

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

Anonymous authors

Paper under double-blind review

ABSTRACT

The success of generative models has reached a unique threshold where their outputs are indistinguishable from real data, leading to the inevitable contamination of future data collection pipelines with synthetic data. While their potential to generate infinite samples initially offers promise for reducing data collection costs and addressing challenges in data-scarce fields, the severe degradation in performance has been observed when iterative loops of training and generation occur—known as “model collapse.” This paper explores a practical scenario in which a pretrained text-to-image diffusion model is finetuned using synthetic images generated from a previous iteration, a process we refer to as the “Chain of Diffusion.” We first demonstrate the significant degradation in image qualities caused by this iterative process and identify the key factor driving this decline through rigorous empirical investigations. 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. ReDiFine mitigates model collapse 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., 2024b), 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 fine-tuning pre-trained text-to-image diffusion models. An end user of text-to-image diffusion models often wants to fine-tune 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 fine-tuned diffusion models are uploaded regularly on platforms like CivitAI¹, 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. 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

¹<https://civitai.com/>

054 feel compelled to include more synthetic images in their datasets to keep up with the data demands of
055 these ever-growing, data-hungry models.

056
057 To this end, we conduct a thorough investigation into how various hyperparameters commonly used
058 during diffusion fine-tuning (e.g., learning rate, diffusion steps, prompts) impact model collapse.
059 From our extensive empirical analysis, we make a crucial observation: the classifier-free guidance
060 (CFG) scale is the most significant factor that influences the rate of model collapse. Moreover, we
061 observe a fascinating phenomenon that the direction of degradation varies with different CFG values—
062 low CFG results in low-frequency degradation in images, while high CFG leads to high-frequency
063 degradation. The critical role of CFG in determining both the rate and direction of model collapse is
064 a novel insight previously unrecognized in the literature.

065 To gain a deeper theoretical understanding of this phenomenon, we draw upon the concept of
066 quantitative trait modeling from statistical genetics. Unlike existing studies that attribute model
067 collapse to limited sample sizes and conclude it leads to zero variance (Alemohammad et al.,
068 2023; Bertrand et al., 2023; Shumailov et al., 2023), we propose that the underlying cause is a
069 truncation-based selection process modulated by CFG. This theoretical model accurately describes
070 our experimental observations. Moreover, we show that other theoretical work on model collapse can
071 also be connected to statistical genetics models, such as random genetic drift and its variants (Fisher,
072 1999; Wright, 1931; Kimura, 1955; Paris et al., 2019). This fresh angle of drawing parallels with
073 statistical genetics offers a new framework to understand the mechanism of model collapse.

074 Finally, inspired by our empirical results and theoretical insights, we introduce the Reusable Diffusion
075 Finetuning (ReDiFine) method. Our experiments show that selecting an appropriate CFG can
076 significantly slow down model collapse. However, finding the optimal CFG is computationally
077 expensive, requiring numerous iterations of finetuning. To address this, we propose ReDiFine as a
078 plug-and-play solution that achieves a comparable reduction in collapse rate without the need for any
079 hyperparameter tuning. The default ReDiFine setting effectively mitigates model collapse across all
080 four datasets we tested. Moreover, ReDiFine is effective at mitigating model collapse not only in a
081 worst-case scenario of having fully-synthetic dataset, but also in a more practical case, where we have
082 a mix of real and synthetic images the training set. Our contributions can be summarized as follows:

- 083 • We performed thorough and extensive empirical investigations into model collapse when we
084 finetune a diffusion model with diffusion-generated images. We tested an exhaustive list of
085 experimental parameters on four datasets (2 digital art & 2 natural images) to identify what affects
086 model collapse. Our analysis reveals that CFG scale is the most influential factor that not only
087 controls the rate of model collapse but also dictates the type of image degradation (Section 3).
- 088 • We provide a novel theoretical analysis of model collapse based on quantitative trait modeling that
089 can accurately predict how power spectra of generated images evolve over iterations (Section 4).
- 090 • We propose a simple yet effective strategy to mitigate model collapse, named ReDiFine, which
091 combines condition drop finetuning and CFG scheduling. Across all four datasets, ReDiFine
092 successfully generates more “reusable” images, which can be reused to finetune a diffusion model
093 without causing a severe model collapse (Section 5).

094 2 RELATED WORK

095
096 The self-consuming training loop and the associated phenomenon known as “model collapse” have
097 become significant areas of study in the past two years (Martínez et al., 2023a;b; Taori & Hashimoto,
098 2023; Alemohammad et al., 2023; Bohacek & Farid, 2023; Guo et al., 2024; Bertrand et al., 2023;
099 Dohmatob et al., 2024a; Briesch et al., 2023; Gillman et al., 2024; Fu et al., 2024; Marchi et al.,
100 2024). Model collapse, defined as “a degenerative process affecting generations of learned generative
101 models, where generated data end up polluting the training set of the next generation of models” in
102 Shumailov et al. (2023), has been observed in both language and image generative models.

103 Empirical studies on LLMs (Briesch et al., 2023) reveal that linguistic diversity collapses, especially
104 in high-entropy tasks (Guo et al., 2024), although this can be mitigated with data accumulation (Ger-
105 strasser et al., 2024). In image generation, several works (Martínez et al., 2023a;b; Alemohammad
106 et al., 2023; Hataya et al., 2023; Bohacek & Farid, 2023; Bertrand et al., 2023) note image degrada-
107 tion when diffusion models are recursively trained with self-generated data. We conduct extensive
empirical experiments to reveal the most significant factor causing model collapse in text-to-image

108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161

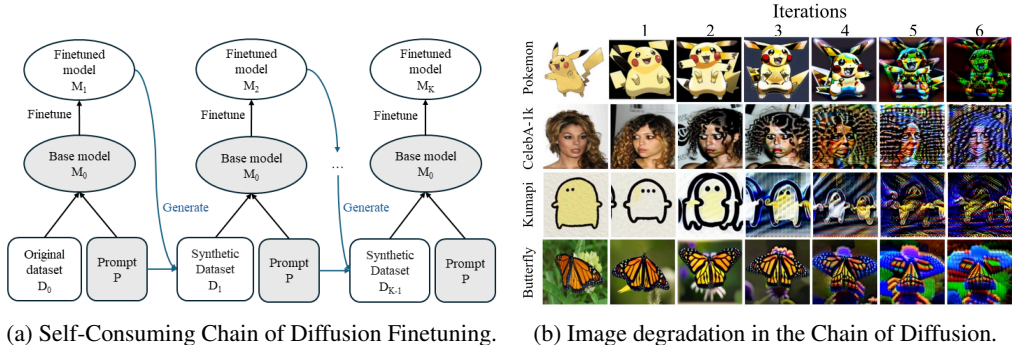


Figure 1: (a) **Overall pipeline of the Chain of Diffusion.** Given a pretrained text-to-image diffusion model M_0 and a prompt set P , a finetuned model M_k is trained using D_{k-1} generated at the previous iteration $k - 1$. Then, M_k generates D_k using the same prompt P , building a fully synthetic loop. Chain of diffusion begins with the original real dataset D_0 . (b) **Image degradation universally occurs during Chain of Diffusion across four datasets.** As the Chain of Diffusion progresses, the severity of image degradation intensifies, exhibiting universal patterns of highly saturated images containing repetitive high-frequency patterns. This consistently holds across four datasets and 10 scenarios that we comprehensively investigate in Section 3.2.

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 have largely focused on the reduction of diversity, typically framed as either decreasing covariance in continuous domains (Alemohammad et al., 2023; Shumailov et al., 2023; Bertrand et al., 2023) or shrinking support in discrete domain (Dohmatob et al., 2024b; Marchi et al., 2024). The finite number of generated samples per iteration has been identified as a key cause of model collapse by several authors (Bertrand et al., 2023; Shumailov et al., 2023; 2024; Dohmatob et al., 2024b; Fu et al., 2024), while others (Alemohammad et al., 2023; Ferbach et al., 2024; Marchi et al., 2024) emphasize that sampling bias reduces the generative model’s effective distribution. In contrast, we present a novel theoretical framework based on quantitative trait modeling, shifting the focus from variance reduction to mean drift induced by the selection process. We further demonstrate that many existing theories of model collapse can be understood as extensions of classical results from statistical genetics.

Regarding mitigation strategies for model collapse, existing works echo the importance of incorporating a large proportion of real data throughout the training loop (Alemohammad et al., 2023; Bertrand et al., 2023; Fu et al., 2024; Ferbach et al., 2024) or accumulating additional data (Gerstgrasser et al., 2024). The only mitigation strategy beyond altering the training data composition is proposed by Gillman et al. (2024), who suggests a self-correcting self-consuming loop using an expert model to correct synthetic outputs. While this approach is demonstrated in human motion generation with a physics simulator, having 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 mitigate model collapse more effectively than mixing more real data.

