MODEL COLLAPSE IN THE SELF-CONSUMING CHAIN OF DIFFUSION FINETUNING: A NOVEL PERSPECTIVE FROM QUANTITATIVE TRAIT MODELING

Youngseok Yoon, Dainong Hu, Iain Weissburg
UCSB
Santa Barbara, CA 93106, USA
{youngseok_yoon,dainong,ixw}@ucsb.edu

Yao Qin
UCSB, Google Deepmind
Santa Barbara, CA 93106, USA
yaoqin@ucsb.edu

Haewon Jeong
UCSB
Santa Barbara, CA 93106, USA
haewon@ucsb.edu

ABSTRACT

Model collapse, the severe degradation of generative models when iteratively trained on their own outputs, has gained significant attention in recent years. This paper examines Chain of Diffusion, where a pretrained text-to-image diffusion model is finetuned on its own generated images. We demonstrate that severe image quality degradation was universal and identify *CFG scale* as the key factor impacting this model collapse. Drawing on an analogy between the Chain of Diffusion and biological evolution, we then introduce a novel theoretical analysis based on quantitative trait modeling. Our theoretical analysis aligns with empirical observations of the generated images in the Chain of Diffusion. Finally, we propose Reusable Diffusion Finetuning (ReDiFine), a simple yet effective strategy inspired by genetic mutations. It operates robustly across various scenarios without requiring any hyperparameter tuning, making it a plug-and-play solution for reusable image generation.

1 INTRODUCTION

Can state-of-the-art AI models learn from their own outputs and improve themselves? As generative AI models (e.g., GPT, Diffusion) now churn out uncountable synthetic texts and images, this question piqued curiosity from many researchers in the past couple of years. While some show positive results of self-improving (Huang et al., 2022; Gerstgrasser et al., 2024), most report an undesirable *"model collapse"*—a phenomenon where a model's performance degrades when it goes through multiple cycles of training with the self-generated data (Bertrand et al., 2023; Gillman et al., 2024; Taori & Hashimoto, 2023; Shumailov et al., 2023; Dohmatob et al., 2024a; Fu et al., 2024; Marchi et al., 2024; Martínez et al., 2023b). When large language models (LLMs) are trained with their own outputs, it begins to produce low-quality text that has a lot of repetitions (Dohmatob et al., 2024c), and its linguistic diversity declines rapidly (Guo et al., 2024; Briesch et al., 2023); image models also show quality degradation (Bohacek & Farid, 2023; Martínez et al., 2023a) and loss of diversity (Alemohammad et al., 2023; Hataya et al., 2023).

The goal of this paper is to investigate model collapse in the practical scenario of finetuning pretrained text-to-image diffusion models. An end user often wants to finetune the latest model to generate images with a very specific style (e.g., creating characters in the style of Pokémon). In fact, hundreds of new finetuned diffusion models are uploaded regularly on platforms like CivitAI[1], each designed to produce different styles of images. When users scrape the Internet to collect images of the style they want, it becomes almost inevitable that synthetic images will be included in their datasets.

[1] https://civitai.com/

This is because the number of real images is limited, while synthetic images can be generated in massive quantities and dominate online sources. As a result, users will feel compelled to include more synthetic images in their datasets to keep up with the data demands of these ever-growing, data-hungry models.

To this end, we conduct a thorough investigation into how various hyperparameters commonly used during diffusion finetuning (e.g., learning rate, diffusion steps, prompts) impact model collapse. From our extensive empirical analysis, we make a crucial observation: the classifier-free guidance (CFG) scale is the most significant factor that influences the rate of model collapse. Moreover, we observe a fascinating phenomenon that the direction of degradation varies with different CFG values—low CFG results in low-frequency degradation in images, while high CFG leads to high-frequency degradation. The critical role of CFG in determining both the rate and direction of model collapse is a novel insight previously unrecognized in the literature.

To gain a deeper theoretical understanding of this phenomenon, we draw upon the concept of quantitative trait modeling from statistical genetics. Unlike existing studies that attribute model collapse to limited sample sizes and conclude it leads to zero variance (Alemohammad et al., 2023; Bertrand et al., 2023; Shumailov et al., 2023), we propose that the underlying cause is a truncation-based selection process modulated by CFG. This theoretical model accurately describes our experimental observations. Moreover, we show that other theoretical work on model collapse can also be connected to statistical genetics models, such as random genetic drift and its variants (Fisher, 1999; Wright, 1931; Kimura, 1955; Paris et al., 2019). This fresh angle of drawing parallels with statistical genetics offers a new framework to understand the mechanism of model collapse.

While we have shown the crucial role of CFG in mitigating model collapse both empirically and theoretically, searching for the optimal CFG scale over multiple finetuning iterations is computationally expensive and cumbersome to implement in practice. We propose a simple solution that bypasses CFG tuning, called Reusable Diffusion Finetuning (ReDiFine). It introduces small modifications in the finetuning and generation algorithm—condition drop and CFG scheduling—and we demonstrate that ReDiFine achieves performance comparable to the optimal CFG scale across all hyperparameter settings we tested.

Our contributions can be summarized as follow:

- We conducted a comprehensive investigation of model collapse when finetuning a diffusion model on its own outputs, testing a wide range of parameters on four datasets (two digital art and two natural images). Our results show that CFG scale is the most critical factor, governing both the rate of model collapse and the type of image degradation (Section 3).
- We provide a novel theoretical analysis of model collapse based on quantitative trait modeling that can accurately predict how power spectra of generated images evolve over iterations (Section 4).
- We propose ReDiFine, a simple yet effective strategy to achieve a near-optimal reusability-fidelity trade-off without any hyperparameter tuning. By combining condition drop finetuning and CFG scheduling, ReDiFine operates robustly across various scenarios—such as synthetic-real mixed datasets and varying initial CFG settings—across all four datasets we evaluated (Section 5).

2 Related Work

The self-consuming training loop and the associated phenomenon known as "model collapse" have become significant areas of study in the past two years (Martínez et al., 2023a;b; Taori & Hashimoto, 2023; Alemohammad et al., 2023; Bohacek & Farid, 2023; Guo et al., 2024; Bertrand et al., 2023; Dohmatob et al., 2024a; Briesch et al., 2023; Gillman et al., 2024; Fu et al., 2024; Marchi et al., 2024). Model collapse, defined as "a degenerative process affecting generations of learned generative models, where generated data end up polluting the training set of the next generation of models" in Shumailov et al. (2023), has been observed in both language and image generative models.

Empirical studies on LLMs (Briesch et al., 2023) reveal that linguistic diversity collapses, especially in high-entropy tasks (Guo et al., 2024), although this can be mitigated with data accumulation (Gerstgrasser et al., 2024). In image generation, several works (Martínez et al., 2023a;b; Alemohammad et al., 2023; Hataya et al., 2023; Bohacek & Farid, 2023; Bertrand et al., 2023) note image degradation when diffusion models are recursively trained with self-generated data. We conduct extensive

experiments to reveal the key factor causing model collapse in text-to-image diffusion models. Our findings reveal that while model collapse is universally observed across various datasets and scenarios, it manifests in three distinct types of image degradation.

Theoretical studies on model collapse largely focused on diversity reduction, typically framed as either decreasing covariance in continuous domains (Alemohammad et al., 2023; Shumailov et al., 2024; Bertrand et al., 2023) or shrinking support in discrete domains (Dohmatob et al., 2024c; Marchi et al., 2024), which aligns with our observations on the decreased recall metric. The finite number of generated samples has been identified as a key cause of model collapse (Bertrand et al., 2023; Shumailov et al., 2024; Dohmatob et al., 2024c; Fu et al., 2024), while others (Alemohammad et al., 2023; Ferbach et al., 2024; Marchi et al., 2024) emphasize that sampling bias reduces the effective distribution. In contrast, we present a novel framework based on quantitative trait modeling, providing a fresh direction based on mean drift induced by the selection process. We further demonstrate that many existing theories of model collapse can be understood through extensions of classical statistical genetics.

Most existing works echo the importance of incorporating a large proportion of real data (Alemohammad et al., 2023; Bertrand et al., 2023; Fu et al., 2024; Ferbach et al., 2024) or accumulating data over iterations (Gerstgrasser et al., 2024) to mitigate model collapse. Gillman et al. (2024) instead suggests a self-correcting self-consuming loop using an expert model, a physics simulator for human motion generation, to correct synthetic outputs. While promising, an expert model may not be feasible and too costly for many real-world applications. In our work, we propose an alternative solution through *reusable image generation*, and show that it can generate high-quality images comparable to the optimal hyperparameter without extensive search.

Several concurrent works seek to address model collapse through their theoretical insights. Dohmatob et al. (2024b) provides a comprehensive analysis of how model collapse behavior varies with model size and data composition. Feng et al. (2024) emphasizes the importance of data verification in preventing performance degradation, while Zhu et al. (2024b) introduces a novel token editing strategy for data curation, both in LLM tasks. On the other hand, Zhu et al. (2024a) and Alemohammad et al. (2024) propose model-specific solutions leveraging real data to counteract the negative effects of synthetic data in image-generative models.

3 Model Collapse in Self-Consuming Chain of Diffusion Finetuning

3.1 Problem setting & Experimental Setup

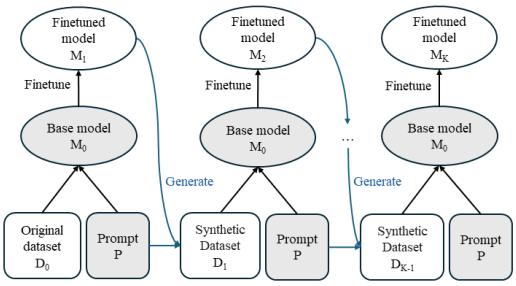
Chain of Diffusion. We begin with formally defining the self-consuming Chain of Diffusion finetuning. Given a pretrained generative model M_0, an original training image set $D_0 = \{x_{0,i} | i \in [0, N-1]\}$, and a prompt set $P = \{y_i | i \in [0, N-1]\}$, where N is the number of total images in the dataset, each image $x_{0,i}$ is paired with a corresponding text prompt y_i. M_{k+1} is a model finetuned from M_0 using the generated image set $D_k = \{x_{k,i} | i \in [0, N-1]\}$ and the prompt set P, which simulates a fully synthetic loop (Alemohammad et al., 2023). Then, M_{k+1} generates a set of images D_{k+1} for the next iteration using the prompt set P:

$$M_{k+1} = \text{Finetune}(M_0, D_k, P), \tag{1}$$

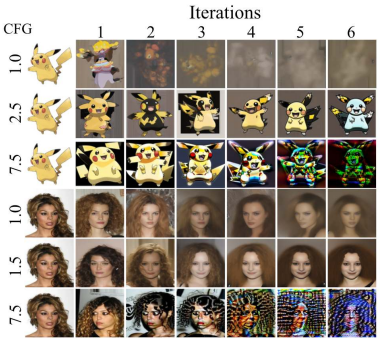
$$D_{k+1} = \text{Generate}(M_{k+1}, P). \tag{2}$$

During the Chain of Diffusion, M_0 and P are fixed across iterations, and one image is generated per one prompt to maintain the dataset size for all iterations. The overall pipeline is shown in Figure 1a. We note that this setting is consistent with Alemohammad et al. (2023); Shumailov et al. (2023); Guo et al. (2024); Fu et al. (2024); Briesch et al. (2023).

Model and datasets. We use Stable Diffusion v1.5 (Rombach et al., 2022) as the pretrained model M_0 and apply LoRA (Hu et al., 2021) to finetune M_0 at each iteration. We build our implementation on kohya-ss and experiment on four datasets: Pokemon (Pokémon, 2023), Kumapi (Ihelon, 2022), Butterfly (Veeralakrishna, 2020), and CelebA-1k (Liu et al., 2015) to investigate various domains including animation, handwriting, and real pictures. M_0 is finetuned for 100 epochs during each iteration. More details can be found in Appendix A.1 and A.2.



(a) Self-Consuming Chain of Diffusion Finetuning.



(b) Image degradation in the Chain of Diffusion.

Figure 1: (a) **Overall pipeline of the Chain of Diffusion.** Given a pretrained text-to-image diffusion model M_0 and a prompt set P, a finetuned model M_k is trained using D_{k-1} generated at the previous iteration $k-1$. Then, M_k generates D_k using the same prompt set P, building a fully synthetic loop. Chain of diffusion begins with the original real dataset D_0. (b) **Image degradation universally occurs across multiple domains, and low and high CFG scales lead to blurry images and high-frequency degradation, respectively.** As the Chain of Diffusion progresses, the severity of image degradation intensifies, which holds consistently across four datasets and 11 scenarios in Section 3.2. CFG 2.5 for Pokemon and 1.5 for CelebA exhibit an ideal middle ground where both types of degradation slow down. More images in Appendix B.

Evaluation metrics. We use Frechet Inception Distance (FID) (Heusel et al., 2017) to measure image quality. Following Stein et al. (2024), we use DiNOv2 (Oquab et al., 2023) as a feature extractor since it is more consistent with our visual inspection than Inception-V3 network (Szegedy et al., 2016). With a slight abuse of terms, we will still refer to the Frechet distance with DiNOv2 as the FID score. Additional experiments with other metrics (CLIP score (Radford et al., 2021), Sample-wise Feature Distance (SFD), and recall) are presented in Appendix E.5.

We propose a new metric to quantify the reusability of generated images. We define collapse rate as the performance degeneration per iteration in the Chain of Diffusion:

$$\text{collapse rate} = \Delta\text{FID} = \frac{\text{FID}_K - \text{FID}_1}{K-1}, \tag{3}$$

where FID_k stands for the FID between k-th iteration set and the original training set. We use FID as a performance metric here, but this can be any other performance metric of interest. Note that a low collapse rate indicates more reusable images since the model does not degrade much. We have $K = 6$ for the rest of the paper.

3.2 Model Collapse in the Chain of Diffusion

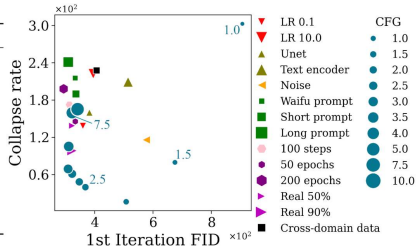
In this section, we make a series of observations regarding the model collapse behavior in the Chain of Diffusion. We conduct extensive investigations to reveal the most impactful factor in the model collapse and analyze how this factor contributes to the Chain of Diffusion.

