multiplied by W_B (W_B simply to keep the overall weight size consistent with ΔW). LoRA can update its weights in the following ways:

$$\Delta W = W_A \cdot W_B$$

$$h = W \cdot x + \Delta W \cdot x = W \cdot x + W_A \cdot W_B \cdot x$$

The original weight $W \in \mathbb{R}^{d \times k}$. The weight $W_A \in \mathbb{R}^{d \times r}$. The weight $W_B \in \mathbb{R}^{r \times k}$.(Note: $r \ll \min(d, k)$.) Instead of storing the intermediate results and the reverse gradient of the original weight W, we only need to store the intermediate results of W_A and W_B and the reverse gradient of W_A and W_B in order to update model. Because $r \ll \min(d, k)$, the use of LoRA can bring a huge improvement in performance, and the training time of model fine-tuning using LoRA is greatly reduced compared with full-parameter fine-tuning, and the requirements for performance are lower. And the training weights are smaller because the original model is frozen and we inject a new trainable layer that can save the weights of the new layer as a file about 3MB in size, which is nearly a thousand times smaller than the original size of the UNet model.

3.1.2 Training Dataset of LoRA

To train a LoRA model effectively, a dataset comprising 10-15 high-resolution photos of a particular individual is required. For demonstration, we used the National University of Singapore (NUS) mascot, LiNUS, as our subject. The training dataset included 15 diverse images of LiNUS, captured from various angles. Accompanying these images were 15 corresponding prompts, each containing the keyword "LiNUS," which helps the model gradually associate this keyword with LiNUS's images over time.

3.1.3 Training Process of LoRA

Once the dataset is prepared, each image undergoes a normalization process to ensure that key features are preserved while maintaining uniformity in size. Additionally, a text document containing descriptive prompts for each image is prepared; these prompts serve as textual guidance during the training phase. The training utilizes the official Stable Diffusion version 1.5 model as its foundation.

3.2 Style and Posture Control

We employed ControlNet to manage both the style and posture of the images of specific individuals. This section outlines our methods for controlling these aspects effectively.

Style Control To precisely dictate the style of the photo (such as ID photos, portrait photos, etc.), we utilized pidinet edges to guide the image generation process. During edge generation, facial information was omitted to prevent the model from overfitting to facial details of the photo template. Given our limited image and computational resources, we fine-tuned an existing ControlNet model based on Pidinet Ilyasviel [2023b].

Posture Control For controlling the posture of the depicted individual, we integrated OpenPose. Existing methods that rely on OpenPose often inaccurately render details like fingers. To address this issue, we enhanced the model by incorporating a hand depth annotated pose model, which adds detailed hand information to OpenPose's posture points. Similarly we fine-tuned an existing ControlNet model based on OpenPose Ilyasviel [2023a] by adding the hand depth information.

In the following sections, we discuss the detailed training methodologies for both ControlNet models.

3.2.1 ControlNet Architecture

Since our two models are built on top of ControlNet. we will briefly discuss the Architecture of ControlNet first.

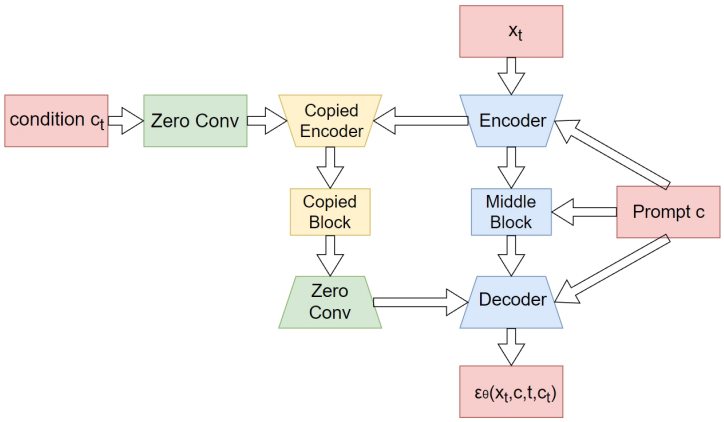


Figure 2: Main structure of ControlNet

Compared with the direct use of text prompt words to specify the pose of the image, the contour map based on edge detection can help the model deepen the understanding of the image and more accurately control the human pose of the generated image. With the inclusion of the additional