3 MODEL COLLAPSE IN SELF-CONSUMING CHAIN OF DIFFUSION

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) for finetuning a pretrained generative model. Then, M_{k+1} generates a set of images D_{k+1} for the next iteration using the prompt set P :

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

162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215

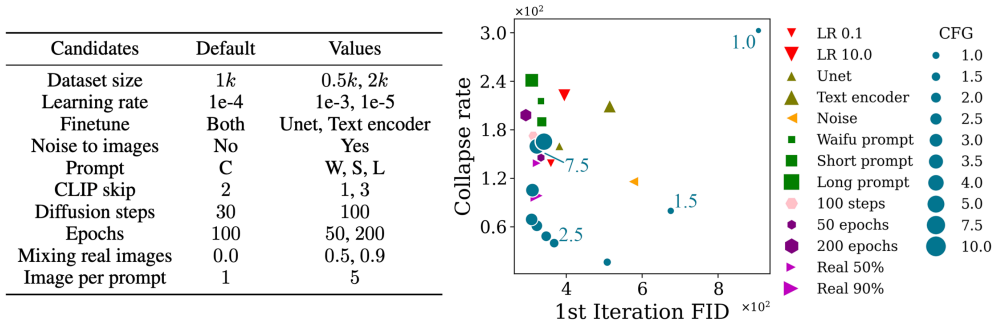


Figure 2: **Left:** Description of 10 potential factors (excluding CFG) that we examined as candidate sources for model collapse. All experiments were conducted on Pokemon except for the dataset size. For dataset size, we use CelebA since its original dataset is bigger, and we can subsample 500, 1000, and 2000 images. For prompt, we concatenate prompts with different lengths.² **Right:** 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

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

During the Chain of Diffusion, M_0 and P are fixed across all the iterations. To maintain the same size of training dataset, we generate one image per text prompt for all iterations. The overall pipeline of the Chain of Diffusion is shown in Figure 1a.

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 the Stable Diffusion at each iteration. We build our implementation on kohya-ss and perform experiments 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. During each iteration, we finetune the pretrained Stable Diffusion M_0 for 100 epochs. More implementation details can be found in Appendix A.1 and A.2.

Evaluation metrics. We use Frechet Inception Distance (FID) (Heusel et al., 2017) to measure image fidelity. Following Stein et al. (2024), we use DiNOv2 (Oquab et al., 2023) as a feature extractor for FID 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 features as the FID score.

In addition to evaluation metrics for generative models presented above, we propose a new metric to quantify the reusability of images generated in the Chain of Diffusion. We define collapse rate as the performance degeneration per iteration in the Chain of Diffusion:

$$\text{collapse rate} = \Delta FID = \frac{FID_K - FID_1}{K - 1}, \tag{3}$$

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 CLIP, or any other performance metric of interest. Note that a low collapse rate indicates more reusable images since it means that the model does not degrade much during the Chain of Diffusion and we have $K = 6$ in the rest of the paper.

²C, 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.

³We 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.

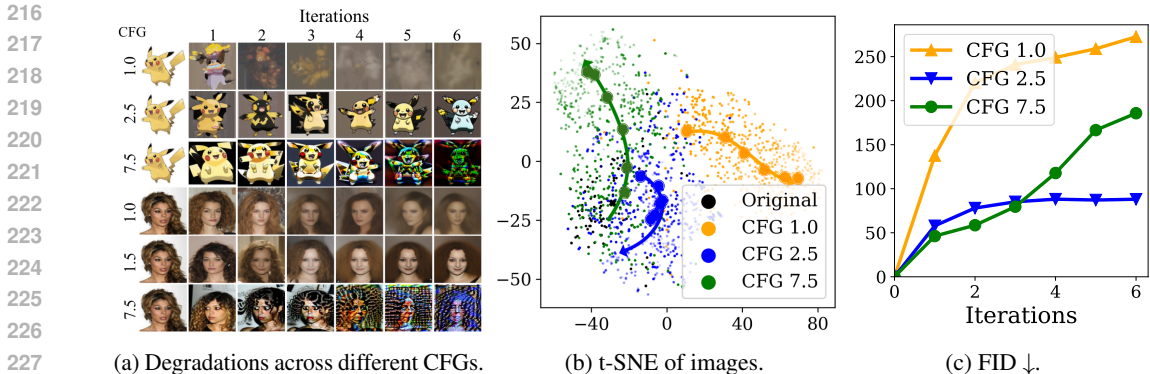


Figure 3: (a) **Low CFG leads to blurry images and high CFG leads to high-frequency degradation in the Chain of Diffusion.** CFG 2.5 for Pokemon and 1.5 for CelebA exhibit an ideal middle ground where both types of degradation slow down. More images in Appendix B. (b) t-SNE plot visualizes how generated 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. (c) Quantitative comparison for FID ↓ (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 (a).

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 (see Figure 1b) in the Chain of Diffusion. The quality of generated images begins to clearly deteriorate in the third iteration and it drops even more rapidly once the visible degradation emerges, reaching an unrecognizable level in two or three additional iterations. We 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 synthetic datasets 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 during finetuning, 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 investigated if the degradation happens for larger Stable Diffusion. *In all settings we tested, image degradation was universally present and very fast.* We plot the results for Pokemon on Figure 2 where y-axis is collapse rate and x-axis is the FID₁ (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 as shown in quantitative values (FID score has been doubled in 6 iterations). Visual inspection for images is provided in Appendix C.

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 in the speed of model collapse. CFG scale was first introduced in Ho & Salimans (2022) to modulate the strength between unconditional and conditional scores at each diffusion step as follows:

$$\text{Total Score} = \text{Unconditional Score} + \text{CFG} \cdot (\text{Conditional Score} - \text{Unconditional Score}). \quad (4)$$

High CFG means that we emphasize the conditional score for the given prompt more, 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.

In Figure 2, we observe 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 low collapse rate while maintaining a good quality in the first iteration as well. Moreover, Figure 3c presents FID for Pokemon to demonstrate how different CFG scales affect the performance. CFG scale 2.5 achieves the most robust performance for all metrics for Pokemon. 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 3a, 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 3b 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 and detailed results can be found in 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 the 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 clearly 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. Leveraging these assumptions, we derive a theorem showing that the mean trait value exhibits linear drift, while the variance stabilizes over time. We then show that this model can successfully capture the three key behaviors observed in Section 3: high-frequency degradation, low-frequency degradation, and optimal CFG scale performance. 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 modeling in statistical genetics explores the evolution of quantitative phenotypes (e.g., height, weight, or color), which are decided by multiple genetic and environmental factors

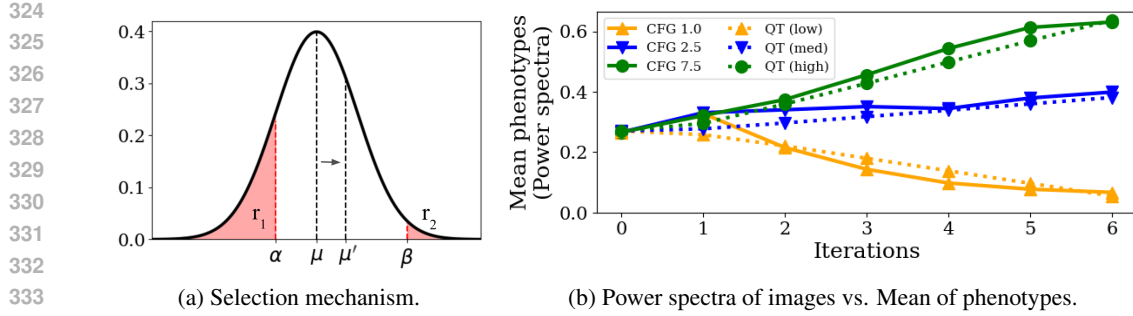


Figure 4: (a) **Selection mechanism with two-sided truncation.** r_1 and r_2 ratios of samples are truncated from the left and right tails, and the remaining $r = 1 - r_1 - r_2$ ratio of samples is selected. (b) **Power spectra for different CFGs (Pokemon) and quantitative trait modeling results of different 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.

in complex ways. These phenotypes 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 $\mathcal{N}(\mu_t, \sigma_{P,t}^2)$ where μ_t and $\sigma_{P,t}^2$ are the mean and variance of quantitative phenotypes. The phenotypic variance $\sigma_{P,t}^2$ is the sum of the (additive) genetic variance $\sigma_{G,t}^2$ and the environmental variance σ_E^2 , i.e., $\sigma_{P,t}^2 = \sigma_{G,t}^2 + \sigma_E^2$ (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 4a, 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) heritability⁵, 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:

$$h_t^2 = \frac{\sigma_{G,t}^2}{\sigma_{P,t}^2} = \frac{\sigma_{G,t}^2}{\sigma_{G,t}^2 + \sigma_E^2} = \frac{\mu_{t+1} - \mu_t}{\mu'_t - \mu_t}, \quad (5)$$

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 at current iteration μ_t , and heritability h_t^2 . Additionally, we assume the genetic variance for the next iteration is determined by the variance of selected individuals $\sigma_{G,t+1}^2 = \sigma_{P,t}^2$. We prove the behaviors of mean and variance of phenotypes under this setting:

Theorem 1 *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 $\frac{1}{1-c_2} \sigma_E^2$ 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 1 is provided in Appendix D.1.