Observation 1: Model collapse is universal in the Chain of Diffusion. We observe significant image degradation in all four datasets in the Chain of Diffusion (see Figure 1b). The quality of generated images begins to deteriorate in the third iteration and drops even more rapidly once the visible degradation emerges, reaching an unrecognizable level in two or three additional iterations. Quantitative evaluation (FID, CLIP score, SFD, recall) also indicates this quality degradation (Appendix E.5). Also we note that we observed significant drops in recall in all our experiments, indicating diversity reduction as discussed in several recent works.

[2]C, W, S, and L for Combine, Waifu, Short, and Long, respectively. We concatenate BLIP and Waifu captions as default setting, referred to as Combine. Short and Long prompts are BLIP captions generated with limitations in the lengths of captions. More details can be found in Appendix C.4.

[3]We do not display scenarios that change the training dataset size, such as Img/Prompt and Dataset Size, as varying sizes result in FID scores on different scales. The related results are presented in the Appendix C, where we observe similar image degradation.

Candidates	Default	Values
Dataset size	1k	0.5k, 2k
Learning rate	1e-4	1e-3, 1e-5
Finetune	Both	Unet, Text encoder
Noise to images	No	Yes
Prompt	C	W, S, L
CLIP skip	2	1, 3
Diffusion steps	30	100
Epochs	100	50, 200
Mixing real images	0.0	0.5, 0.9
Image per prompt	1	5
Cross-domain data	1	4



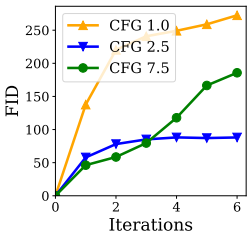


Figure 2: **Left:** Description of 11 potential factors (excluding CFG) that we examined as candidate sources for model collapse. All experiments were conducted on Pokemon except for the dataset size. For dataset size, we use CelebA since its original dataset is bigger, and we can subsample 500, 1000, and 2000 images. For prompt, we concatenate prompts with different lengths[2]. **Center:** All hyperparameter settings other than changing CFG show a high *collapse rate* greater than 1.0, i.e., FID score increases by $\sim$2x in 6 iterations, indicating severe image degradation. The x-axis represents FID_1, quantifying the generation performance at the first iteration and the y-axis represents the *collapse rate* defined in equation 3. For both FID_1 and *collapse rate*, lower is better[3]. **Right:** Quantitative comparison for FID $\downarrow$ (Pokemon). CFG 2.5 achieves the most robust performance. CFG 1.0 degrades from the beginning of the chain while CFG 7.5 begins to degrade in the third iteration, which aligns with the visual inspection in Figure 1b.

We then investigated a variety of different scenarios (summarized in Figure 2) to see if this degradation is an anomaly of specific hyperparameter settings or if it is a ubiquitous phenomenon. We tested various dataset sizes for D_0 and D_k, increasing the size of D_k by generating more than one image per prompt, mixing real images from D_0 to D_k, changing the descriptiveness of prompts, freezing U-Net or text encoder, and various other hyperparameters (# sampling steps, # epochs, learning rate, and CLIP skip). We also tested adding a small Gaussian noise in each image in the original set D_0 to see if having small random perturbations can improve reusability, and merging all four datasets into one to see whether the extended domain can delay model collapse (Cross-domain data). *In all settings we tested, image degradation was universally present and very fast.* We plot the trade-off for Pokemon on Figure 2 where y-axis is collapse rate and x-axis is the FID_1 (better when closer to the origin). While adding noise to images and mixing 90% real images to every iteration as proposed by Alemohammad et al. (2023); Bertrand et al. (2023); Fu et al. (2024); Ferbach et al. (2024); Gerstgrasser et al. (2024) show the lowest collapse rate, they still exhibit significant degradation (FID score has been doubled in 6 iterations).

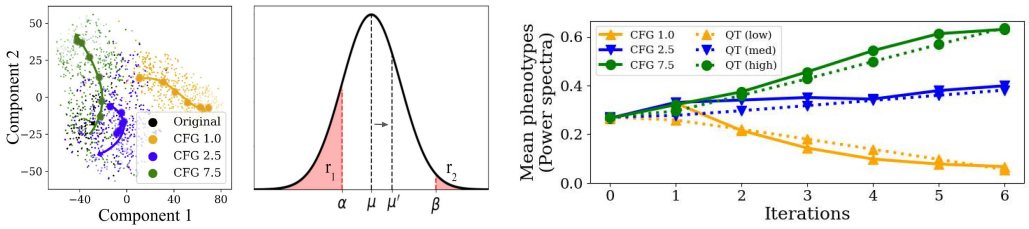
Moreover, this degradation is unavoidable even for Stable Diffusion XL (Appendix C.11). We further examined slightly different settings from the literature where all previously generated images are accumulated (data accumulation (Gerstgrasser et al., 2024), Appendix C.12) or the same model is iteratively finetuned (iterative retraining (Bertrand et al., 2023; Martínez et al., 2023a), Appendix E.4), and consistently observed model collapse.

Observation 2: CFG is the most significant factor that impacts the model collapse. Throughout all our experiments, classifier-free guidance (CFG) had the biggest impact on the speed of model collapse. CFG scale was first introduced in Ho & Salimans (2022) to modulate the balance between unconditional score (Uncond) and conditional score (Cond) at each diffusion step as follows:

$$\text{Total Score} = \text{Uncond} + \text{CFG} \cdot (\text{Cond} - \text{Uncond}). \tag{4}$$

High CFG emphasizes the conditional score for the given prompt, which pushes the generation to align better with the prompt and often leads to higher-fidelity images. On the other hand, lower CFG places less weight on the conditional score and provides more diversity in generated images. For those familiar with temperature sampling (Ackley et al., 1985), CFG plays a similar role as temperature, which adjusts the trade-off between fidelity and diversity.

Figure 2 shows that as we increase the CFG scale, the image quality in the first iteration improves (smaller FID_1 on x-axis), which is expected from our understanding of CFG. A more surprising part is that this comes at the cost of a worse collapse rate (an increase on the y-axis). Also, when the CFG scale is as high as 7.5 or 10.0, the improvement in FID_1 plateaus, and increased CFG worsens both FID_1 and collapse rate. Similarly, when the CFG scale is too low—below 2.0—the improvement in collapse rate plateaus, and both FID_1 and collapse rate begin to increase. There is an optimal region of CFG values (near 2.5, specific to Pokemon), where we achieve a low collapse rate while



(a) t-SNE of images. (b) Selection mechanism. (c) Power spectra of images vs. Mean of phenotypes.

Figure 3: (a) **t-SNE plot visualizes how generated images evolve from the original distribution (black) and shows distinct paths for three CFG scales (Pokemon).** Different CFG scales and iterations are differentiated with different colors and transparency. Arrows indicate how the distributions of generated images move for different CFG scales. While CFG 2.5 (blue) stays near the original images (black), high and low CFG scales (1.0 and 7.5) deviate fast, indicating image degradation. (b) **Selection mechanism with two-sided truncation.** r_1 and r_2 ratios of samples are truncated from the left and right tails, and the remaining $r = 1 - r_1 - r_2$ ratio of samples is selected. (c) **Power spectra for different CFGs (Pokemon) and quantitative trait modeling results align well for different selection strategies.** Directional selections with truncation can effectively explain our observations in Section 3: the behaviors of high- and low-frequency degradation and optimal CFG scale. Detailed settings are provided in Appendix D.2.

maintaining a good quality in the first iteration as well. Moreover, Figure 2 presents FID for Pokemon to demonstrate how different CFG scales affect performance. CFG scale 2.5 achieves the most robust performance for all metrics. Interestingly, optimal CFG scales differ for different styles: 2.5 for animated or hand-writing datasets (Pokemon and Kumapi) and 1.5 for photo datasets (CelebA and Butterfly) as shown in Appendix B.

Observation 3: High CFG scales cause high-frequency degradation and low CFG scales cause low-frequency degradation. CFG scale does not only affect the speed of model collapse, but also the pattern of model collapse. As shown in Figure 1b, CFG 1.0 makes the images progressively more blurry in the Chain of Diffusion, eventually collapsing to images without any structure, which we refer to low-frequency degradation. On the other hand, for CFG 7.5 how images degrade looks completely different: some features start to be emphasized excessively, repetitive patterns begin to appear, and the overall color distribution becomes saturated. The t-SNE plot in Figure 3a clearly demonstrates that the distribution shift over iterations follows distinct paths for high, low, and medium CFG scales. These patterns were consistent in all four datasets (Appendix B). In Section 4, we detail how different patterns of power spectra in high-frequency regions from different CFG scales can be understood using the framework from genetic biology.

Implications of our observations. Our extensive investigations show that a high CFG of 7.5, a common choice to generate visually appealing images, significantly increases collapse rate to achieve slightly better FID_1. Sampling for maximizing the perceptual quality was coined as 'sampling bias' in Alemohammad et al. (2023). While they reported a monotonic increase in collapse rate as CFG increased from 1.0 to 2.0, we show that the holistic picture is not entirely monotonic when we look at a wider range of CFG scales from 1.0 to 10.0. It shows an intriguing trade-off between perceptual quality and reusability. This suggests that developers concerning reusability of images can substantially improve future generations by carefully choosing CFG.

4 Quantitative Trait Models for Fully-Synthetic Training Loops

This section introduces a novel perspective to understand model collapse in generative models by drawing parallels between genetic biology and the Chain of Diffusion. Distinguishable iterations in the Chain of Diffusion—where each iteration is separable with no duplicated individual, and the current iteration originates from the previous one—mirror genetic processes involving successive iterations of parents and offspring. By applying quantitative trait modeling from statistical genetics, we provide a framework to describe how images evolve across iterations in the Chain of Diffusion.

We begin by introducing quantitative trait modeling and its underlying mathematical assumptions. Based on them, we derive a theorem showing that the mean trait value exhibits linear drift, while the

variance stabilizes over time. This model can successfully capture the three key behaviors observed in Section 3: high-frequency degradation, low-frequency degradation, and optimal CFG scale. This suggests that different CFG scales in the Chain of Diffusion can be viewed as varying selection strategies, with power spectra corresponding to quantitative traits. We use the term "iteration" instead of "generation" for genetic generation to avoid confusion with image generation.

4.1 Quantitative Trait Modeling

Quantitative trait (QT) modeling in statistical genetics explores the evolution of quantitative phenotypes (e.g., height, weight, or color), which are complexly decided by multiple genetic and environmental factors. They are typically assumed to follow a Gaussian distribution, with discrete iterations where parents and offspring are distinguishable (t- and $t+1$-th iterations are clearly separable).

Let the distribution of phenotypes at the t-th iteration be denoted as $\mathcal{N}(\mu_t, \sigma^2_{P,t})$ where μ_t and $\sigma^2_{P,t}$ are the mean and variance of quantitative phenotypes. The phenotypic variance $\sigma^2_{P,t}$ is the sum of the (additive) genetic variance $\sigma^2_{G,t}$ and the environmental variance σ^2_E, i.e., $\sigma^2_{P,t} = \sigma^2_{G,t} + \sigma^2_E$ (Falconer, 1996)[4]. When selection occurs in each iteration, whether natural (e.g., faster animals surviving predators) or artificial (e.g., breeding livestock for higher milk production), it affects the distribution of the effective population that influences the next iteration. We consider directional selection with truncation as shown in Figure 3b, where r ratio of the samples is selected by truncating r_1 from the left and r_2 from the right side of the distribution. Here, $r + r_1 + r_2 = 1$ and larger phenotype values are preferred when $r_1 > r_2$.

The (narrow-sense) heritability[5], which is defined as the proportion of phenotypic variance attributable to additive genetic factors, can also be represented using the Breeder's Eqn. (Lush, 2013) as:

$$h^2_t = \frac{\sigma^2_{G,t}}{\sigma^2_{P,t}} = \frac{\sigma^2_{G,t}}{\sigma^2_{G,t} + \sigma^2_E} = \frac{\mu_{t+1} - \mu_t}{\mu'_t - \mu_t}, \tag{5}$$

where the mean phenotype of the next iteration μ_{t+1} is represented using the mean phenotype of selected individuals μ'_t, the mean phenotype of the entire population μ_t, and heritability h^2_t. We assume the genetic variance for the next iteration is determined by the variance of selected individuals $\sigma^2_{G,t+1} = \sigma'^2_{P,t}$. We prove the behaviors of mean and variance of phenotypes under this setting: Suppose the distributions of phenotypes follow Gaussian distribution and directional selection truncates individuals on both sides with ratios r_1 and r_2. Then mean of phenotypes asymptotically increases (decreases) by $\frac{c_1 c_2}{\sqrt{1-c_2}}\sigma_E$ per iteration and the variance converges to $\frac{1}{1-c_2}\sigma^2_E$ when $r_1 > r_2$ ($r_2 < r_1$), where c_1 and c_2 are constants that depend on r_1 and r_2. The proof of Theorem 4.1 is provided in Appendix D.1.

4.2 Explaining the Chain of Diffusion with Quantitative Trait Modeling

Just as phenotypes evolve through hidden genotypes in QT modeling, image features similarly evolve in the Chain of Diffusion. We identify high-frequency power spectra as a crucial phenotype, influenced by CFG scale as a selection mechanism. Using 2D Fourier transforms, we analyze high-frequency components above a certain threshold. Different CFG scales correspond to different selection strategies: a high CFG selects individuals in the right tail of the distribution, favoring more detailed features with reduced diversity, while a low CFG selects from the left tail.

Figure 3c compares the evolution of power spectra in the Chain of Diffusion (solid line) with the phenotype mean modeled by Eqn. 10 (dotted lines). The simulation parameters, including initial mean μ_0 and genetic deviation $\sigma_{G,0}$, are set to match the original image set at iteration 0. Three ratio configurations, r_1 and r_2, effectively model the power spectra distribution: CFG 7.5 is modeled as selecting the top 5%, CFG 2.5 selects 50% of the samples slightly favoring higher frequency, and CFG 1.0 selected the lower 30%. This modeling accurately captures power spectra evolution over six iterations of the Chain of Diffusion. More details can be found in Appendix D.2.2.