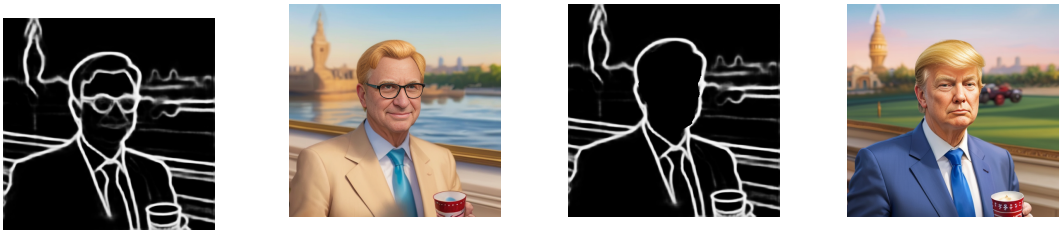
condition c_f, the distribution of target images can be expressed as $P(x, z|c, c_f) = P(x|z)P(z|c, c_f)$. By providing c and c_f, we initially encode both conditions to derive the latent variable z, followed by decoding z to generate the predicted noise for x_t.

Figure 2 illustrates the core architecture of ControlNet. This model integrates the original UNet structure from a pre-trained diffusion model alongside a secondary UNet, primarily tasked with controlling the pose and patterns of generated images. The encoder and middle blocks of the secondary UNet are inherited from the original UNet, while the decoder block comprises convolutional blocks with zero-initialized weights. To preserve the learned parameters within pre-trained models, ControlNet locks the original backbone and solely updates the parameters within the secondary UNet. The predicted noise for images at time t can be expressed as $\epsilon_\theta(x_t, \phi(c), t, \eta_\lambda(c_f))$, where λ represents the parameter of the secondary UNet. Initially, due to the zero-initialized weights of the decoder in the secondary UNet, its influence on the overall network output is minimal during the initial training stages. As training progresses, ControlNet gradually adapts the generated image while aligning with the progressively incorporated condition c_t.

3.2.2 Training Method of the Style Control Model

The style control model was designed to generate photos across various styles, such as portrait, ID, business, and casual styles. Given the scarcity of high-quality studio-level photos, we manually curated a set of 38 images representing different styles. We observed that existing edge-based models tended to overfit the head of the template. To counter this, we fine-tuned an existing ControlNet model based on edges but removed face information during training to encourage the model to learn to render faces without relying on direct edge data.

For instance, overfitting issues are evident when comparing model outputs. As illustrated, the output in Figure 3b, derived from the complete edge detection in Figure 3a, is less representative of Trump compared to the output in Figure 3d, which is based on the masked edge detection in Figure 3c. This discrepancy is primarily due to the inclusion of non-characteristic features like glasses in the first face outline, which complicates accurate facial rendering in the subsequent image.



(a) Full Edges

(b) Prompt: Donald Trump

(c) Face Masked Edges

(d) Prompt: Donald Trump

Figure 3: Select different edge plots for data comparison

Training Data for the Style Control Model The training of the style control model involved three types of data:

1. *Edges of the Original Image:* Generated through the application of the pidinet (https://github.com/hellozhuo/pidinet). The facial information of the person were removed during the training, this is critical because we only want the model to learn the style instead of face.
2. *Image-Specific Prompts:* Generated using the CLIP OpenAI [2021] model developed by OpenAI, which provided contextual prompts for each image.
3. *Original Image:* Original image were used as the target image during the training

Training Process for the Style Control Model Figure 4 illustrates the training architecture. The edges of the original image and the corresponding prompt were used as inputs, with the original image serving as the target during training.