4.2 EXPLAINING THE CHAIN OF DIFFUSION WITH QUANTITATIVE TRAIT MODELING

Just as phenotypes evolve through complex functions of hidden genotypes across multiple iterations in quantitative trait modeling, various continuous features of images evolve similarly during the

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

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

Chain of Diffusion. A key discovery in our work is that the high-frequency power spectra is a crucial phenotype that evolves over iterations and CFG scale acts as different selection mechanisms. These power spectra are computed using 2D Fourier transforms, focusing on frequencies above a certain threshold to capture high-frequency components. Different CFG scales during generation 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 4b shows the evolution of power spectra in the Chain of Diffusion (in solid line) and the phenotype mean modeled with Eqn. 9 (in dotted lines). The initial mean μ_0 and genetic deviation $\sigma_{G,0}$ in the simulation are set to match those of the original image set at iteration 0. Three different configurations of ratios, r_1 and r_2 , successfully model the power spectra distribution: CFG 7.5 is modeled as selecting from upper 5% of samples, CFG 2.5 selects 50% of the samples slightly favoring higher frequency, and CFG 1.0 results from selecting the lower 30%. We can see that this modeling can capture the evolution of power spectra very accurately over 6 iterations of Chain of Diffusion. Details about power spectra computation and simulation parameters are provided in Appendix D.2.2.

4.3 STATISTICAL GENETICS AS A UNIFYING LENS TO MODEL COLLAPSE

Quantitative trait modeling, a tool borrowed from statistical genetics, serves as a solid theoretical framework for interpreting our experimental results. As genetic processes and self-consuming training of generative models share a lot of similarities, we find that other concepts in mathematical genetics also have close ties to 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 effectively. However, the optimal CFG value that minimizes the collapse rate is different for each dataset (e.g., CFG 1.5 for CelebA, CFG 2.5 for Pokemon). There is no efficient way of finding an optimal CFG other than performing a grid search over a wide range of values and iteratively fine-tuning the diffusion model to evaluate the collapse rate for each configuration. In practical scenarios, it is unlikely that end users who are not machine learning experts will go through such a process during finetuning, just to prevent a potential model collapse that can happen in the next generation. This raises the question: *How can we design a finetuning and generation strategy that is user-friendly and can slow down model collapse effectively?*

To address this question, we again draw inspiration from the evolution process in nature. In biology, mutations naturally counteract genetic drift and preserve diversity. Furthermore, selection in nature is often a soft process rather than a hard truncation, as illustrated in Figure 4a. 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 one⁶. We propose **Reusable Diffusion Finetuning (ReDiFine)** method that incorporates these two ideas and show that it is highly effective at mitigating model collapse without the need for any finetuning.

⁶We 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 showing the image generation results with ReDiFine. We refer the curious readers to the appendix.

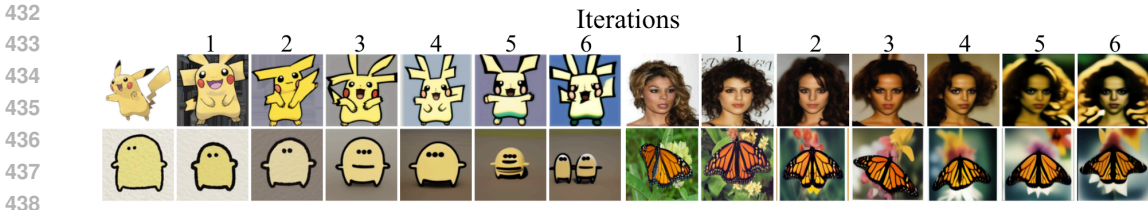


Figure 5: **ReDiFine effectively mitigates image degradation in the Chain of Diffusion.** ReDiFine successfully preserves the characteristics and features without further dataset-specific hyperparameter search. Artifacts observed in high-frequency degradation do not exist for all four datasets.

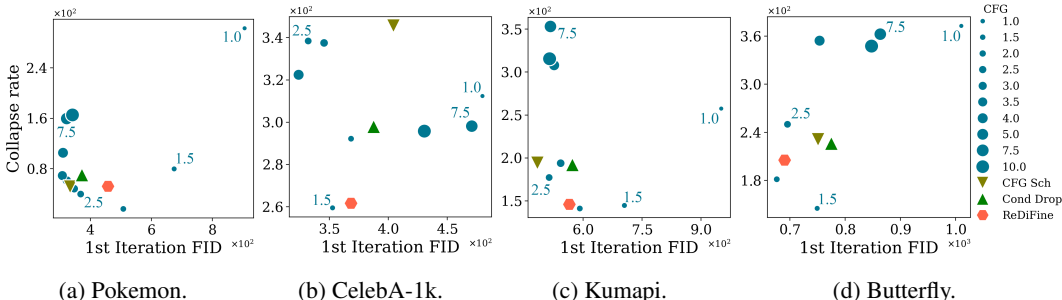


Figure 6: **ReDiFine effectively mitigates collapse-FID trade-off across all four datasets.** While the optimal CFG scale varies for different datasets, ReDiFine consistently achieves low collapse rate and FID at the same time (lower is better). Note that the differences in FID are relatively smaller than those in collapse rate, supporting the necessity to evaluate collapse rate in the Chain of Diffusion.

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 condition 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 38b).

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. High CFG lets us capture high-level semantic information accurately during the initial diffusion steps while lower CFG in later steps can prevent generating unnecessary high-frequency details.

ReDeFine Results. We present the generated images and quantitative metrics for ReDiFine. We use the initial CFG scale $s_0 = 7.5$, diffusion steps $T = 30$, decay rate $\alpha = 2$, and condition drop probability 0.2 for all of our experiments, except in the robustness comparison and ablation study. While ReDiFine can be further optimized through hyperparameter tuning (e.g., changing condition drop probability), our goal is to demonstrate that ReDiFine robustly mitigates model collapse with these default parameter settings for all datasets. In Figure 5, we can clearly see that ReDiFine significantly improves the image quality at later iterations compared to the baseline (Figure 1b) for all four datasets. In addition to the visual comparison, Figure 6 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 across all four datasets, demonstrating its effectiveness as a universal solution for mitigating model collapse. In contrast, using a fixed CFG scale that works well for one dataset often



Figure 7: **Comparison of ReDiFine and baseline (CFG scale 7.5) with training sets mixed with original and synthetic images from the previous iteration.** While mixing real images as much as 90% still results in model collapse for the baseline (as shown in the first row), the ability of ReDiFine to mitigate model collapse is even more highlighted by mixing real images (second row).

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 but performs poorly for Pokemon and Kumapi.

We also compare ReDiFine with real image mixing, which is suggested as a mitigation strategy in several previous papers (Alemohammad et al., 2023; Bertrand et al., 2023; Fu et al., 2024; Ferbach et al., 2024). In Figure 7, we can observe that even infusing the dataset with 90% real images, it shows severe image degradation within six iterations. By visual comparison, we can see that Pikachu images in Figure 5 have less degradation than those with real image mixing. Furthermore, applying ReDiFine to these mixed datasets, as shown in Figure 7, significantly improves image quality compared to the baseline. This demonstrates that ReDiFine is an effective solution for various settings and can be used alongside other mitigation strategies such as real data mixing.

Ablation study & Further analyses. We conducted an ablation study to understand the contributions of condition drop finetuning and CFG scheduling to the success of ReDiFine. We plot the results of using only condition drop and CFG scheduling in green triangles on Figure 6. 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, using either one of them showed significantly worse performance than ReDiFine. These results suggest that combining both 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.

Reusable image generation as a responsible AI practice. In this section, we demonstrated that ReDiFine significantly slows the collapse rate without sacrificing image quality, similar to how fair classifiers reduce bias without compromising accuracy (Alghamdi et al., 2022). Adopting ReDiFine is a responsible, environment-conscious practice to prevent polluting the world with images that might look visually appealing, but totally unusable to improve future AI models.

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 an effective mitigation strategy that generates reusable images for future training.

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

- 540
541
542 David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann
543 machines. *Cognitive science*, 9(1):147–169, 1985.
- 544
545 Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein
546 Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G Baraniuk. Self-consuming generative
547 models go mad. *arXiv preprint arXiv:2307.01850*, 2023.
- 548
549 Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, Peter Michalak, Shahab Asoodeh, and
550 Flavio Calmon. Beyond adult and compas: Fair multi-class prediction via information projection.
Advances in Neural Information Processing Systems, 35:38747–38760, 2022.
- 551
552 Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky,
553 Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will
554 Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael
555 Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos,
556 Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian
557 Puhersch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil
558 Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou,
559 Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. PyTorch 2: Faster
560 Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation.
561 In *29th ACM International Conference on Architectural Support for Programming Languages and
562 Operating Systems, Volume 2 (ASPLOS '24)*. ACM, April 2024. doi: 10.1145/3620665.3640366.
URL <https://pytorch.org/assets/pytorch2-2.pdf>.
- 563
564 Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten
565 Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with
566 an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.
- 567
568 Quentin Bertrand, Avishek Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel.
569 On the stability of iterative retraining of generative models on their own data. *arXiv preprint
arXiv:2310.00429*, 2023.
- 570
571 Matyas Bohacek and Hany Farid. Nepotistically trained generative-ai models collapse. *arXiv preprint
arXiv:2311.12202*, 2023.
- 572
573 Martin Briesch, Dominik Sobania, and Franz Rothlauf. Large language models suffer from their own
574 output: An analysis of the self-consuming training loop. *arXiv preprint arXiv:2311.16822*, 2023.
- 575
576 Brian Charlesworth. How long does it take to fix a favorable mutation, and why should we care? *The
577 American Naturalist*, 195(5):753–771, 2020.
- 578
579 Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing
580 properties of synthetic images: from generative adversarial networks to diffusion models. In
581 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.
582 973–982, 2023a.
- 583
584 Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa
585 Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP
586 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,
pp. 1–5. IEEE, 2023b.
- 587
588 Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. Model Collapse Demystified: The Case of
589 Regression, April 2024a. URL <http://arxiv.org/abs/2402.07712>. arXiv:2402.07712
590 [cs, stat].
- 591
592 Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model
593 collapse as a change of scaling laws. *arXiv preprint arXiv:2402.07043*, 2024b.
- DS Falconer. *Introduction to quantitative genetics*. Pearson Education India, 1996.