[4]Genetic variance is composed of additive, dominance, and interaction variance. Here, we only consider additive variance, which is a common assumption in the field.

[5]The heritability is narrow-sense when the genetic variance is restricted to additive variance.

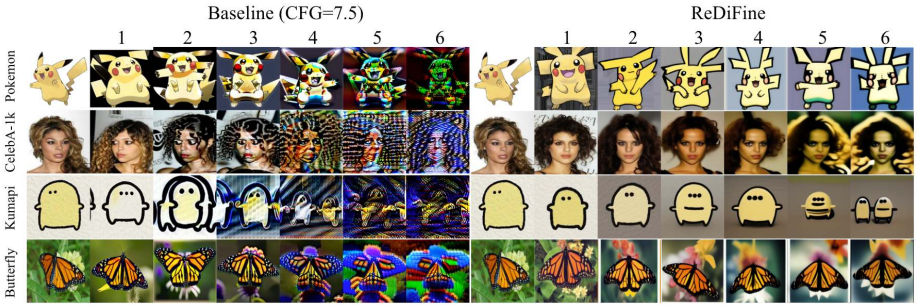


Figure 4: **ReDiFine significantly improves the baseline without extensive hyperparameter tuning in the Chain of Diffusion.** It performs robustly across multiple datasets, resolving high-frequency degradation observed in the baseline with the default configuration.

4.3 Statistical Genetics as a Lens to Model Collapse

QT modeling, a tool borrowed from statistical genetics, serves as a solid theoretical framework for interpreting our experimental results. Given the parallels between genetic processes and the Chain of Diffusion, we find that key concepts from mathematical genetics align closely with existing theoretical work on model collapse. For instance, the traditional Wright-Fisher model (Wright, 1931) describes how traits evolve from one generation to the next in a finite population. Due to randomness in the sampling process, the composition of traits within a finite population drifts over time and eventually collapses to a single phenotype. The Wright-Fisher model and its continuous variants (Tataru et al., 2017) are exactly equivalent to the collapse behavior modeled with simple categorical or Gaussian distributions in Alemohammad et al. (2023); Bertrand et al. (2023); Shumailov et al. (2023; 2024). Given its widespread use in genetics, many extensions of the Wright-Fisher model continue to be an active area of research, such as including different mechanisms of selection (He et al., 2017; Kaj et al., 2024) or mutations (Charlesworth, 2020). Beyond the current work and existing theoretical studies on model collapse, we believe that statistical genetics offers a unified perspective to understand model collapse by reducing the complex dynamics of the self-consuming loop of generative model training to a small number of parameters, from which we can gain valuable insights.

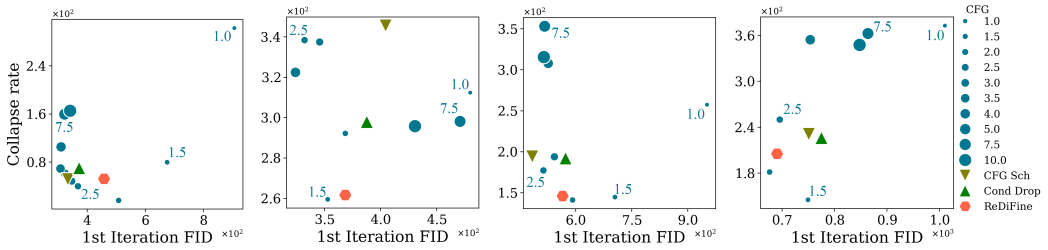
5 Reusable Image Generation with ReDiFine

In the previous sections, we have discovered a significant role of CFG in model collapse and that a good choice of CFG can mitigate the collapse while preserving the first iteration FID. However, the optimal CFG value differs for each dataset (e.g., 1.5 for CelebA, 2.5 for Pokemon). There is no efficient way of finding an optimal CFG other than iteratively finetuning the diffusion model to evaluate the collapse rate on each configuration to perform a grid search. In practical scenarios, it is unlikely that non-expert end users will go through such a process just to prevent a potential model collapse. This raises the question: *How can we design a user-friendly finetuning and generation strategy that can slow down model collapse without CFG tuning?*

To address this question, we again draw inspiration from the evolution process in nature where mutations naturally counteract genetic drift and preserve diversity. Furthermore, selection in nature is often a soft process rather than a hard truncation illustrated in Figure 3b. This soft selection allows for the inclusion of outliers, thereby maintaining the overall genetic diversity. We connect these biological inspirations with two strategies: *condition drop finetuning* to include more randomness and *CFG scheduling* during generation to transform hard selection to a softer one[6]. We propose **Re**usable **Di**ffusion **Fine**tuning (ReDiFine) which integrates these two ideas and achieves a collapse rate similar to the optimal CFG, producing reusable images with minimal extra effort.

Condition drop finetuning. We introduce condition drop finetuning, which randomly drops the text condition during finetuning to update both the conditional and unconditional scores. Although condi-

[6]We show a modified quantitative trait modeling result with these two modifications in Appendix D.2 and show that it captures the ReDiFine experimental results effectively. In this section, we focus on the image generation results with ReDiFine. We refer the curious readers to the appendix.



(a) Pokemon. (b) CelebA-1k. (c) Kumapi. (d) Butterfly.

Figure 5: **ReDiFine performs comparably to the optimal CFG scale, outperforming the baseline (CFG=7.5) on collapse-FID trade-off across all four datasets**. While the optimal CFG scale varies for different datasets, ReDiFine consistently achieves low collapse rate and FID at the same time (lower is better). Note that the differences in FID are relatively smaller than those in collapse rate, supporting the necessity to evaluate collapse rate in the Chain of Diffusion.

tion drop was suggested in the original CFG paper (Ho & Salimans, 2022), it is not a common practice during finetuning since we can achieve good images without it in the first iteration where model collapse is yet to happen (see Figure 1b). However, the small Diff (= Cond Score − Uncond Score) can accumulate over multiple iterations, leading to a significant gap as the Chain of Diffusion progresses. On the other hand, condition drop finetuning with drop probability 0.2 preserves the norm of Diff over the iterations (see Figure 36b).

CFG Scheduling. We propose gradually reducing the CFG scale during diffusion steps to mitigate the impacts of overemphasizing the conditional score in later stages, which can lead to high-frequency degradation. Specifically, we exponentially decrease the CFG scale s during T diffusion steps as $s = s_0 \times e^{-\alpha \times t/T}$, where s_0 is the initial CFG scale and α is the rate of exponential decay. This scheduling approach is consistent with findings by Balaji et al. (2022), which suggest that different diffusion steps contribute uniquely to the generation process. We note here that ReDiFine adopts exponential decrease of CFG scale for simplicity, but any advanced CFG scheduling methods (Sadat et al., 2023; Karras et al., 2024; Kynkäänniemi et al., 2024) can be explored in future research.

ReDeFine Results. We present the generated images and quantitative metrics for ReDiFine, showing its robust performance without hyperparameter tuning in all four datasets. We use the default initial CFG scale $s_0 = 7.5$, decay rate $\alpha = 2$, and condition drop probability 0.2, except in the robustness comparison and ablation study. Figure 4 shows that ReDiFine generates significantly better images compared to the baselines (CFG=7.5), and performs comparably to the optimal CFG scales. In addition to the visual comparison, Figure 5 quantitatively shows the collapse-FID trade-off of ReDiFine. In all four datasets, ReDiFine shows substantially lower collapse rate (y-axis), compared to the baseline case (CFG=7.5). Furthermore, the performance of ReDiFine is close to the optimal Pareto curve spanned by different CFG scales, achieving similar performance as the optimal CFG values, demonstrating its effectiveness as a universal and user-friendly solution. In contrast, using a fixed CFG scale that works well for one dataset often fails on others: CFG 2.5 is optimal for Pokemon but performs poorly for CelebA-1k and Butterfly, and CFG 1.5 is optimal for CelebA-1k but performs poorly for Pokemon and Kumapi. Further experimental results of ReDiFine for six additional iterations, cross-domain data, and iterative retraining can be found in Appendix E.

Ablation study & Further analyses. We conducted an ablation study to understand the contributions of condition drop finetuning and CFG scheduling to ReDiFine. We plot the results of using only condition drop and CFG scheduling in green triangles on Figure 5. In Pokemon, using only one of those strategies outperforms ReDiFine, but in all other datasets, using only one strategy shows higher collapse rate than ReDiFine. Especially in CelebA-1k, using either one of them showed significantly worse performance than ReDiFine. It suggests that combining condition drop and CFG scheduling builds robustness to the method, making ReDiFine effective across all tested datasets. We conducted further analyses on ReDiFine, examining the distribution of latent features, the evolution of the norm of Diff, the power spectra using 2D Fourier transforms, and forensic fingerprints based on prior studies (Corvi et al., 2023a;b). Our analysis shows that ReDiFine effectively preserves the latent distribution and the norms of Diff over six iterations, with forensic fingerprints closely resembling those of the optimal CFG case. Detailed results are provided in Appendix G.

6 Conclusion

The influx of AI-generated data into the world is inevitable and training sets that consist of synthetic data will be part of the AI development pipeline. In this paper, we empirically and theoretically studied the scenario of finetuning a model with its own generated data, where a gradual degradation called "model collapse" happens. We (1) identify the most impactful factor through comprehensive empirical investigations, (2) develop a novel theoretical perspective inspired by statistical genetics to explain model collapse, and (3) propose ReDiFine strategy for diffusion finetuning to slow down image quality degradation in model collapse.

We started this paper with a question: can current AI models learn from their own output and improve themselves? Our paper shows a glimpse that widely-used text-to-image models are not ready to improve from their own creation quite yet. While we presented one solution focused on generating *reusable data*, many open directions remain, such as developing algorithms that can distinguish between real and synthetic data and apply different learning techniques accordingly.

References

David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G Baraniuk. Self-consuming generative models go mad. *arXiv preprint arXiv:2307.01850*, 2023.

Sina Alemohammad, Ahmed Imtiaz Humayun, Shruti Agarwal, John Collomosse, and Richard Baraniuk. Self-improving diffusion models with synthetic data. *arXiv preprint arXiv:2408.16333*, 2024.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, April 2024. doi: 10.1145/3620665.3640366. URL https://pytorch.org/assets/pytorch2-2.pdf.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Quentin Bertrand, Avishek Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. On the stability of iterative retraining of generative models on their own data. *arXiv preprint arXiv:2310.00429*, 2023.

Matyas Bohacek and Hany Farid. Nepotistically trained generative-ai models collapse. *arXiv preprint arXiv:2311.12202*, 2023.

Martin Briesch, Dominik Sobania, and Franz Rothlauf. Large language models suffer from their own output: An analysis of the self-consuming training loop. *arXiv preprint arXiv:2311.16822*, 2023.

Brian Charlesworth. How long does it take to fix a favorable mutation, and why should we care? *The American Naturalist*, 195(5):753–771, 2020.

Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 973–982, 2023a.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023b.

Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. Model Collapse Demystified: The Case of Regression, April 2024a. URL http://arxiv.org/abs/2402.07712. arXiv:2402.07712 [cs, stat].

Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. Strong model collapse. *arXiv preprint arXiv:2410.04840*, 2024b.

Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. *arXiv preprint arXiv:2402.07043*, 2024c.

DS Falconer. *Introduction to quantitative genetics*. Pearson Education India, 1996.

Yunzhen Feng, Elvis Dohmatob, Pu Yang, Francois Charton, Julia Kempe, and FAIR Meta. Beyond model collapse: Scaling up with syn-thesized data requires verification. *arXiv preprint arXiv:2406.07515*, 2024.

Damien Ferbach, Quentin Bertrand, Avishek Joey Bose, and Gauthier Gidel. Self-consuming generative models with curated data provably optimize human preferences. *arXiv preprint arXiv:2407.09499*, 2024.

Ronald Aylmer Fisher. *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press, 1999.

Shi Fu, Sen Zhang, Yingjie Wang, Xinmei Tian, and Dacheng Tao. Towards theoretical understandings of self-consuming generative models. *arXiv preprint arXiv:2402.11778*, 2024.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, et al. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*, 2024.

Nate Gillman, Michael Freeman, Daksh Aggarwal, Chia-Hong Hsu, Calvin Luo, Yonglong Tian, and Chen Sun. Self-Correcting Self-Consuming Loops for Generative Model Training, April 2024. URL http://arxiv.org/abs/2402.07087. arXiv:2402.07087 [cs, stat].

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text, April 2024. URL http://arxiv.org/abs/2311.09807. arXiv:2311.09807 [cs].

Hakurei. Waifu diffusion v1.4. https://huggingface.co/hakurei/waifu-diffusion-v1-4, 2022. Accessed: 2023-04-18.

Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will Large-scale Generative Models Corrupt Future Datasets? In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20498–20508, Paris, France, October 2023. IEEE. ISBN 9798350307184. doi: 10.1109/ICCV51070.2023.01879. URL https://ieeexplore.ieee.org/document/10376575/.

Zhangyi He, Mark Beaumont, and Feng Yu. Effects of the ordering of natural selection and population regulation mechanisms on wright-fisher models. *G3: Genes, Genomes, Genetics*, 7(7):2095–2106, 2017.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022.

Ihelon. Illustrations kumapi390, 2022. URL `https://www.kaggle.com/datasets/ihelon/illustrations-kumapi390`. Accessed: 2023-04-19.

Ingemar Kaj, Carina F Mugal, and Rebekka Müller-Widmann. A wright–fisher graph model and the impact of directional selection on genetic variation. *Theoretical Population Biology*, 159:13–24, 2024.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *arXiv preprint arXiv:2406.02507*, 2024.

Motoo Kimura. Solution of a process of random genetic drift with a continuous model. *Proceedings of the National Academy of Sciences*, 41(3):144–150, 1955.

kohya-ss. kohya-ss trainer. URL `https://github.com/kohya-ss/sd-scripts`.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *arXiv preprint arXiv:2404.07724*, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. URL `https://arxiv.org/abs/2201.12086`.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Jay L Lush. *Animal breeding plans*. Read Books Ltd, 2013.