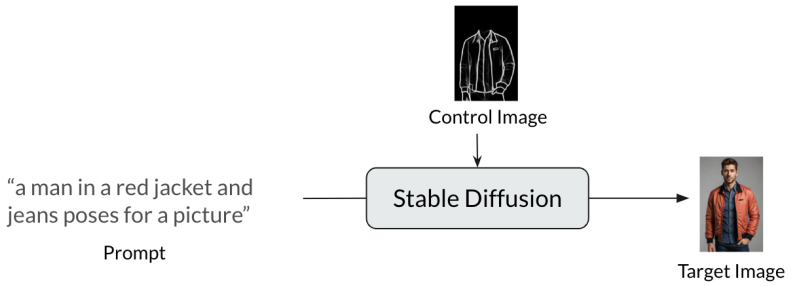


Figure 4: The training process of the style control model

3.2.3 Training Method of the Posture Control Model

The COCO dataset Lin et al. [2015] was used to train our posture control model. The COCO2017 contains 64115 real-world images categorized as humans for training and 2693 human images for validation, which has sufficient human postures and make it a good fit for training this model. We remove the low resolution and low detail images and filter about 11000 images from training. Below is the detailed features we extracted out from this dataset:

1. *Depth Map of the Hand:* The hand's depth was ascertained using the MeshGraphormer depth reconstruction library.
2. *Posture Points of the Image:* The posture data was extracted from the original image using the OpenPose library, which identifies human pose.
3. *Image-Specific Prompts:* Similar to the segmented depth model, the CLIP model from OpenAI was utilized to create relevant prompts for each image.
4. *Original Image:* Original image were used as the target image during the training

Both 1 and 2 are then merged as a single controlled image during training.

4 Results

4.1 Fine-tune Stable Diffusion for Face Rendering of Specific Individuals

The LoRA model starts to recongize the NUS mascot LiNUS after about 4000 steps of training. Figure 5 is the LiNUS generated by the LoRA fine-tuned model after 6000 steps of training. One thing to note is that the "NUS" characters are clearly rendered by our model, which proves the effectiveness of the LoRA model, as we know that it's typically difficult for stable diffusion to render characters correctly.

(a) LiNUS at beach (b) LiNUS on grass (c) LiNUS doing ski

Figure 5: LiNUS mascot generated under specific scene using different prompts

4.1.1 Alternative Methods

In addition to our primary methodology, we conducted evaluations using alternative techniques such as DreamBooth. Notably, our LoRA model is significantly more compact, with a file size of 145 MB, compared to the 2.1 GB required by the DreamBooth model to learn the same individual's characteristics. Furthermore, the training duration for DreamBooth is approximately three times longer than that for our LoRA model. While DreamBooth may offer superior facial detail with less training data, we contend that LoRA is more suited to our practical requirements. Specifically, in a deployment scenario where application efficiency is critical, the substantial storage savings offered by LoRA—avoiding an excess of 2 GB per user on our servers—is a decisive advantage.

4.2 Style Control

Figure 6 shows our fine-tuned ControlNet model for generating an ID photo by providing a suit as the template.

(a) Style Image Provided

(b) The Edge detected

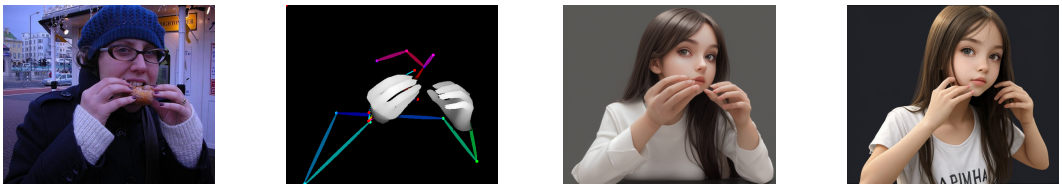
(c) Image generated by existing ControlNet

(d) Our Fine-tuned Model Output

Figure 6: Comparison between existing ControllNet and our fine-tuned model

Figure 6c presents an image generated by the current PIDiNet-based ControlNet model Ilyasviel [2023b], using the input from Figure 6b. Conversely, Figure 6d displays an image produced by our fine-tuned model based on the same input, the improvements are evident in the more natural connection between the neck and head. This highlights the effectiveness of our fine-tuning approach.

4.3 Posture control



(a) Input reference image.

(b) Depth annotated pose image.

(c) Image generated by existing ControlNet.

(d) Image generated by our model.

Figure 7: Examples of depth annotated pose generated by our fine-tuned model.