- 594 Damien Ferbach, Quentin Bertrand, Avishek Joey Bose, and Gauthier Gidel. Self-consuming
595 generative models with curated data provably optimize human preferences. *arXiv preprint*
596 *arXiv:2407.09499*, 2024.
- 597
598 Ronald Aylmer Fisher. *The genetical theory of natural selection: a complete variorum edition*.
599 Oxford University Press, 1999.
- 600 Shi Fu, Sen Zhang, Yingjie Wang, Xinmei Tian, and Dacheng Tao. Towards theoretical understandings
601 of self-consuming generative models. *arXiv preprint arXiv:2402.11778*, 2024.
- 602
603 Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes,
604 Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, et al. Is model collapse in-
605 evitable? breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint*
606 *arXiv:2404.01413*, 2024.
- 607 Nate Gillman, Michael Freeman, Daksh Aggarwal, Chia-Hong Hsu, Calvin Luo, Yonglong Tian, and
608 Chen Sun. Self-Correcting Self-Consuming Loops for Generative Model Training, April 2024.
609 URL <http://arxiv.org/abs/2402.07087>. arXiv:2402.07087 [cs, stat].
- 610
611 Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. The Curious Decline of
612 Linguistic Diversity: Training Language Models on Synthetic Text, April 2024. URL <http://arxiv.org/abs/2311.09807>. arXiv:2311.09807 [cs].
- 613
614 Hakurei. Waifu diffusion v1.4. [https://huggingface.co/hakurei/
615 waifu-diffusion-v1-4](https://huggingface.co/hakurei/waifu-diffusion-v1-4), 2022. Accessed: 2023-04-18.
- 616
617 Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will Large-scale Generative Models Corrupt Future
618 Datasets? In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20498–
619 20508, Paris, France, October 2023. IEEE. ISBN 9798350307184. doi: 10.1109/ICCV51070.
620 2023.01879. URL <https://ieeexplore.ieee.org/document/10376575/>.
- 621
622 Zhangyi He, Mark Beaumont, and Feng Yu. Effects of the ordering of natural selection and population
623 regulation mechanisms on wright-fisher models. *G3: Genes, Genomes, Genetics*, 7(7):2095–2106,
624 2017.
- 625
626 Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans
627 trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural
628 information processing systems*, 30, 2017.
- 629
630 Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*,
631 2022.
- 632
633 Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,
634 and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint*
635 *arXiv:2106.09685*, 2021.
- 636
637 Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han.
638 Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022.
- 639
640 Ihelon. Illustrations kumapi390, 2022. URL [https://www.kaggle.com/datasets/
641 ihelon/illustrations-kumapi390](https://www.kaggle.com/datasets/ihelon/illustrations-kumapi390). Accessed: 2023-04-19.
- 642
643 Ingemar Kaj, Carina F Mugal, and Rebekka Müller-Widmann. A wright–fisher graph model and the
644 impact of directional selection on genetic variation. *Theoretical Population Biology*, 159:13–24,
645 2024.
- 646
647 Motoo Kimura. Solution of a process of random genetic drift with a continuous model. *Proceedings
648 of the National Academy of Sciences*, 41(3):144–150, 1955.
- 649
650 kohya-ss. kohya-ss trainer. URL <https://github.com/kohya-ss/sd-scripts>.
- 651
652 Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved
653 precision and recall metric for assessing generative models. *Advances in neural information
654 processing systems*, 32, 2019.

- 648 Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image
649 pre-training for unified vision-language understanding and generation, 2022. URL <https://arxiv.org/abs/2201.12086>.
650
- 651 Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In
652 *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
653
- 654 Jay L Lush. *Animal breeding plans*. Read Books Ltd, 2013.
655
- 656 Matteo Marchi, Stefano Soatto, Pratik Chaudhari, and Paulo Tabuada. Heat death of generative
657 models in closed-loop learning. *arXiv preprint arXiv:2404.02325*, 2024.
- 658 Gonzalo Martínez, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juárez, and
659 Rik Sarkar. Combining generative artificial intelligence (ai) and the internet: Heading towards
660 evolution or degradation? *arXiv preprint arXiv:2303.01255*, 2023a.
- 661 Gonzalo Martínez, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juárez, and Rik
662 Sarkar. Towards understanding the interplay of generative artificial intelligence and the internet. In
663 *International Workshop on Epistemic Uncertainty in Artificial Intelligence*, pp. 59–73. Springer,
664 2023b.
665
- 666 Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov,
667 Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning
668 robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- 669 Cyriel Paris, Bertrand Servin, and Simon Boitard. Inference of selection from genetic time series using
670 various parametric approximations to the wright-fisher model. *G3: Genes, Genomes, Genetics*, 9
671 (12):4073–4086, 2019.
672
- 673 Pokémon. Pokédex. <https://www.pokemon.com/us/pokedex>, 2023. Accessed: 2023-04-
674 19.
- 675 Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-
676 resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF confer-
677 ence on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- 678 Iliia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Ander-
679 son. The curse of recursion: Training on generated data makes models forget. *arXiv preprint
680 arXiv:2305.17493*, 2023.
681
- 682 Iliia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai
683 models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.
- 684 StabilityAI. Sd vae ft mse original, 2022. URL [https://huggingface.co/stabilityai/
685 sd-vae-ft-mse-original](https://huggingface.co/stabilityai/sd-vae-ft-mse-original). Accessed: 2023-04-19.
686
- 687 George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Vilecroze,
688 Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of
689 generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in
690 Neural Information Processing Systems*, 36, 2024.
- 691 Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking
692 the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer
693 vision and pattern recognition*, pp. 2818–2826, 2016.
- 694 Rohan Taori and Tatsunori Hashimoto. Data feedback loops: Model-driven amplification of dataset
695 biases. In *International Conference on Machine Learning*, pp. 33883–33920. PMLR, 2023.
696
- 697 Paula Tataru, Maria Simonsen, Thomas Bataillon, and Asger Hobolth. Statistical inference in the
698 wright–fisher model using allele frequency data. *Systematic biology*, 66(1):e30–e46, 2017.
- 699 Veeralakrishna. Butterfly dataset. [https://www.kaggle.com/datasets/
700 veeralakrishna/butterfly-dataset](https://www.kaggle.com/datasets/veeralakrishna/butterfly-dataset), 2020. Accessed: 2024-01-03.
701
- Sewall Wright. Evolution in mendelian populations. *Genetics*, 16(2):97, 1931.

702 APPENDIX

703

704 A EXPERIMENTAL SETUP

705

706

707 A.1 HYPERPARAMETERS

708 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.

714

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

716

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

735

736

737 A.2 DATASETS

738

739 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.

744

745 **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.

746

747

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

749

750


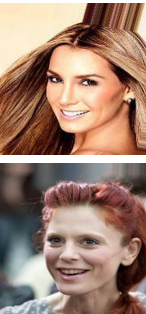

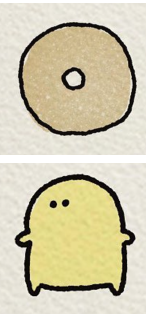
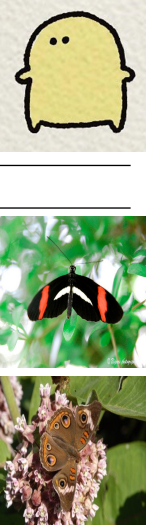
751 **Kumapi.** Kumapi (Ihelon, 2022) includes 391 handwriting-style images, with Waifu-generated prompts and manual adjustments.

752

753

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

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

756	Table 2: Sample images and prompts for Pokemon, CelebA-1k, Kumapi, and Butterfly datasets.	
757		
758	Pokemon	
759	<p data-bbox="297 325 1154 499">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 leaf, 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</p>	
760	<p data-bbox="297 495 1154 674">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 head, 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</p>	
761		
762		
763		
764		
765		
766		
767		
768		
769		
770		
771	CelebA-1k	
772	<p data-bbox="297 858 1154 896">a woman with brown hair smiling and posing for a picture in front of a mirror and gold and white stripes</p>	
773		
774		
775		
776		
777		
778		
779	<p data-bbox="297 968 1154 984">a woman with a very long red hair smiles and laughs on a city street and other people in the background</p>	
780		
781		
782		
783	Kumapi	
784		
785		
786	<p data-bbox="297 1092 1154 1108">solo, simple background, food, donut, grey background, no humans, food focus, still life, Kumapi style</p>	
787		
788		
789		
790		
791	<p data-bbox="297 1155 1154 1171">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</p>	
792		
793		
794		
795	Butterfly	
796		
797	<p data-bbox="297 1266 1154 1283">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</p>	
798		
799		
800		
801	<p data-bbox="297 1316 1154 1333">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.</p>	
802		
803		
804		
805		
806		
807		
808		
809		

B DIFFERENT CFG SCALES

This section presents images from Chain of Diffusion at various CFG scales for the four datasets. Figure 8, 9, 10, and 11 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.