Matteo Marchi, Stefano Soatto, Pratik Chaudhari, and Paulo Tabuada. Heat death of generative models in closed-loop learning. *arXiv preprint arXiv:2404.02325*, 2024.

Gonzalo Martínez, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juarez, and Rik Sarkar. Combining generative artificial intelligence (ai) and the internet: Heading towards evolution or degradation? *arXiv preprint arXiv:2303.01255*, 2023a.

Gonzalo Martínez, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juarez, and Rik Sarkar. Towards understanding the interplay of generative artificial intelligence and the internet. In *International Workshop on Epistemic Uncertainty in Artificial Intelligence*, pp. 59–73. Springer, 2023b.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Cyriel Paris, Bertrand Servin, and Simon Boitard. Inference of selection from genetic time series using various parametric approximations to the wright-fisher model. *G3: Genes, Genomes, Genetics*, 9 (12):4073–4086, 2019.

Pokémon. Pokédex. `https://www.pokemon.com/us/pokedex`, 2023. Accessed: 2023-04-19.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

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*, pp. 10684–10695, 2022.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. *arXiv preprint arXiv:2310.17347*, 2023.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

StabilityAI. Sd vae ft mse original, 2022. URL `https://huggingface.co/stabilityai/sd-vae-ft-mse-original`. Accessed: 2023-04-19.

George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Rohan Taori and Tatsunori Hashimoto. Data feedback loops: Model-driven amplification of dataset biases. In *International Conference on Machine Learning*, pp. 33883–33920. PMLR, 2023.

Paula Tataru, Maria Simonsen, Thomas Bataillon, and Asger Hobolth. Statistical inference in the wright–fisher model using allele frequency data. *Systematic biology*, 66(1):e30–e46, 2017.

Veeralakrishna. Butterfly dataset. `https://www.kaggle.com/datasets/veeralakrishna/butterfly-dataset`, 2020. Accessed: 2024-01-03.

Sewall Wright. Evolution in mendelian populations. *Genetics*, 16(2):97, 1931.

Huminhao Zhu, Fangyikang Wang, Tianyu Ding, Qing Qu, and Zhihui Zhu. Analyzing and improving model collapse in rectified flow models. *arXiv preprint arXiv:2412.08175*, 2024a.

Xuekai Zhu, Daixuan Cheng, Hengli Li, Kaiyan Zhang, Ermo Hua, Xingtai Lv, Ning Ding, Zhouhan Lin, Zilong Zheng, and Bowen Zhou. How to synthesize text data without model collapse? *arXiv preprint arXiv:2412.14689*, 2024b.

A Experimental setup

A.1 Hyperparameters

We finetune Stable Diffusion v1.5 (Rombach et al., 2022) using LoRA (Hu et al., 2021) at each iteration, with ft-MSE (StabilityAI, 2022) as a fixed VAE to project images into latent space. Horizontal flip is the only image augmentation applied. Our implementation follows kohya-ss and is built on PyTorch v2.2.2 (Ansel et al., 2024), with torchvision 0.17.2, running on CUDA 12.4 using NVIDIA A-100 and L40S GPUs. All default hyperparameters are listed in Table 1.

Table 1: Default hyperparameters used for the Chain of Diffusion.

Hyperparameter	**Value**
Optimizer	AdamW
Learning Rate - Unet	0.0001
Learning Rate - CLIP	0.00005
LoRA Weight Scaling	8
LoRA Rank	32
Batch Size	6
Max Epochs	100
CLIP Skip	2
Noise Offset	0.0
Mixed Precision	fp16
Loss Function	MSE
Min SNR gamma	5.0
Max Gradient Norm Clipping	1.0
Caption Dropout Rate	0.0
Sampler	Euler A
Classifier-Free Guidance Scale	7.5
Number of Diffusion Steps	30
Number of Images per Prompt	1

A.2 Datasets

We use four image datasets to demonstrate the universal nature of degradation: Pokemon (Pokémon, 2023), Kumapi (Ihelon, 2022), Butterfly (Veeralakrishna, 2020), and CelebA-1k (Liu et al., 2015), covering animation, handwriting, and real images. All images are resized to 512×512 pixels. Text prompts are generated using BLIP captioner (Li et al., 2022) and Waifu Diffusion v1.4 tagger (Hakurei, 2022). Sample images and prompts can listed in Table 2.

Pokemon. The Pokemon dataset (Pokémon, 2023) contains 1008 images indexed by number, with prompts combining BLIP captions (length 50-75 words) and Waifu Diffusion tagger.

CelebA-1k. CelebA-1k is a subsample of 1000 images from CelebA (Liu et al., 2015), with BLIP captions (25-50 words).

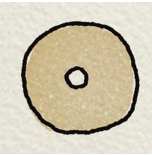
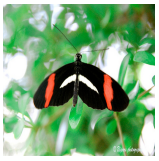
Kumapi. Kumapi (Ihelon, 2022) includes 391 handwriting-style images, with Waifu-generated prompts and manual adjustments.

Butterfly. Butterfly (Veeralakrishna, 2020) includes 832 real picture images across 8 species, using BLIP captions (length 50-75 words) and species descriptions.

A.3 Evaluation metrics

FID The Fréchet Inception Distance (FID) (Heusel et al., 2017) is a widely used metric for evaluating the quality of generative models. It measures the similarity between the distributions of generated images and real images by comparing their feature representations extracted from a

Table 2: Sample images and prompts for Pokemon, CelebA-1k, Kumapi, and Butterfly datasets.

Pokemon	
a green pokemon with red eyes and a leaf on the back of its head and tail, an image of the pokemon character with a red eye and big green tail, all set up to look like it is holding a lea, ultra-detailed, high-definition, high quality, masterpiece, sugimori ken (style), solo, smile, open mouth, simple background, red eyes, white background, standing, full body, pokemon (creature), no humans, fangs, transparent background, claws, Bulbasaur	
a very cute looking pokemon with a big leaf on its back and a big leaf on its head, a very cute little pokemon character with leaves in the back ground around his chest and hea, white background, ultra-detailed, high-definition, high quality, masterpiece, sugimori ken (style), solo, red eyes, closed mouth, standing, full body, pokemon (creature), no humans, fangs, transparent background, claws, outline, white outline, animal focus, fangs out	
CelebA-1k	
a woman with brown hair smiling and posing for a picture in front of a mirror and gold and white stripes	
a woman with a very long red hair smiles and laughs on a city street and other people in the background	
Kumapi	
solo, simple background, food, donut, grey background, no humans, food focus, still life, Kumapi style	
solo, looking at viewer, cute yellow figure, two tiny hands and feet, simple background, black dot eyes, white background, grey background, no humans, Kumapi style	
Butterfly	
a crimson patched longwing butterfly with a red and black stripe on its wings, wings are long, narrow, rounded, black, crossed on fore wing by broad crimson patch, and on hind wing by narrow yellow line	
a Common Buckeye butterfly is sitting on a flower in the sun, wings scalloped and rounded except at drawn-out fore wing tip, on hind wing, 1 large eyespot near upper margin and 1 small eyespot below it. Eyespots are black, yellow-rimmed, with iridescent blue and lilac irises, on fore wing, 1 very small near tip and 1 large eyespot in white fore wing bar.	

pretrained Inception network. Following Stein et al. (2024), we use DiNOv2 feature extractor (Oquab et al., 2023) instead of Inception, as it aligns better with human evaluation for high-quality images.

CLIP score The CLIP score is a metric used to evaluate the alignment between an image and a text description, leveraging the pretrained CLIP (Contrastive Language-Image Pre-training) model (Radford et al., 2021). CLIP is designed to encode both images and text into a shared embedding space, where semantically related image-text pairs have high cosine similarity. Based on this embedding space, the CLIP score measures the text-alignment capability of text-to-image generative models by computing cosine similarities of generated images and text descriptions used for generations.

SFD We suggest and compute the average sample-wise feature distance (SFD) between a pair of images corresponding to the same text prompt as the fidelity metric applicable for text-to-image generation. It computes text-prompt-wise similarities of the real image and generated image, more directly comparing two pairs of images while other metrics compares the overall distribution of datasets. We calculate

$$\mathrm{SFD}_k = \frac{1}{N} \sum d(f(x_{k,i}), f(x_{0,i})) \tag{6}$$

to evaluate each iteration k. SFD overcomes the problem of FID being sensitive to the number of images to compare.

Recall As a key metric for evaluating the diversity of generative models, recall measures the fraction of the real data distribution that is captured by the generated data distribution. Following the method proposed by Kynkäänniemi et al. (2019), we utilize the DiNOv2 feature extractor and set the number of neighbors to 5 for our evaluations.

B Different CFG Scales

This section presents images from Chain of Diffusion at various CFG scales for the four datasets. Figure 6, 7, 8, and 9 corresponds to Pokemon, CelebA-1k, Kumapi, and Butterfly, respectively. Each dataset shows results for five CFG scales, covering high, medium, and low values. The optimal medium CFG scale is 2.5 for Pokemon and Kumapi, and 1.5 for CelebA-1k and Butterfly. Lower-than-optimal CFG scales lead to low-frequency degradations, while higher-than-optimal scales result in high-frequency degradations with saturated colors and repetitive patterns. Notably, the optimal CFG for Pokemon and Kumapi causes severe degradation in CelebA-1k and Butterfly.

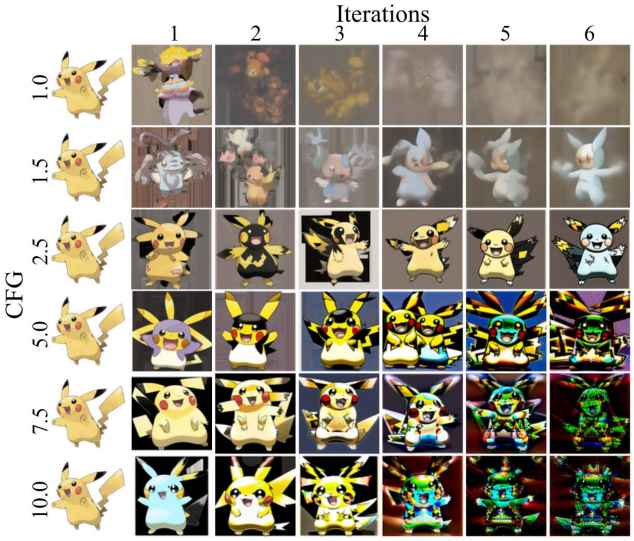


Figure 6: Chain of Diffusion for Pokemon at various CFG scales. The optimal CFG scale is 2.5.

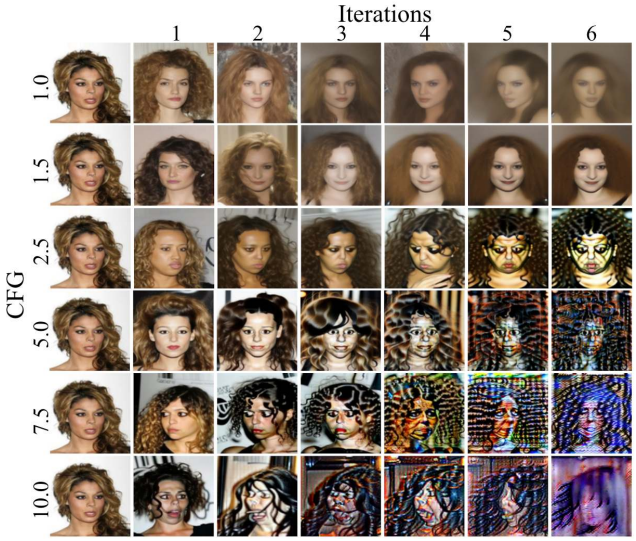


Figure 7: Chain of Diffusion for CelebA-1k at various CFG scales. The optimal CFG scale is 1.5.

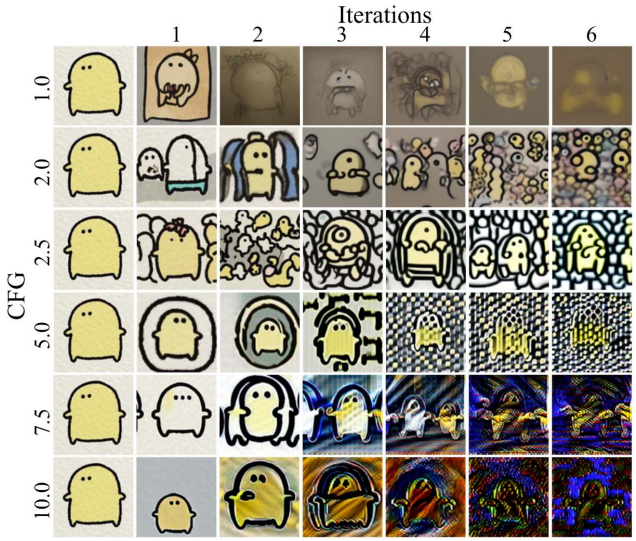


Figure 8: Chain of Diffusion for Kumapi at various CFG scales. The optimal CFG scale is 2.5.

C Hyperparameter investigations to unveil the most significant factor of degradation

This section presents experimental results identifying the most significant factors contributing to degradation in the Chain of Diffusion. We systematically vary each hyperparameter from Figure 2, using the default settings from Table 1, to assess their impact on degradation. The CFG scale is fixed at 7.5 for all cases.

C.1 Training set size

We used CelebA dataset (Liu et al., 2015) to examine how training set size (both D_0 and D_k) impacts degradation in the Chain of Diffusion. By subsampling, we adjusted the training set to 100, 250, 500, and 2000 images. Degradation occurs regardless of dataset size, as shown in Figure 10, but appears earlier with smaller sets. By the 6th iteration, images degrade severely for all cases. The number of parameter updates was kept constant across all dataset sizes.