Figure 7 showcases an example from our Depth Annotated Poseture Control model, illustrating its capability to accurately generate images with the desired posture, where the hands and arms are seamlessly connected, unlike the outcomes from existing approaches.

Current methods for simulating human posture with accurate hand shapes typically rely on a two-step process: a posture control model is applied first, followed by a separate depth model for the hands. This approach often results in a noticeable disconnection between the hands and arms, a consequence of the two models being developed and trained independently, which can lead to inconsistencies when they are used together. Figure 7c displays the results of such a method, clearly showing the unnatural separation between the hand and arm regions.

5 Conclusion

In this project, we have successfully deployed the Low-Rank Adaptation (LoRA) fine-tuning method alongside two specialized ControlNet models to enhance personalized image generation. Our approach has improved the generation of stylized and posture-specific images while maintaining high fidelity and contextual accuracy.

We have shown that the LoRA fine-tuned model has a substantial reduction in model size—145 MB compared to the 2+ GB typically required by other methods like Dreambooth fine-tuning, while it can achieve similar results and accurately renders the figure of specifical individuals.

For style control, our fine-tuned ControlNet model has outperformed existing models by producing more stable and precise ID photos and styled images. The benefits of this model include enhanced adaptation to specific style prompts and improved stability in image generation, making it a valuable tool for creating personalized visual content.

In terms of posture control, our Depth Annotated Pose Control model has addressed common issues found in current methodologies, such as the unnatural separation between hands and arms. By integrating depth annotations directly within the training process, our model ensures a seamless and realistic representation of human postures.

Collectively, these advancements underscore our project's contribution to reducing the gap between AI capabilities and practical applications in personalized media creation. Looking forward, these developments lay a solid foundation for further research.

6 Limitations and Future work

This project has encountered several limitations that present opportunities for future research and development. A primary concern is the realism of the generated images, particularly those depicting humans. Despite the advancements made, these images can still be readily identified as AI-generated due to their lack of authenticity. Increasing the resolution of generated images may partially mitigate this issue, enhancing the perceived realness and reducing the distinguishability from actual photographs.

Another significant limitation is the quantity and quality of the training data available. Due to constraints in manpower and computational resources, our ability to procure and preprocess large volumes of high-quality data has been limited. The training dataset used, primarily sourced from the COCO dataset, features images that are low in details, particularly in body and facial features. This deficiency has occasionally resulted in the generation of awkward postures and anatomically incorrect limb positioning. Also, fine-tuning a ControlNet model with 11,000 images for one epoch takes approximately 22 hours on our desktop machine equipped with a RTX 3080 Ti graphics card. This significant time investment limits the feasibility of extensive model training and iterative experimentation.

For future work, we propose several avenues to address these challenges. First, enhancing the dataset quality by incorporating higher-resolution images from more diverse sources such as Instagram, where users frequently share high-quality portrait photos, could improve the model's output. While resolution is an important factor, the focus should primarily be on the richness of details in the training images, which is crucial for generating more realistic and accurate depictions.

Additionally, training the network from scratch, rather than relying solely on fine-tuning existing models, may provide further improvements. This approach would allow the models to better adapt to the nuances of high-quality, detailed images, potentially overcoming some of the current limitations in realism and anatomical accuracy.

Finally, expanding computational resources or optimizing model training methods to reduce the time and resource requirements could enable more extensive training and experimentation, leading to more refined results.

References

Shanghua Gao, Zhijie Lin, Xingyu Xie, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Editanything: Empowering unparalleled flexibility in image editing and generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9414–9416, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Ilyasviel. Controlnet model based on openpose. `https://huggingface.co/lllyasviel/sd-controlnet-openpose`, 2023a. Accessed: 2024-04-29.

Ilyasviel. Controlnet model based on pidinet. `https://huggingface.co/lllyasviel/control_v11p_sd15_softedge`, 2023b. Accessed: 2024-04-29.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

OpenAI. Clip. `https://github.com/openai/CLIP`, 2021. Accessed: 2024-04-29.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

Appendix

Code Repository : `https://github.com/allenwq/CILDE`

Acknowledgments

AI tools assisted with grammar and rewording in specific sections.