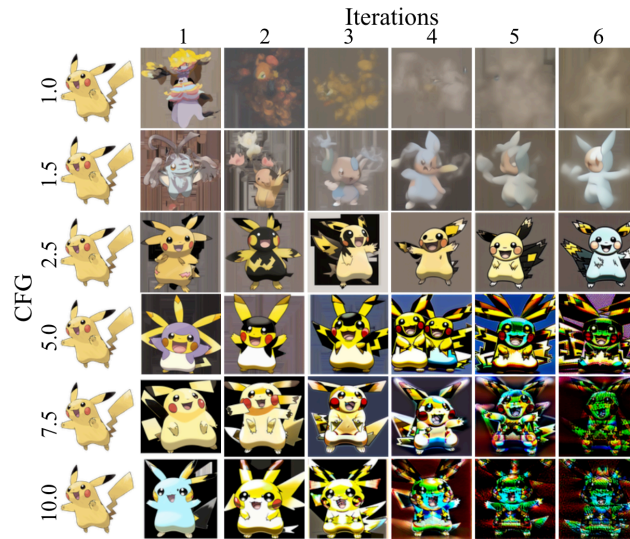


Figure 8: Chain of Diffusion for Pokemon at various CFG scales. The optimal CFG scale is 2.5.

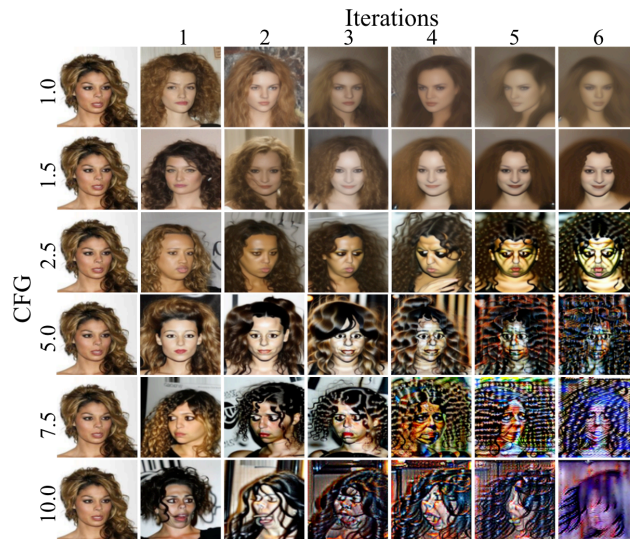
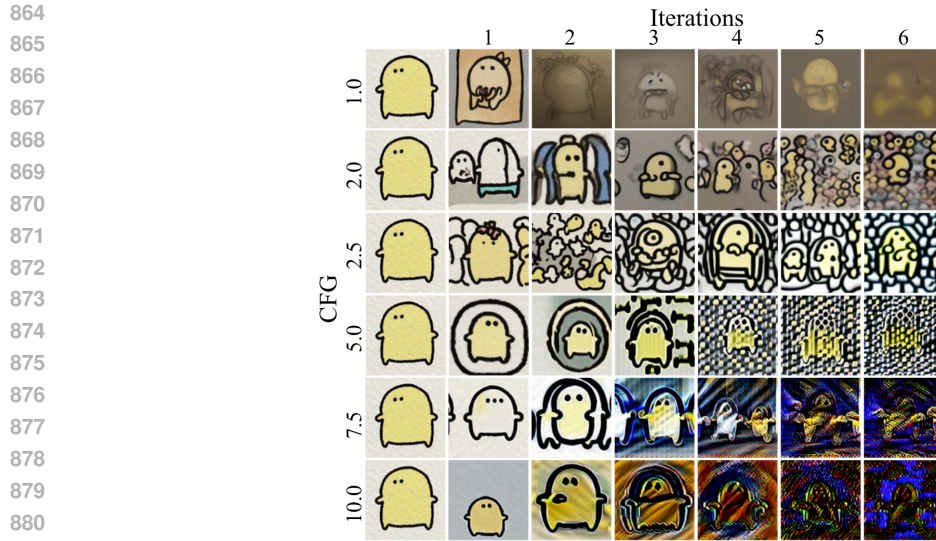


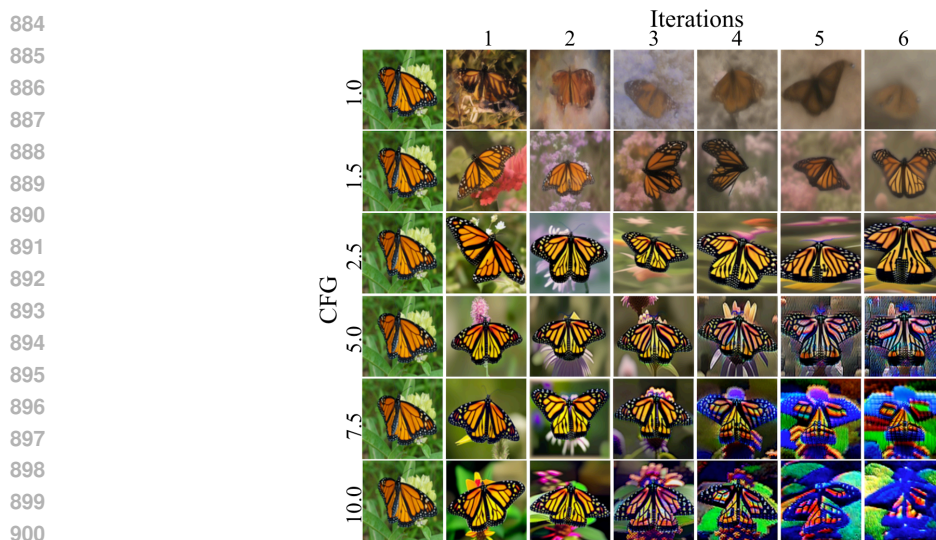
Figure 9: Chain of Diffusion for CelebA-1k at various CFG scales. The optimal CFG scale is 1.5.

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 Table ??,



882 Figure 10: Chain of Diffusion for Kumapi at various CFG scales. The optimal CFG scale is 2.5.



902 Figure 11: Chain of Diffusion for Butterfly at various CFG scales. The optimal CFG scale is 1.5.

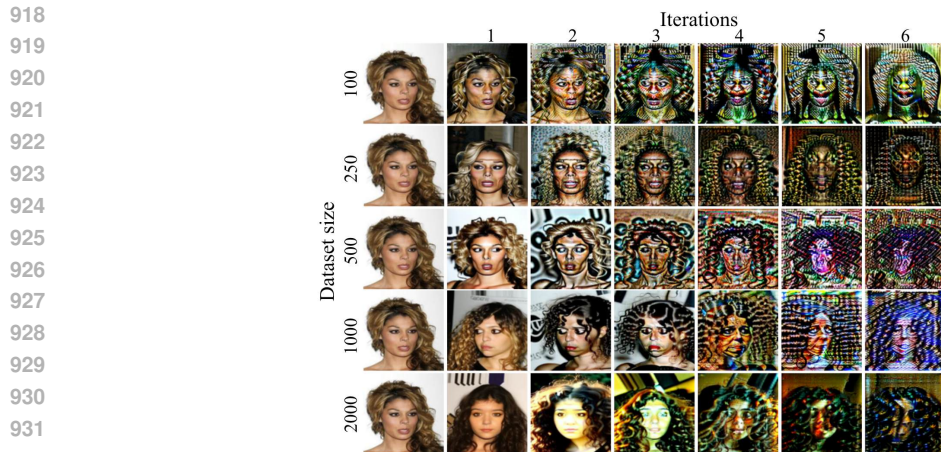
904 using the default settings from Table 1, to assess their impact on degradation. The CFG scale is fixed
905 at 7.5 for all cases.

907 C.1 TRAINING SET SIZE

908
909 We used CelebA dataset (Liu et al., 2015) to examine how training set size (both D_0 and D_k) impacts
910 degradation in the Chain of Diffusion. By subsampling, we adjusted the training set to 100, 250, 500,
911 and 2000 images. Degradation occurs regardless of dataset size, as shown in Figure 12, but appears
912 earlier with smaller sets. By the 6th iteration, images degrade severely for all cases. The number of
913 parameter updates was kept constant across all dataset sizes.

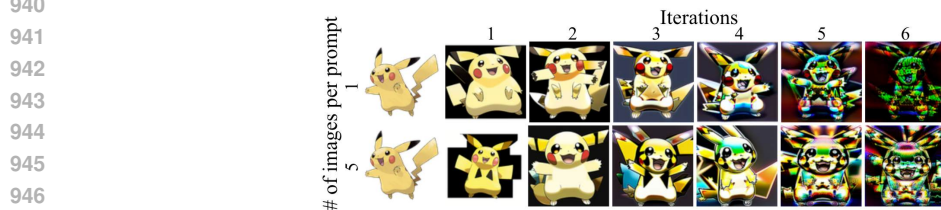
915 C.2 NUMBER OF IMAGES PER PROMPT

916
917 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



933 Figure 12: Chain of Diffusion on CelebA dataset with varying training set sizes. Degradation occurs
934 faster with smaller sets, but all result in severe degradation by the 6th iteration.

935
936
937 generating 5 times more images. As shown in Figure 13, while degradation is slightly delayed (by one
938 iteration), it remains unmitigated, and the high computational cost makes this approach impractical.