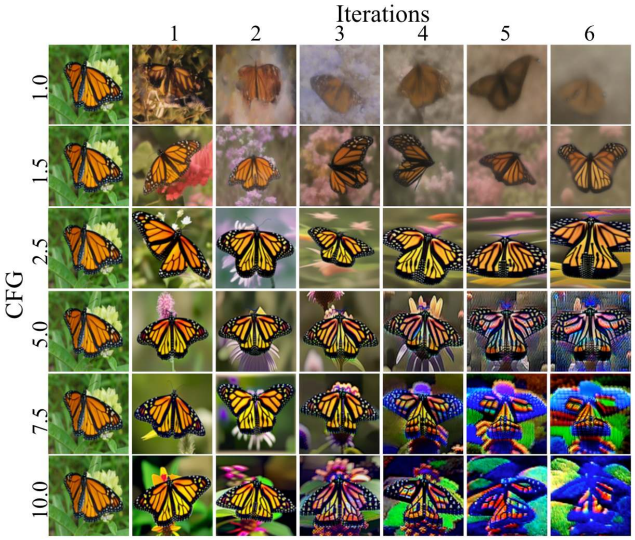


Figure 9: Chain of Diffusion for Butterfly at various CFG scales. The optimal CFG scale is 1.5.

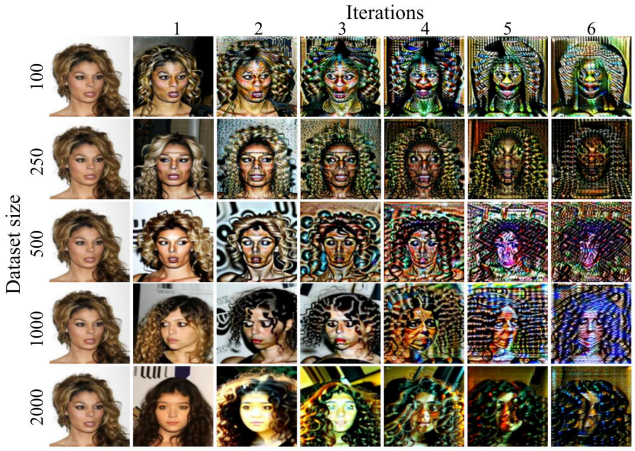


Figure 10: Chain of Diffusion on CelebA dataset with varying training set sizes. Degradation occurs faster with smaller sets, but all result in severe degradation by the 6th iteration.

C.2 Number of images per prompt

Generating multiple images per prompt is a simple way to increase training set diversity and can be considered as a solution to mitigate degradation in the Chain of Diffusion. We tested this by generating 5 times more images. As shown in Figure 11, while degradation is slightly delayed (by one iteration), it remains unmitigated, and the high computational cost makes this approach impractical.

Figure 11: Chain of Diffusion on Pokemon dataset with multiple images generated per prompt. Increasing the training set size in this way does not mitigate degradation.

C.3 MIXING REAL IMAGES TO SYNTHETIC SETS

Many previous works suggest augmenting the training set with real images to mitigate degradation during iterative training. We investigated whether mixing images from the original training set into the synthetic set at each iteration could alleviate this issue. At each iteration, images are randomly replaced with corresponding real images. Figure 12 and 13 show how degradation in the Chain of Diffusion varies when 50%, 90%, 95%, 98% and 99% of images are replaced for Pokemon and CelebA-1k datasets, respectively. Notably, even 5% synthetic images are sufficient to induce degradation, and 50% replacement rarely slows it down. CelebA-1k dataset appears to be significantly more susceptible to degradation. Importantly, using only 5% and 2% synthetic images is sufficient to trigger model collapse in Pokemon and CelebA-1k datasets. This demonstrates the catastrophic impact of synthetic data on diffusion finetuning and raises the question of whether synthetic data detection can be a safe solution.

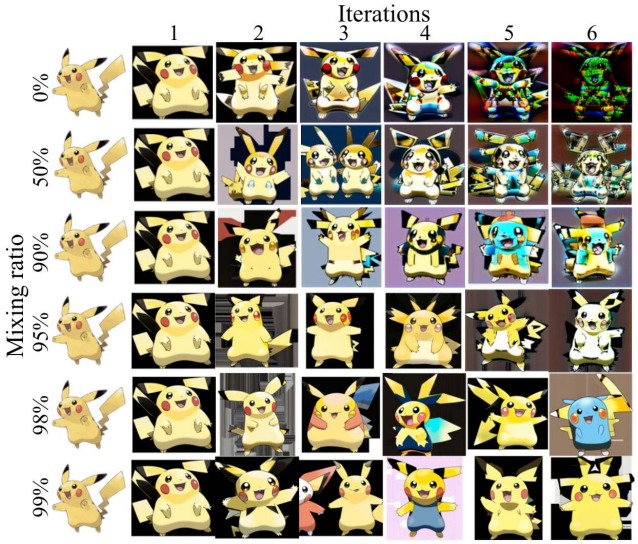


Figure 12: Chain of Diffusion on Pokemon dataset with real images randomly replacing synthetic images at each iteration. A 50% replacement rarely slows degradation, while 10% synthetic images are sufficient to initiate it.

C.4 PROMPT SET

We hypothesized that the descriptiveness of prompts influences degradation in the Chain of Diffusion. We tested various prompt sets for Pokemon and CelebA-1k datasets. For Pokemon dataset, the default prompt set consists of concatenated Waifu and BLIP captions, with BLIP captions ranging from 50 to 75 words. Figure 14 shows how using only Waifu prompts and varying BLIP caption lengths affect degradation. Notably, different styles of high-frequency degradation were observed; shorter prompts reduced repetitive patterns but decreased diversity. In CelebA-1k dataset, varying BLIP caption lengths resulted in similar degradation levels, but prompts that were either insufficiently or excessively descriptive caused the images to deviate from the originals, as shown in Figure 15.

C.5 U-NET AND TEXT-ENCODER

Figure 16 illustrates the Chain of Diffusion with either the U-Net or text encoder finetuned. When the text encoder is not updated (second row), similar degradation occurs. However, the degradation pattern changes when the U-Net is not updated, as the model's ability to generate images remains unchanged. In contrast, updating the text encoder results in a loss of image content preservation.

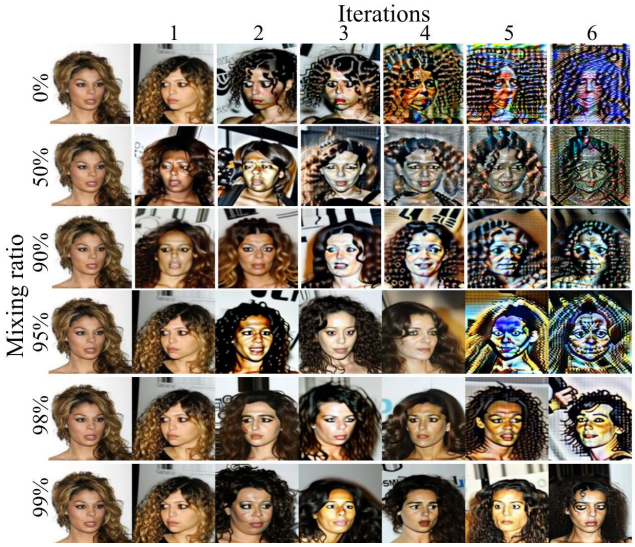


Figure 13: Chain of Diffusion on CelebA-1k dataset with real images randomly replacing synthetic images at each iteration. A 50% replacement rarely slows degradation, while 5% synthetic images are sufficient to initiate it. It suffers from more severe degradation than Pokemon dataset as compared with Figure 12.

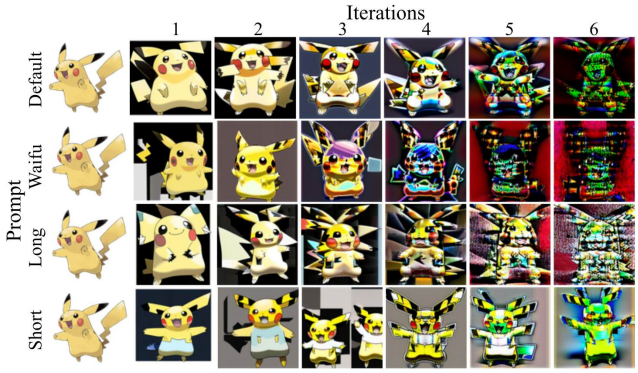


Figure 14: Chain of Diffusion on Pokemon dataset with different prompts. The default prompts are concatenations of BLIP captions (50-75 words) and Waifu captions. We compare the Chain of Diffusion using default captions, Waifu captions, short (less than 25 words) and long (50-75 words) BLIP captions.

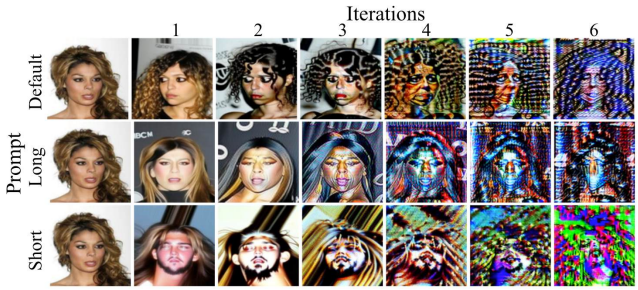


Figure 15: Chain of Diffusion on CelebA-1k dataset with different prompts. The default prompts range from 25 to 50 words. We compare the Chain of Diffusion using longer prompts (over 50 words) and shorter prompts (under 25 words).

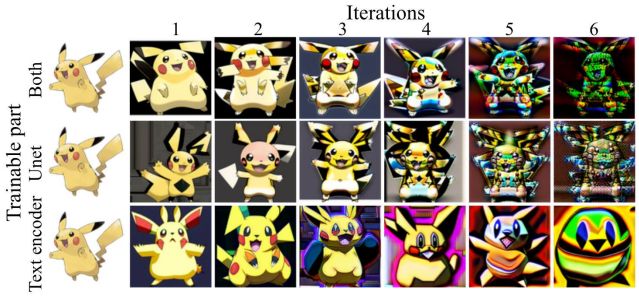


Figure 16: Chain of Diffusion on Pokemon dataset with either the U-Net or text encoder finetuned.

C.6 Number of diffusion steps

We investigated whether an insufficient number of diffusion steps during generation contributes to degradation. Figure 17 shows that increasing the number of diffusion steps does not enhance the Chain of Diffusion.

C.7 Number of Epochs

We also examined whether insufficient or excessive training affects our default setting. As shown in Figure 17, images from the initial iterations exhibit similar quality, resulting in comparable degradations. We set the default finetuning to 100 epochs since loss values continue to decrease after 50 epochs.

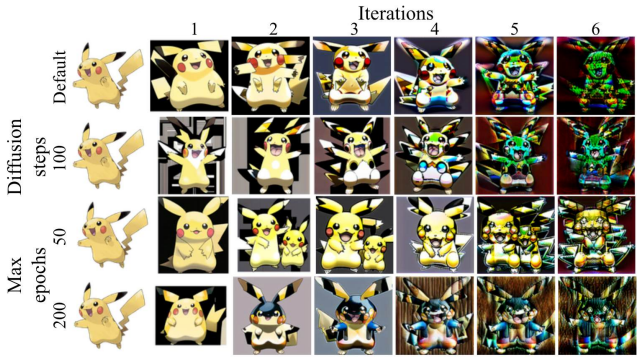


Figure 17: Chain of Diffusion on Pokemon dataset with varying diffusion steps and training epochs. Both increased diffusion steps and differing training epochs fail to mitigate degradation, resulting in similar patterns.

C.8 Learning Rate

Similarly, we assessed finetuning adequacy in Figure 18 by adjusting the learning rates for the U-Net and text encoder by x10 and x0.1. Images from the initial iterations show that the default values are suitable for finetuning. Although the styles are different, degradation consistently occurs across different learning rates.

C.9 CLIP Skip

We investigated the impact of the CLIP skip hyperparameter on degradation. The CLIP skip determines which intermediate feature from the CLIP text encoder is used as the text embedding for conditional generation, with smaller values selecting features closer to the output and larger values selecting those nearer to the input text. As shown in Figure 19, this hyperparameter has minimal effect on degradation patterns.

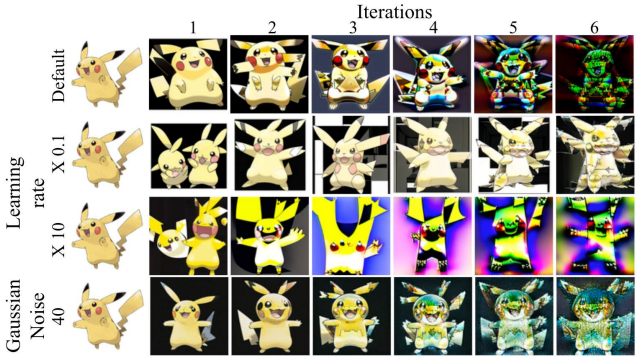


Figure 18: Chain of Diffusion on Pokemon dataset with varying learning rates and added Gaussian noise to the original training set.

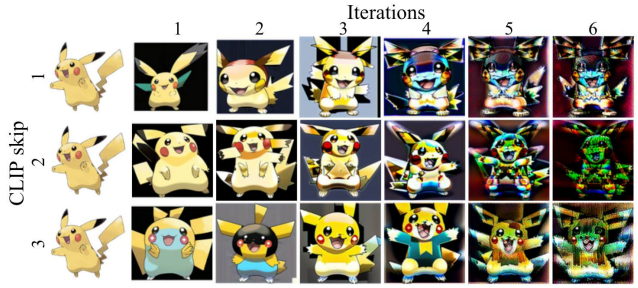


Figure 19: Chain of Diffusion on Pokemon dataset using different CLIP skip hyperparameters. The CLIP skip hyperparameter shows a negligible effect on degradation.

C.10 Adding Gaussian noise to the original training set

We examined whether differences between real and synthetic images contribute to degradation by adding random Gaussian noise to the original training set D_0. Figure 18 illustrates that while the characteristics of the original training set have some effect, degradation still occurs. This supports our findings that degradations are universal across real, animation, and handwritten images.

C.11 Stable Diffusion XL

We investigated image degradation for a different Stable Diffusion model, noting that the optimal hyperparameters for finetuning SDXL using LoRA are not well established. Consequently, we applied the same hyperparameters used for Stable Diffusion v1.5, which may be suboptimal. To manage space complexity, we reduced the batch size to 2 and maintained a resolution of 512×512, as the first iteration images exhibit impressive quality. Results are presented in Figure 20.