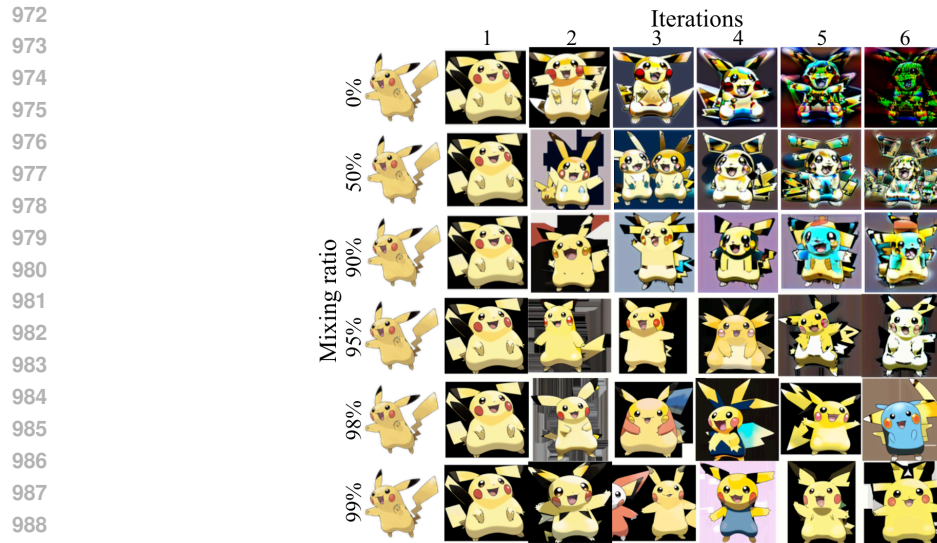
948 Figure 13: Chain of Diffusion on Pokemon dataset with multiple images generated per prompt.
949 Increasing the training set size in this way does not mitigate degradation.

952 C.3 MIXING REAL IMAGES TO SYNTHETIC SETS

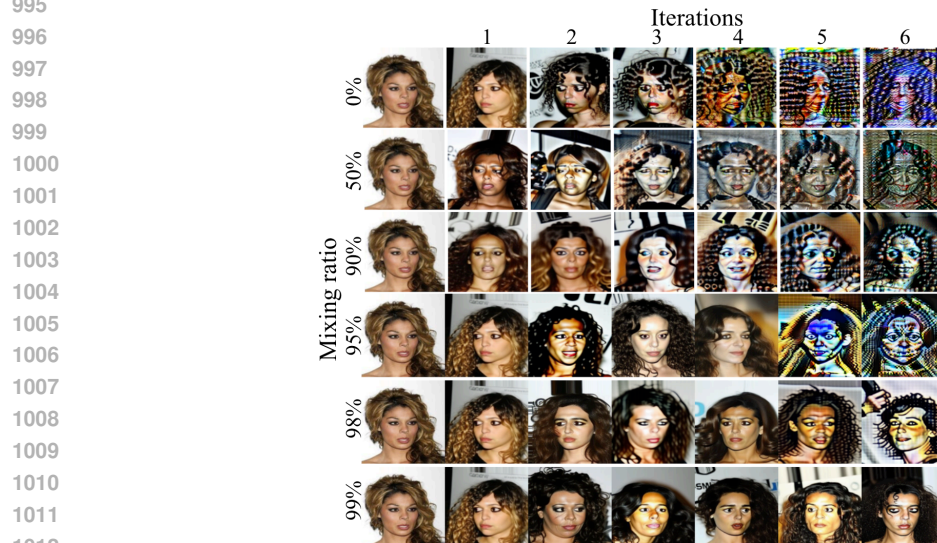
954 Many previous works suggest augmenting the training set with real images to mitigate degradation
955 during iterative training. We investigated whether mixing images from the original training set into
956 the synthetic set at each iteration could alleviate this issue. At each iteration, images are randomly
957 replaced with corresponding real images. Figure 14 and 15 show how degradation in the Chain
958 of Diffusion varies when 50%, 90%, 95%, 98% and 99% of images are replaced for Pokemon
959 and CelebA-1k datasets, respectively. Notably, even 5% synthetic images are sufficient to induce
960 degradation, and 50% replacement rarely slows it down. CelebA-1k dataset appears to be significantly
961 more susceptible to degradation.

963 C.4 PROMPT SET

964
965 We hypothesized that the descriptiveness of prompts influences degradation in the Chain of Diffusion.
966 We tested various prompt sets for Pokemon and CelebA-1k datasets. For Pokemon dataset, the default
967 prompt set consists of concatenated Waifu and BLIP captions, with BLIP captions ranging from
968 50 to 75 words. Figure 16 shows how using only Waifu prompts and varying BLIP caption lengths
969 affect degradation. Notably, different styles of high-frequency degradation were observed; shorter
970 prompts reduced repetitive patterns but decreased diversity. In CelebA-1k dataset, varying BLIP
971 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 17.



990 Figure 14: Chain of Diffusion on Pokemon dataset with real images randomly replacing synthetic
991 images at each iteration. A 50% replacement rarely slows degradation, while 10% synthetic images
992 are sufficient to initiate it.

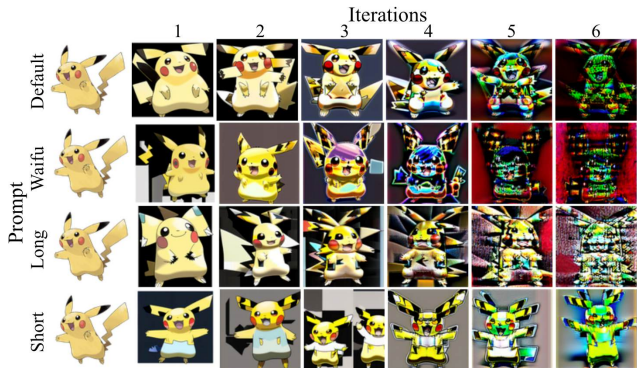


1013 Figure 15: Chain of Diffusion on CelebA-1k dataset with real images randomly replacing synthetic
1014 images at each iteration. A 50% replacement rarely slows degradation, while 5% synthetic images
1015 are sufficient to initiate it. It suffers from more severe degradation than Pokemon dataset as compared
1016 with Figure 14.

1020 C.5 U-NET AND TEXT-ENCODER

1021
1022
1023 Figure 18 illustrates the Chain of Diffusion with either the U-Net or text encoder finetuned. When
1024 the text encoder is not updated (second row), similar degradation occurs. However, the degradation
1025 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.

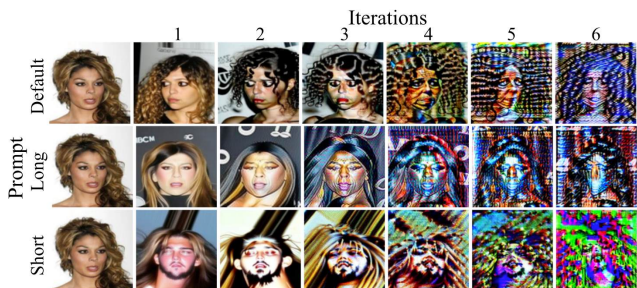
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037



1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052

Figure 16: 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.

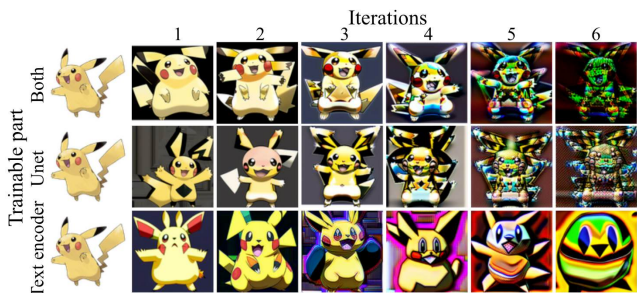
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052



1053
1054
1055
1056
1057

Figure 17: 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).

1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068



1069
1070

Figure 18: Chain of Diffusion on Pokemon dataset with either the U-Net or text encoder finetuned.

1071
1072
1073
1074

1075 C.6 NUMBER OF DIFFUSION STEPS

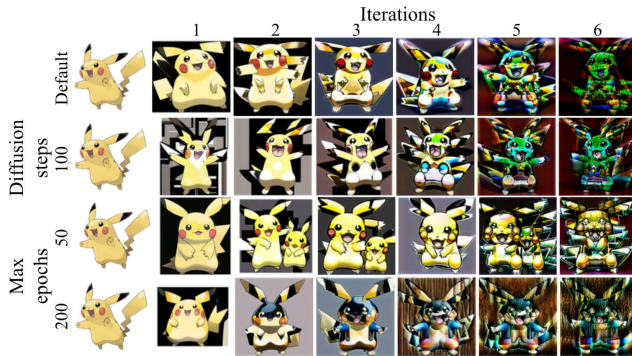
1076
1077
1078
1079

We investigated whether an insufficient number of diffusion steps during generation contributes to degradation. Figure 19 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 19, 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.

1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091



1092 Figure 19: Chain of Diffusion on Pokemon dataset with varying diffusion steps and training epochs.
1093 Both increased diffusion steps and differing training epochs fail to mitigate degradation, resulting in
1094 similar patterns.

1095
1096

1097 C.8 LEARNING RATE

1098

1099 Similarly, we assessed finetuning adequacy in Figure 20 by adjusting the learning rates for the U-Net
1100 and text encoder by $\times 10$ and $\times 0.1$. Images from the initial iterations show that the default values
1101 are suitable for finetuning. Although the styles are different, degradation consistently occurs across
1102 different learning rates.

1103

1104

1105

1106

1107

1108

1109

1110

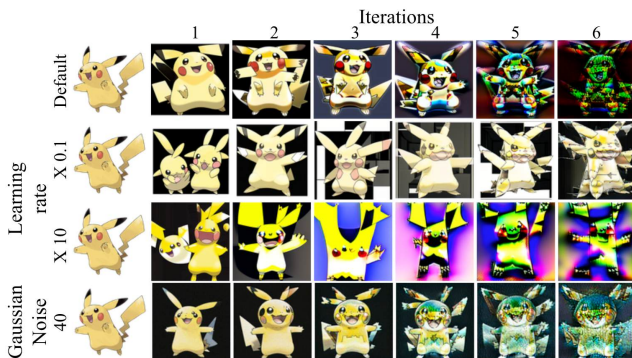
1111

1112

1113

1114

1115



1116 Figure 20: Chain of Diffusion on Pokemon dataset with varying learning rates and added Gaussian
1117 noise to the original training set.

1118

1119

1120

1121 C.9 CLIP SKIP

1122

1123 We investigated the impact of the CLIP skip hyperparameter on degradation. The CLIP skip de-
1124 termines which intermediate feature from the CLIP text encoder is used as the text embedding for
1125 conditional generation, with smaller values selecting features closer to the output and larger values
1126 selecting those nearer to the input text. As shown in Figure 21, this hyperparameter has minimal
1127 effect on degradation patterns.

1128

1129 C.10 ADDING GAUSSIAN NOISE TO THE ORIGINAL TRAINING SET

1130

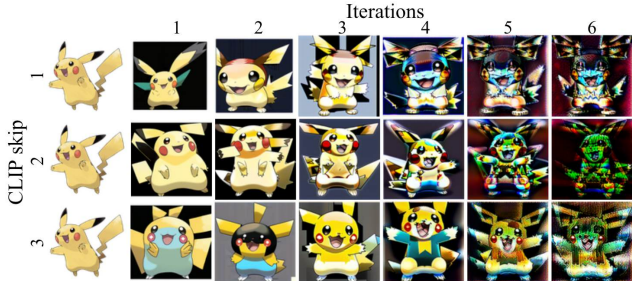
1131

1132

1133

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 20 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.

1134
1135
1136
1137
1138
1139
1140
1141
1142
1143



1144 Figure 21: Chain of Diffusion on Pokemon dataset using different CLIP skip hyperparameters. The
1145 CLIP skip hyperparameter shows a negligible effect on degradation.

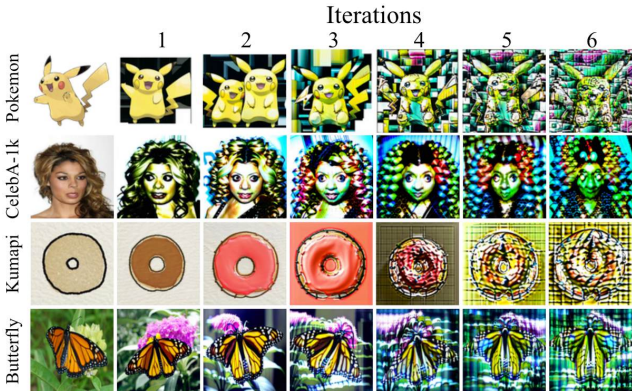
1146
1147

C.11 STABLE DIFFUSION XL

1148
1149
1150
1151
1152
1153

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 22.

1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166



1167 Figure 22: Chain of Diffusion of SDXL on four datasets. Due to insufficient investigation into optimal
1168 hyperparameters for finetuning SDXL, our experiments largely rely on those from Stable Diffusion
1169 v1.5.

1170
1171

C.12 ITERATION ACCUMULATION

1172
1173
1174
1175
1176
1177

Iteration 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.

D QUANTITATIVE TRAIT MODELING

D.1 PROOF OF THEOREM 1

1182

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

1183

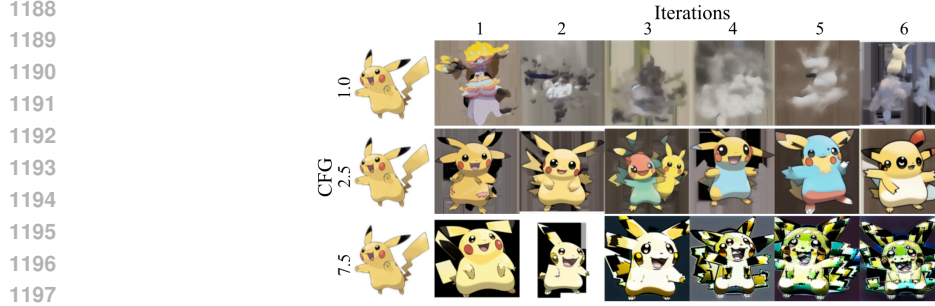
$$\Delta\mu_t = \mu_{t+1} - \mu_t = h_t^2(\mu'_t - \mu_t). \tag{6}$$

1184
1185

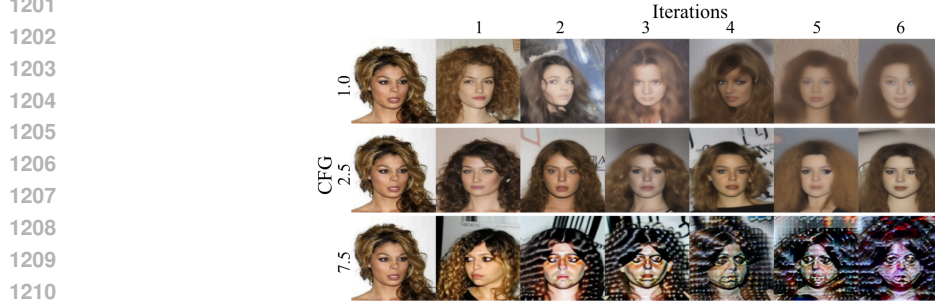
When Gaussian distribution with mean μ and variance σ^2 is truncated on both sides with r_1 and r_2 ratios, the mean and variance of truncated Gaussian distribution are expressed as:

1186
1187

$$\mu_{trun} = \mu - \frac{\varphi(\beta) - \varphi(\alpha)}{\Phi(\beta) - \Phi(\alpha)}\sigma, \tag{7}$$



1198 Figure 23: Chain of Diffusion on Pokemon dataset when training set is accumulated from previous
1199 iterations. All concepts from previous iterations are preserved for finetuning.
1200



1211 Figure 24: Chain of Diffusion on CelebA-1k dataset when training set is accumulated from previous
1212 iterations. All concepts from previous iterations are preserved for finetuning.
1213

1214
1215

$$\sigma_{trun}^2 = \left(1 - \frac{\beta\varphi(\beta) - \alpha\varphi(\alpha)}{\Phi(\beta) - \Phi(\alpha)} - \left(\frac{\varphi(\beta) - \varphi(\alpha)}{\Phi(\beta) - \Phi(\alpha)}\right)^2\right)\sigma^2, \quad (8)$$

1217 where φ and Φ are the probability density function (PDF) and cumulative distribution function (CDF)
1218 of the standard normal distribution, and $\alpha = \Phi^{-1}(r_1)$ and $\beta = \Phi^{-1}(1 - r_2)$. Accordingly, given
1219 $\mu = \mu_t$ 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

1220
1221
1222

$$\Delta\mu_t = h_t^2(\mu'_t - \mu_t) = \frac{\sigma_{G,t}^2}{\sigma_{P,t}^2} c_1 \sigma_{P,t} = c_1 \frac{\sigma_{P,t-1}^2}{\sigma_{P,t}^2} \sigma_{P,t} = c_1 c_2 \frac{\sigma_{P,t-1}^2}{\sigma_{P,t}^2} \sigma_{P,t}, \quad (9)$$

1223 where $c_1 = \left|\frac{\varphi(\beta) - \varphi(\alpha)}{\Phi(\beta) - \Phi(\alpha)}\right|$ and $c_2 = 1 - \frac{\beta\varphi(\beta) - \alpha\varphi(\alpha)}{\Phi(\beta) - \Phi(\alpha)} - \left(\frac{\varphi(\beta) - \varphi(\alpha)}{\Phi(\beta) - \Phi(\alpha)}\right)^2$ when $r_1 > r_2$. On the other
1224 hand, the mean phenotypes decreases when $r_1 < r_2$ as:

1225
1226
1227

$$\Delta\mu_t = -c_1 c_2 \frac{\sigma_{P,t-1}^2}{\sigma_{P,t}^2} \sigma_{P,t}. \quad (10)$$

1228 Furthermore, the phenotype variance converges over time as:

1229
1230

$$\sigma_{P,t}^2 = \sigma_{G,t}^2 + \sigma_E^2 = \sigma_{P,t-1}^2 + \sigma_E^2 = c_2 \sigma_{P,t-1}^2 + \sigma_E^2. \quad (11)$$

1231
1232
1233

$$\sigma_{P,t}^2 - \frac{\sigma_E^2}{1 - c_2} = c_2 \left(\sigma_{P,t-1}^2 - \frac{\sigma_E^2}{1 - c_2}\right) = \dots = c_2^t \left(\sigma_{P,0}^2 - \frac{\sigma_E^2}{1 - c_2}\right). \quad (12)$$

1234 As a result, the phenotype variance converges to $\frac{\sigma_E^2}{1 - c_2}$ for $0 < c_2 < 1$ (the variance of truncated
1235 distribution is smaller than the variance of the original distribution), and the mean asymptotically
1236 increases (decreases) by $\frac{c_1 c_2}{\sqrt{1 - c_2}} \sigma_E$ per iteration when $r_1 > r_2$ ($r_2 > r_1$). \square
1237

1238 D.2 SIMULATION SETUP

1239
1240 Our simulation aims to demonstrate that our experimental results can be modeled using quantitative
1241 trait modeling. We show that the radial sum of power spectra of images is one of the phenotypes
explained by our theoretical analysis.