C.12 Data Accumulation

Data accumulation experiments aim to investigate whether concept overfitting and disappearing are major reasons for model collapse. The training set for iteration t is the combination of all previously generated sets, including the original training set. The number of training epochs is controlled accordingly to maintain the total number of updates.

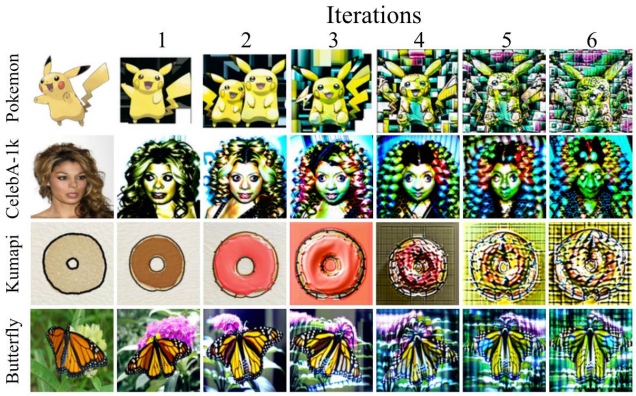


Figure 20: Chain of Diffusion of SDXL on four datasets. Due to insufficient investigation into optimal hyperparameters for finetuning SDXL, our experiments largely rely on those from Stable Diffusion v1.5.

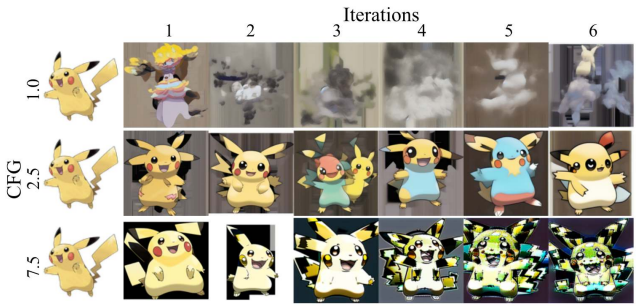


Figure 21: Chain of Diffusion on Pokemon dataset when training set is accumulated from previous iterations. All concepts from previous iterations are preserved for finetuning.

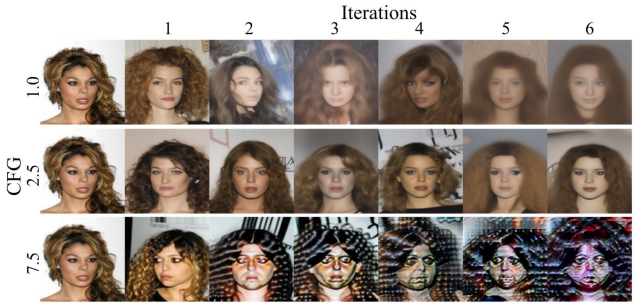


Figure 22: Chain of Diffusion on CelebA-1k dataset when training set is accumulated from previous iterations. All concepts from previous iterations are preserved for finetuning.

D Quantitative Trait Modeling

D.1 Proof of Theorem 4.1

Proof. The difference between the successive means in quantitative trait modeling is given by:

$$\Delta\mu_t = \mu_{t+1} - \mu_t = h_t^2(\mu_t' - \mu_t). \tag{7}$$

When Gaussian distribution with mean μ and variance σ^2 is truncated on both sides with r_1 and r_2 ratios, the mean and variance of truncated Gaussian distribution are expressed as:

$$\mu_{trun} = \mu - \frac{\varphi(\beta) - \varphi(\alpha)}{\Phi(\beta) - \Phi(\alpha)}\sigma, \tag{8}$$

$$\sigma_{trun}^2 = (1 - \frac{\beta\varphi(\beta) - \alpha\varphi(\alpha)}{\Phi(\beta) - \Phi(\alpha)} - (\frac{\varphi(\beta) - \varphi(\alpha)}{\Phi(\beta) - \Phi(\alpha)})^2)\sigma^2, \tag{9}$$

where φ and Φ are the probability density function (PDF) and cumulative distribution function (CDF) of the standard normal distribution, and $\alpha = \Phi^{-1}(r_1)$ and $\beta = \Phi^{-1}(1 - r_2)$. Accordingly, given $\mu = \mu_t$ and $\sigma^2 = \sigma_{P,t}^2$ with $\mu_t' = \mu_{trun}$ and $\sigma_{G,t+1}^2 = \sigma_{P,t}'^2 = \sigma_{trun}^2$, we have

$$\Delta\mu_t = h_t^2(\mu_t' - \mu_t) = \frac{\sigma_{G,t}^2}{\sigma_{P,t}^2} c_1 \sigma_{P,t} = c_1 \frac{\sigma_{P,t-1}'^2}{\sigma_{P,t}^2} \sigma_{P,t} = c_1 c_2 \frac{\sigma_{P,t-1}^2}{\sigma_{P,t}^2} \sigma_{P,t}, \tag{10}$$

where $c_1 = |\frac{\varphi(\beta)-\varphi(\alpha)}{\Phi(\beta)-\Phi(\alpha)}|$ and $c_2 = 1 - \frac{\beta\varphi(\beta)-\alpha\varphi(\alpha)}{\Phi(\beta)-\Phi(\alpha)} - (\frac{\varphi(\beta)-\varphi(\alpha)}{\Phi(\beta)-\Phi(\alpha)})^2$ when $r_1 > r_2$. On the other hand, the mean phenotypes decreases when $r_1 < r_2$ as:

$$\Delta\mu_t = -c_1 c_2 \frac{\sigma_{P,t-1}^2}{\sigma_{P,t}^2} \sigma_{P,t}. \tag{11}$$

Furthermore, the phenotype variance converges over time as:

$$\sigma_{P,t}^2 = \sigma_{G,t}^2 + \sigma_E^2 = \sigma_{P,t-1}'^2 + \sigma_E^2 = c_2 \sigma_{P,t-1}^2 + \sigma_E^2. \tag{12}$$

$$\sigma_{P,t}^2 - \frac{\sigma_E^2}{1 - c_2} = c_2(\sigma_{P,t-1}^2 - \frac{\sigma_E^2}{1 - c_2}) = \cdots = c_2^t(\sigma_{P,0}^2 - \frac{\sigma_E^2}{1 - c_2}). \tag{13}$$

As a result, the phenotype variance converges to $\frac{\sigma_E^2}{1-c_2}$ for $0 < c_2 < 1$ (the variance of truncated distribution is smaller than the variance of the original distribution), and the mean asymptotically increases (decreases) by $\frac{c_1 c_2}{\sqrt{1-c_2}}\sigma_E$ per iteration when $r_1 > r_2$ ($r_2 > r_1$). □

D.2 Simulation setup

Our simulation aims to demonstrate that our experimental results can be modeled using quantitative trait modeling. We show that the radial sum of power spectra of images is one of the phenotypes explained by our theoretical analysis.

D.2.1 Computing the radial sum of power spectra.

The power spectra of images are computed as the square of the magnitude of the 2D Fourier transform. Here, we demonstrate how to compute the radial sum of power spectra in order to use the sum of radial power spectra above a certain threshold. Given a set of images $\{I_i\}$ where $i \in \{1, 2, \cdots, N\}$, the 2D Discrete Fourier Transform (DFT) of an image I_i of size $M \times N$ is computed as:

$$F_i(u, v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I_i(x, y) e^{-2\pi j(\frac{ux}{M} + \frac{vy}{N})} \tag{14}$$

where $F_i(u, v)$ represents the frequency component at coordinates (u, v). The power spectrum of an image I_i at (u, v) is the square of the magnitude of its Fourier transform $P_i(u, v) = |F_i(u, v)|^2$. We compute the radial sum of the power spectra using a norm of each frequency component (u, v) as $r(u, v) = \sqrt{(\frac{u}{M})^2 + (\frac{v}{N})^2}$. Given the threshold frequency τ, we compute the radial sum of the high-frequency power spectra of an image I_i as

$$S_i = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} P_i(u, v) \cdot \mathrm{I}(r(u, v) > \tau) \tag{15}$$

where $\mathrm{I}(\cdot)$ is an indicator function. We compute the sum of high-frequency components because low-frequency components tend to be noisy and use the threshold frequency of 0.02. Then, the total sum of power spectra for all images can be written as:

$$S = \sum_{i=0}^{N} S_i = \sum_{i=0}^{N} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} P_i(u, v) \cdot \mathrm{I}(r(u, v) > \tau). \tag{16}$$

Practically, we shift the Fourier transform maps so that the frequency norm larger than $\frac{1}{\sqrt{2}}$ is considered to be in the opposite direction. Code for detailed implementation comes from the official code of Corvi et al. (2023a).

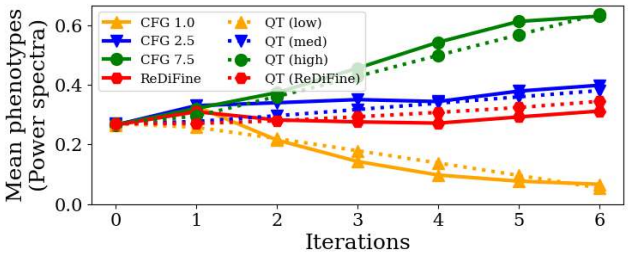


Figure 23: Results comparing the simulation for mutations and power spectra of images generated by ReDiFine (with CFG scale 7.5). Power spectra and simulation are plotted in solid and dotted lines, respectively. Our modifications to heritability and selection process successfully demonstrate changes occur to images by ReDiFine.

D.2.2 Simulation parameters matching different CFG scales.

We simulate different selection strategies using truncation ratios r_1 and r_2. To model high, medium, and low CFG scales from our experiments, we apply $(0.025, 0.675)$, $(0.5, 0.09)$, and $(0.95, 0.0002)$ for (r_1, r_2), respectively, which correspond to CFG scales of 7.5, 2.5, and 1.0. For initial values in the simulation, we compute the mean (0.027) and standard deviation (0.056) from the original training set (iteration 0), using them as the initial values for mean and genetic standard deviation for our simulation. We set the environmental standard deviation to 0.25.

D.2.3 ReDiFine and mutations.

ReDiFine combines condition drop finetuning and CFG scheduling, inspired by the mutation mechanism that compensates for distribution shift and maintains genetic distributions in population genetics. We apply two modifications to our theoretical analysis of Section 4—adding mutation variance to heritability and smoothing truncations—to simulate the effects of ReDiFine in the Chain of Diffusion. Specifically, we add the mutation standard deviation σ_M of 0.1 to heritability as:

$$h_t^2 = \frac{\sigma_{G,t}^2}{\sigma_{P,t}^2 + \sigma_M^2} = \frac{\sigma_{G,t}^2}{\sigma_{G,t}^2 + \sigma_E^2 + \sigma_M^2}, \tag{17}$$

representing the randomness added to each iteration due to mutations. This influences the effects of previous iteration to current iteration, which reflect finetuning. Moreover, we apply exponential tails to truncations instead of cut-off thresholds where samples outside the truncation area can be randomly selected with exponential distribution ($e^{-\alpha d(x)}$ where $d(x)$ is a distance to truncation zone and we use $\alpha = 0.1$). This modified selection simulates the effect of CFG scheduling during image generation.

Figure 23 shows the effects of two modifications to our simulation results, and they closely align with the power spectra of images generated by ReDiFine. This demonstrates that the effects of ReDiFine can be understood as interference similar to mutations in population genetics. This suggests further research on model collapse motivated from other fields like biology.

E ReDiFine

E.1 Visual Inspections

Figure 24, 25, 26, and 27 show how robust ReDiFine is to different CFG scales. It successfully mitigates the high-frequency degradations for a wide range of CFG scales.

E.2 Further iterations

We conduct additional experiments to compare the baseline with the optimal CFG scale and ReDiFine over extended iterations. As shown in Figure 28, ReDiFine consistently generates images of similar quality up to 12 iterations, whereas the optimally tuned CFG scale fails to sustain image quality.

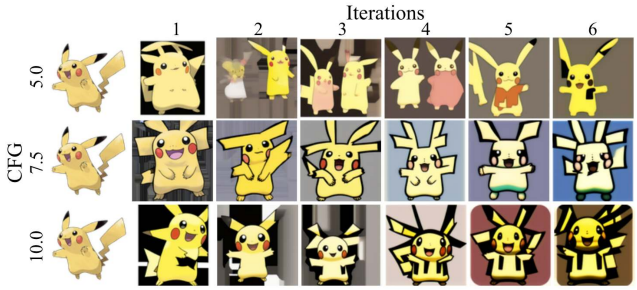


Figure 24: Chain of Diffusion of ReDiFine with different CFG scales on Pokemon dataset. ReDiFine successfully achieves robust image qualities for varying CFG scales.

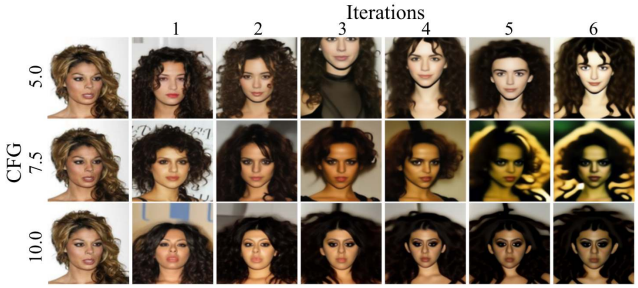


Figure 25: Chain of Diffusion of ReDiFine with different CFG scales on CelebA-1k dataset. ReDiFine successfully achieves robust image qualities for varying CFG scales.

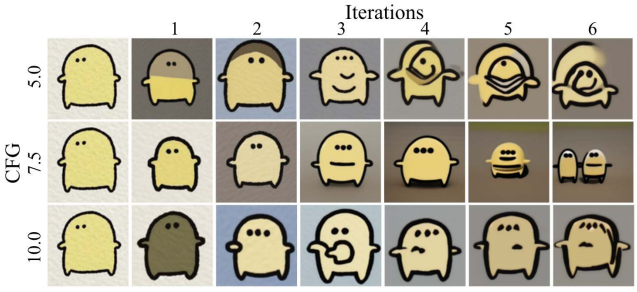


Figure 26: Chain of Diffusion of ReDiFine with different CFG scales on Kumapi dataset. ReDiFine successfully achieves robust image qualities for varying CFG scales.

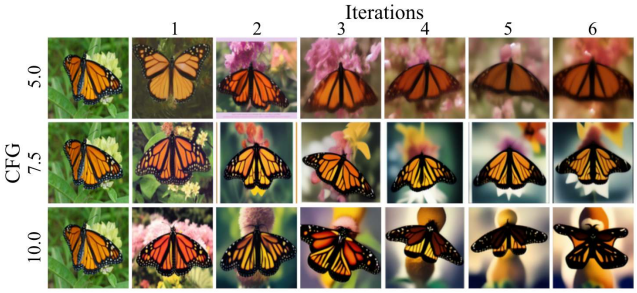


Figure 27: Chain of Diffusion of ReDiFine with different CFG scales on Butterfly dataset. ReDiFine successfully achieves robust image qualities for varying CFG scales.

This decline suggests that repeated hyperparameter searches are necessary to identify suitable CFG scales for subsequent iterations. Such an approach becomes increasingly impractical as the number of

iterations grows, highlighting the limitations of relying on the optimal CFG scale to mitigate model collapse.

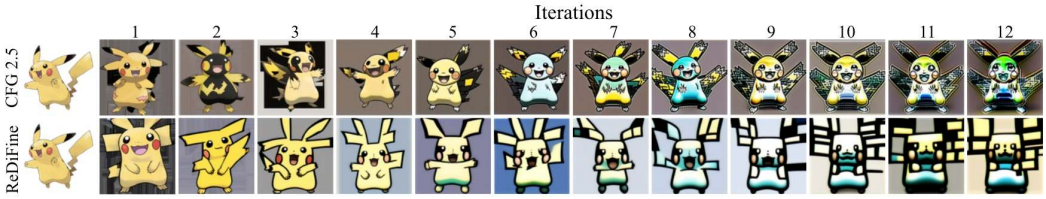


Figure 28: Chain of Diffusion of baseline with the optimal CFG scale and ReDiFine for more iterations. The generation quality of ReDiFine is preserved for additional iterations while optimally found CFG scale 2.5 fails to maintain the image qualities. Similar high-frequency degradation is observed.

E.3 Cross-domain Data

We additionally conduct experiments when the training dataset is the cross-domain set of four datasets (Pokemon, CelebA-1k, Kumapi, and Butterfly) to investigate whether having a broader range of concepts impact model collapse. The merged dataset serves as the original training set, while the combined captions are used for image generation at each iteration. The results, presented in Figure 29, show that despite the increased number of images and the inclusion of diverse concepts and domains, model collapse persists at both low and high CFG scales. Moreover, no single CFG scale (e.g., 1.5 or 2.5) can consistently produce high-quality, reusable images across all datasets, highlighting the limitations of relying on an optimal CFG scale for diverse domains. In contrast, ReDiFine leverages the increased conceptual diversity in the original training set, generating more reliable images across all datasets. To ensure a fair comparison, we control the number of training epochs to maintain consistent updates across experiments.

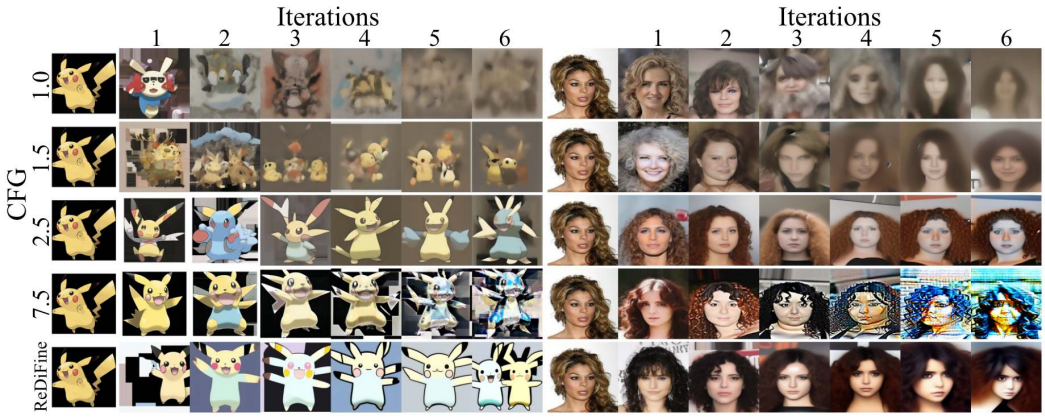


Figure 29: Chain of Diffusion with the multi-domain set of datasets (Pokemon, CelebA-1k, Kumapi, and Butterfly) used for finetuning. Model collapse remains evident at low and high CFG scales. A single CFG scale (e.g., 1.5 or 2.5) fails to achieve optimal performance across both the Pokemon and CelebA-1k datasets. In contrast, ReDiFine successfully generates high-quality, reusable images simultaneously. While images for Kumapi and Butterfly datasets are not displayed due to space constraints, they are included in the finetuning process along with the other datasets.

E.4 Iterative Retraining

Some prior works on model degeneration examine scenarios in which a single model is continually trained on synthetic data it has generated. To adapt our Chain of Diffusion framework to this setting, we consider a setup where the same model is finetuned iteratively across multiple iterations. At each iteration, the model generates a fixed number of images using a predefined prompt set, and these

generated images are then used to further finetune the model. Figure 30 demonstrates how images degrade under this setting across different CFG scales and ReDiFine. While severe degradation is observed for low and high CFG scales, the optimal CFG scale and ReDiFine are able to mitigate model collapse, generating high-quality images. This indicates that the effect of ReDiFine is maintained even when a single model is continually finetuned.

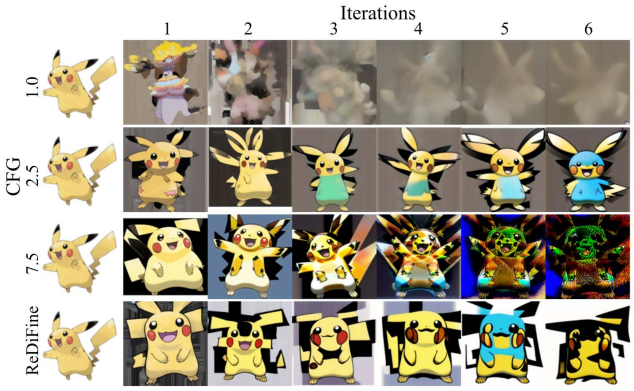


Figure 30: Chain of Diffusion where a single model is continually finetuned across multiple iterations. Model collapse is observed consistently at CFG scales of 1.0 and 7.5, while the baseline with a CFG scale of 2.5 and ReDiFine effectively mitigate model collapse. This demonstrates that model collapse is a universal phenomenon across different settings, and the effect of ReDiFine is effective robustly.

E.5 Quantitative results

In addition to visual inspections for images generated using ReDiFine, we compare the quantitative results of ReDiFine to baselines with different CFG scales across FID, CLIP score, SFD, and recall. Details about each metric are provided in Appendix A.3. DiNOv2 features are used to compute FID, recall, and SFD. We follow Kynkäänniemi et al. (2019) to compute recall and set the number of neighbors for computing recall 5. The results, shown in Figure 31, demonstrate that ReDiFine achieves performance comparable to the optimal CFG scales (2.5 for Pokemon and Kumapi, 1.5 for CelebA-1k and Butterfly) across different datasets and metrics.

F Ablation study

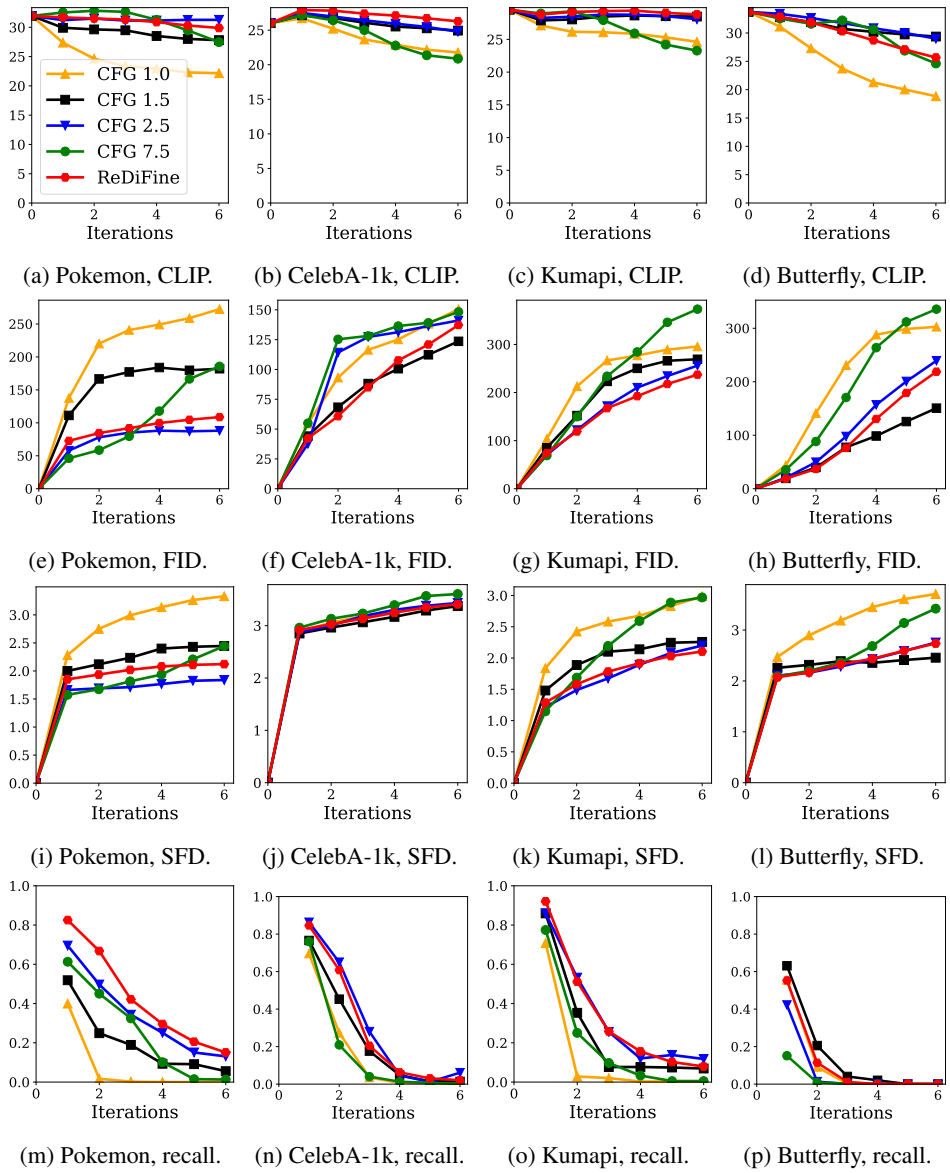
This section provides an ablation study to understand how condition drop finetuning and CFG scheduling contribute to the success of ReDiFine.

F.1 Condition drop finetuning

We conducted an ablation study to understand how the probability of dropping text embedding during finetuning affects the image quality in the Chain of Diffusion. We examine 0.1, 0.2, and 0.4 as Stable Diffusion is trained using 0.1 or 0.2. For both Pokemon and CelebA-1k datasets, a probability of 0.2 works the best, as shown in Figure 32 and Figure 33, respectively. Interestingly, condition drop finetuning helps to mitigate the color saturation problem, but its effect decreases with a higher probability. For both of these datasets, condition drop finetuning can mitigate image degradation to some degree, but still, there is a large quality degradation that needs to be improved.

F.2 CFG scheduling

We also evaluated how different CFG scale decreasing strategies impact image degradation in the Chain of Diffusion. We experimented with two different exponential decay rates and compared them with a linear decreasing strategy. Figure 34 demonstrates that CFG scheduling is effective for Pokemon dataset, generating high-quality images comparable to those generated by ReDiFine. However, as shown in Figure 35, it fails to enhance image quality on CelebA-1k dataset. This



(a) Pokemon, CLIP. (b) CelebA-1k, CLIP. (c) Kumapi, CLIP. (d) Butterfly, CLIP.

(e) Pokemon, FID. (f) CelebA-1k, FID. (g) Kumapi, FID. (h) Butterfly, FID.

(i) Pokemon, SFD. (j) CelebA-1k, SFD. (k) Kumapi, SFD. (l) Butterfly, SFD.

(m) Pokemon, recall. (n) CelebA-1k, recall. (o) Kumapi, recall. (p) Butterfly, recall.

Figure 31: Quantitative results of ReDiFine and baselines (different CFG scales).

highlights the necessity of condition drop finetuning for achieving universal improvements in the Chain of Diffusion across various datasets.

G Analysis

In this section, we present a series of analyses of images generated through the Chain of Diffusion. Specifically, we examine the distribution of latent values and the differences between conditional and unconditional scores. Additionally, we analyze the power spectra of the images using 2D Fourier transforms and explore fingerprints through forensic analysis (Corvi et al., 2023a;b).

G.1 Latent analysis

Figure 36a illustrates how the distribution of latent values evolves across different iterations. The histograms show the final latent vectors before decoding into pixel space, comparing various CFG

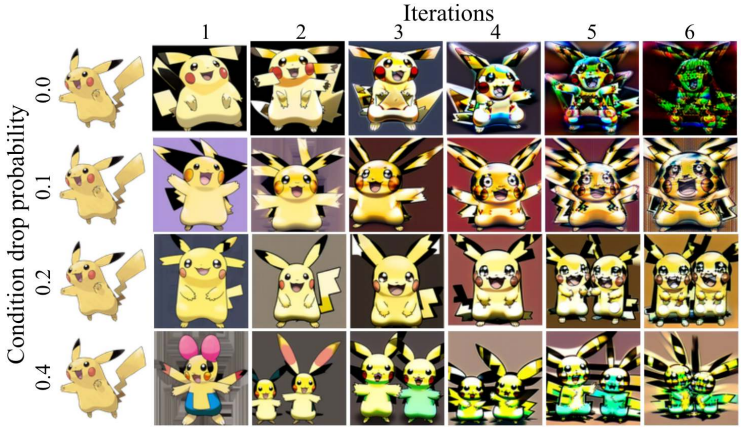


Figure 32: Chain of Diffusion with condition drop finetuning on Pokemon dataset.

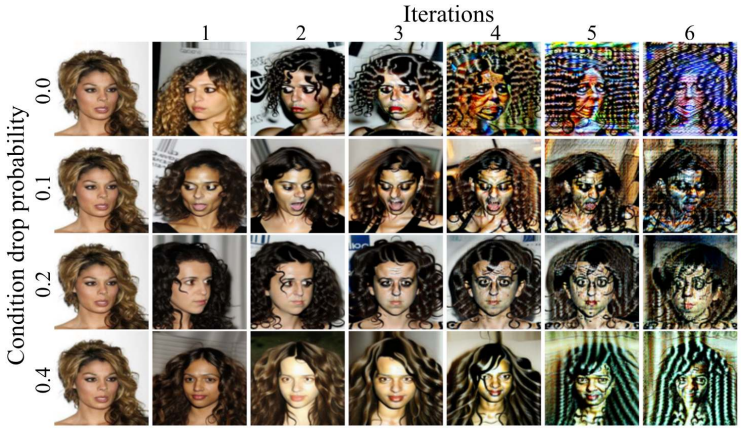


Figure 33: Chain of Diffusion with condition drop finetuning on CelebA-1k dataset.

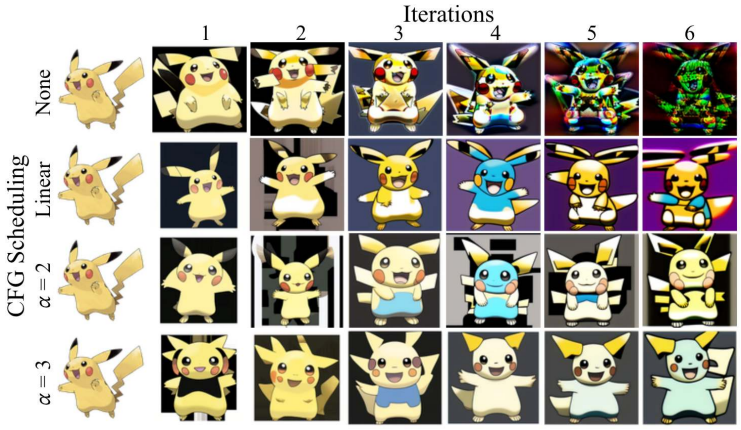


Figure 34: Chain of Diffusion with CFG scheduling on Pokemon dataset.

scales and ReDiFine. For a CFG of 1.0, the latent distribution rapidly converges into a Gaussian-like shape, with its variance shrinking over iterations. This behavior is consistent with previous work (Bertrand et al., 2023; Alemohammad et al., 2023; Dohmatob et al., 2024c), which theoretically predicted that the self-consuming loop progressively trims the tails of the distribution, reducing output diversity until it collapses to a single mode. We hypothesize that this narrowing in the latent space

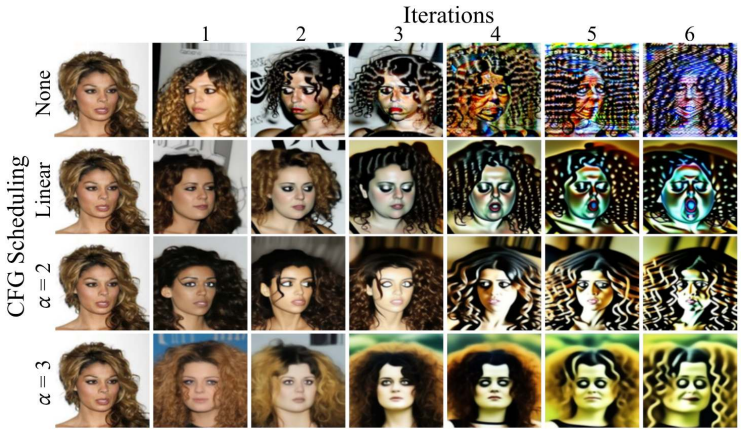
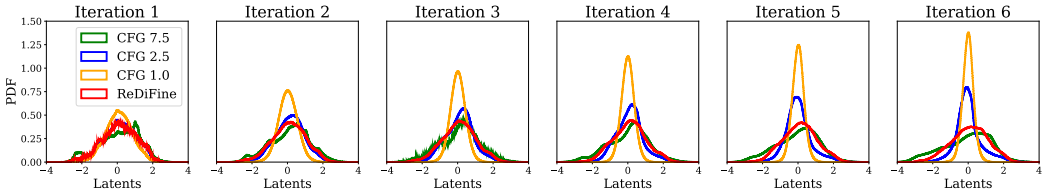
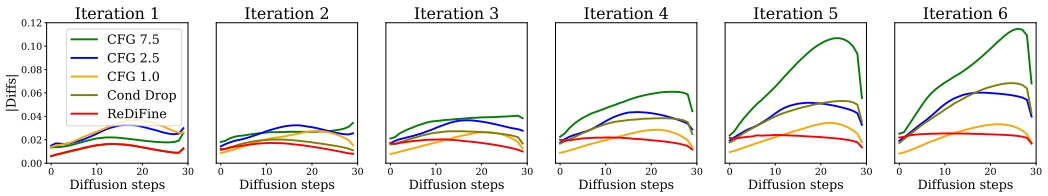


Figure 35: Chain of Diffusion with CFG scheduling on CelebA-1k dataset.



(a) Probability density function (PDF) of values of latent vectors at the last diffusion step.



(b) Norm of differences between conditional and unconditional scores during diffusion steps.

Figure 36: Histogram of latent values and Diffs during diffusion steps for Pokemon dataset. (a) **Latent distribution shrinks over iteration for low CFG and expands with high CFG.** Larger values in latent vectors are more likely to occur with high CFG, gradually increasing the tail of the distribution. (b) **Differences between conditional and unconditional scores increase as the training set is more degraded.** Especially, high differences in the later diffusion steps can be a cause of high-frequency degradation.

leads to blurrier, more homogeneous outputs in pixel space. Conversely, at a CFG scale of 7.5, the latent distribution develops longer tails and tends toward a more uniform spread across space. A CFG scale of 2.5, which demonstrates the best reusability among the three, better preserves the latent distribution over iterations. ReDiFine further enhances this preservation, maintaining the histogram from the first to the last iteration, thus achieving both high fidelity in the first iteration and better reusability.

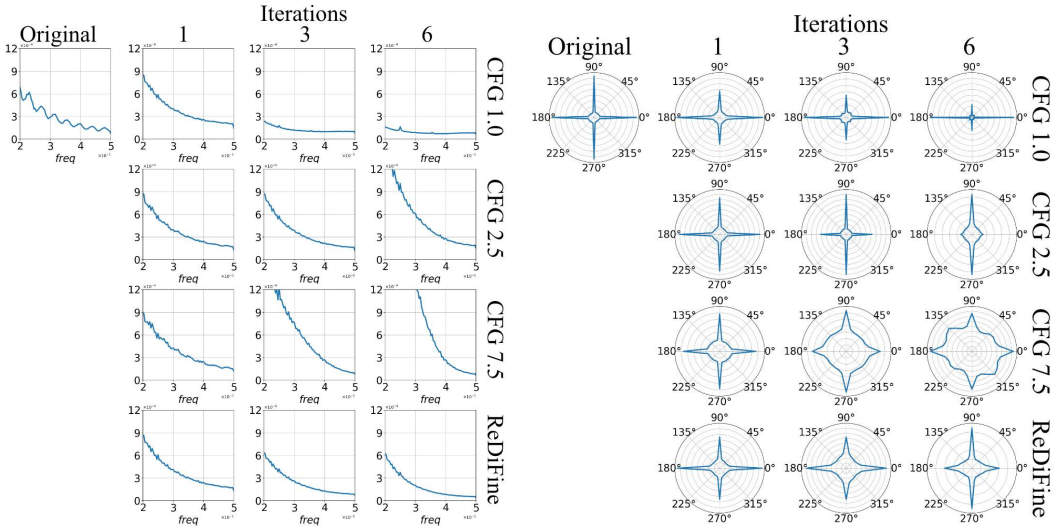
G.2 Differences between conditional and unconditional scores

Next, we plot the evolution of the average norm of Diff (= Cond Score − Uncond Score) across diffusion steps for different iterations in Figure 36b. In the first iteration, the highest Diff value is observed for CFG 1.0, followed by CFG 2.5 and CFG 7.5. This behavior can be interpreted as the models' adaptive behavior to preserve the values added to the latent vectors, Diff multiplied by CFG scale, at each step. However, this trend shifts in later iterations. The Diff value for CFG 7.5 continues to grow with each iteration, and by iteration 6, we see elevated Diff values throughout the entire diffusion steps, creating a significant gap compared to CFG 2.5 and 1.0. We conjecture that

this accumulation of Diff is the responsible for the high-frequency degradation in images generated with CFG 7.5. In contrast, the Diff value for CFG 1.0 remains relatively stable or even decreases across iterations. The deviation of Diff among different iterations is minimized by ReDiFine, which explains its ability to preserve image quality in later iterations. While condition drop finetuning helps reduce the Diff in the earlier iterations, it fails to prevent accumulations in later iterations. This limitation is also evident in the ablation study, where condition drop finetuning alone was insufficient to prevent model collapse. Notably, ReDiFine produces significantly smaller Diff values compared to the baseline with CFG scale 2.5, comparable to CFG scale 1.0 even when using a high CFG scale 7.5. This underscores the importance of combining condition drop finetuning with CFG scheduling.

G.3 Power spectra of 2D Fourier transforms

Figure 37 demonstrates the radial and angular spectrum power density of both the original and synthetic images. It is evident that ReDiFine closely maintains the radial spectrum power density of the original training set, whereas even a CFG scale 2.5 falls short. Additionally, ReDiFine demonstrates stable angular spectra throughout the Chain of Diffusion, even though they differ from those of the original training set. Pokemon dataset is used for power spectra analysis.



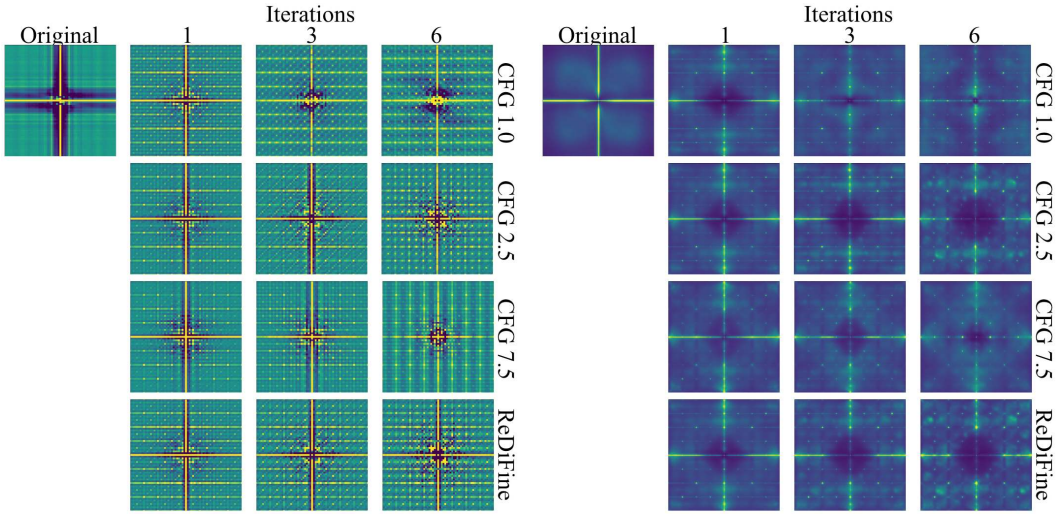
(a) Radial spectrum power density.

(b) Angular spectrum power density.

Figure 37: Power spectrum density of the original training set and synthetic sets for Pokemon dataset. **Images generated by ReDiFine maintain power density distribution during Chain of Diffusion while baselines fail. Even CFG scale** 2.5 **cannot maintain the distribution for the last iteration.** (a) Radial spectrum power density. ReDiFine shows a density distribution similar to that of the original training set. (b) Angular spectrum power density. Power density of generated images by ReDiFine remains during the iterations while baselines cannot maintain angular distribution.

G.4 Fingerprints for forensic analysis

Several works (Corvi et al., 2023a;b) aim to identify fingerprints of synthetic images. High-quality synthetic images from different generative models have clearly distinct fingerprints, showing the potential to be used for synthetic image detection. We analyze fingerprints of synthetic images for different CFG scales and iterations, and compare them to fingerprints of the original training set. Both autocorrelation and average power spectra show clear differences between the original training set and synthetic images, as shown in Figure 38. Moreover, how the fingerprints of synthetic images evolve throughout the Chain of Diffusion differ for ReDiFine and different CFG scales. Specifically, fingerprints of synthetic images from ReDiFine are similar to those of images from CFG scale 2.5, while other CFG scales (1.0 and 7.5) make fingerprints different from the first iteration as iterations proceed. Horizontal and vertical lines in autocorrelation gradually disappear and central regions in power spectra vary for further iterations. Also, the varying central regions in power spectra imply that



(a) Autocorrelation.

(b) Average power spectra.

Figure 38: **The fingerprints of the original training set and synthetic sets show clear differences, and ReDiFine produces fingerprints similar to CFG scale 2.5.** (a) Autocorrelation of image fingerprints. Horizontal and vertical lines gradually disappear for CFG scales 1.0 and 7.5 while they are maintained for CFG scale 2.5 and ReDiFine. (b) Average power spectra of images. Central regions are amplified or diminished for CFG scales 1.0 and 7.5, demonstrating low and high-frequency degradation.

low frequency features increase and decrease for CFG scale 1.0 and 7.5, respectively, aligning with visual inspections. Generating images with fingerprints similar to those of the original real images can be an interesting future direction to reduce the degradation in the Chain of Diffusion. Pokemon dataset is used for fingerprint analysis.