1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295

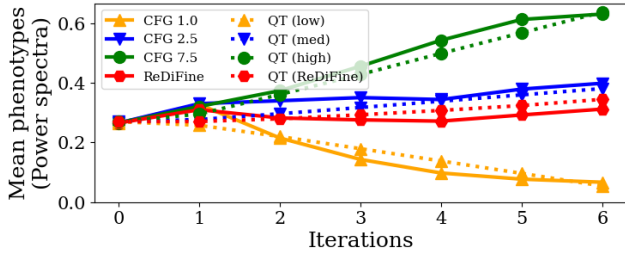


Figure 25: Results comparing the simulation for mutations and power spectra of images generated by ReDiFine (with CFG scale 7.5). Power spectra and simulation are plotted in solid and dotted lines, respectively. Our modifications to heritability and selection process successfully demonstrate changes occur to images by ReDiFine.

D.2.1 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 $i \in \{1, 2, \dots, N\}$, the 2D Discrete Fourier Transform (DFT) of an image I_i of size $M \times N$ is computed as:

$$F_i(u, v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I_i(x, y) e^{-2\pi j(\frac{ux}{M} + \frac{vy}{N})} \tag{13}$$

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

$$S_i = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} P_i(u, v) \cdot \mathbb{I}(r(u, v) > \tau) \tag{14}$$

where $\mathbb{I}(\cdot)$ 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:

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

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).

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:

$$h_t^2 = \frac{\sigma_{G,t}^2}{\sigma_{P,t}^2 + \sigma_M^2} = \frac{\sigma_{G,t}^2}{\sigma_{G,t}^2 + \sigma_E^2 + \sigma_M^2}, \tag{16}$$

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 25 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 26, 27, 28, and 29 show how robust ReDiFine is to different CFG scales. It successfully mitigates the high-frequency degradations for a wide range of CFG scales.

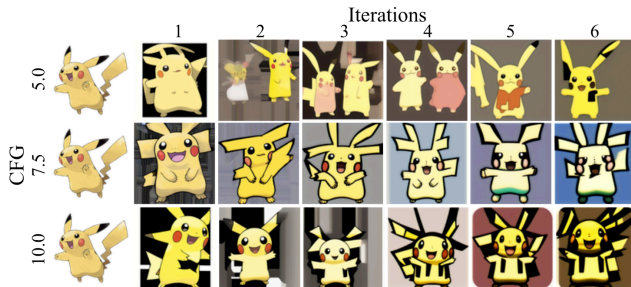


Figure 26: Chain of Diffusion of ReDiFine with different CFG scales on Pokemon dataset. ReDiFine successfully achieves robust image qualities for varying CFG scales.

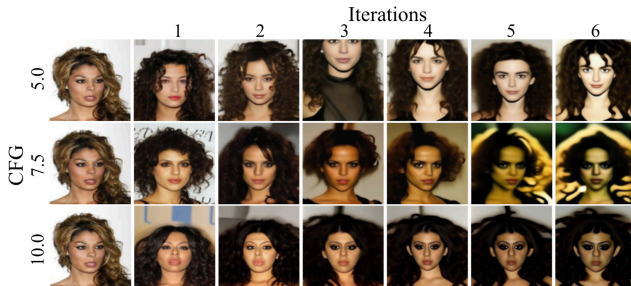


Figure 27: Chain of Diffusion of ReDiFine with different CFG scales on CelebA-1k dataset. ReDiFine successfully achieves robust image qualities for varying CFG scales.

E.2 MORE ITERATIONS

We conduct additional experiments to compare the baseline with the optimal CFG scale and ReDiFine over extended iterations. As shown in Figure 30, ReDiFine consistently generates images of similar

1350
1351
1352
1353
1354
1355
1356
1357
1358
1359

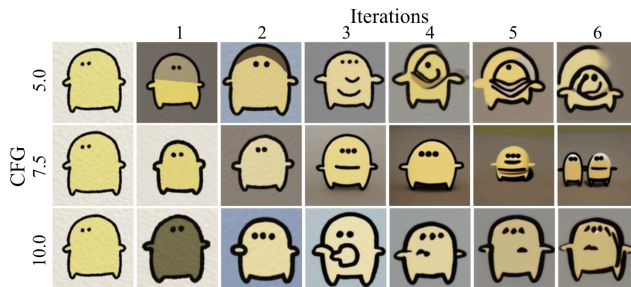


Figure 28: Chain of Diffusion of ReDiFine with different CFG scales on Kumapi dataset. ReDiFine successfully achieves robust image qualities for varying CFG scales.

1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372

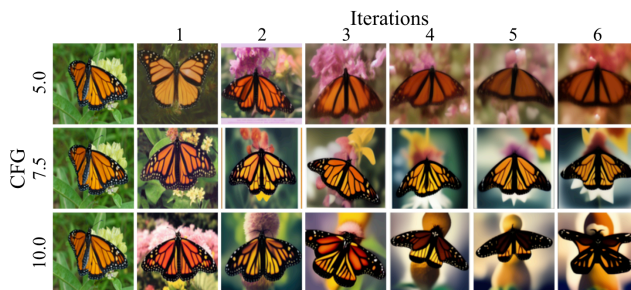


Figure 29: Chain of Diffusion of ReDiFine with different CFG scales on Butterfly dataset. ReDiFine successfully achieves robust image qualities for varying CFG scales.

1373
1374
1375
1376
1377

quality up to 12 iterations, whereas the optimally tuned CFG scale fails to sustain image quality. 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.

1381
1382
1383
1384
1385
1386
1387
1388

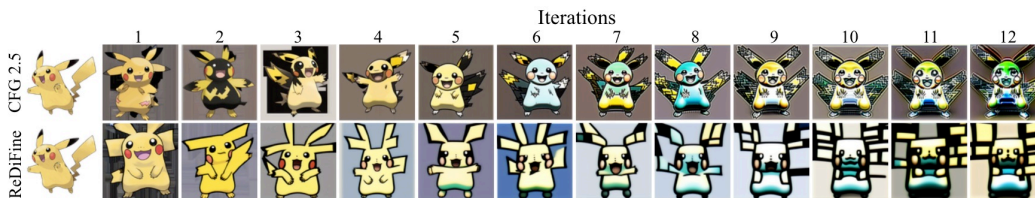


Figure 30: Chain of Diffusion of baseline with the optimal CFG scale and ReDiFine for more iterations. The generation quality of ReDiFine is preserved for additional iterations while optimally found CFG scale 2.5 fails to maintain the image qualities. Similar high-frequency degradation is observed.

1393
1394
1395

E.3 DATASET ACCUMULATION

1396
1397
1398
1399
1400
1401
1402
1403

We additionally conduct experiments when the training dataset is the accumulation of four datasets (Pokemon, CelebA-1k, Kumapi, and Butterfly) to investigate whether having a broader range of concepts impact model collapse. The accumulated dataset serves as the original training set, while the combined captions are used for image generation at each iteration. The results, presented in Figure 31, 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.

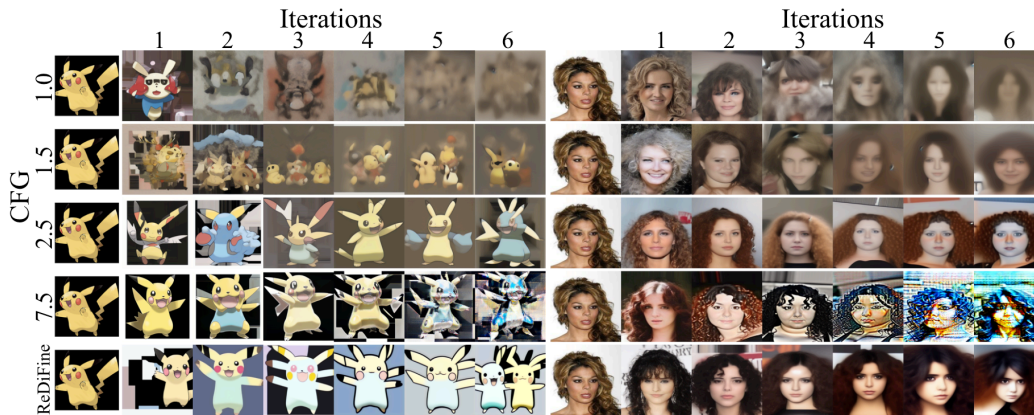


Figure 31: Chain of Diffusion with the accumulation of four datasets (Pokemon, CelebA-1k, Kumapi, and Butterfly) used for finetuning. Model collapse remains evident at low and high CFG scales. A single CFG scale (e.g., 1.5 or 2.5) fails to achieve optimal performance across both the Pokemon and CelebA-1k 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 32 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.

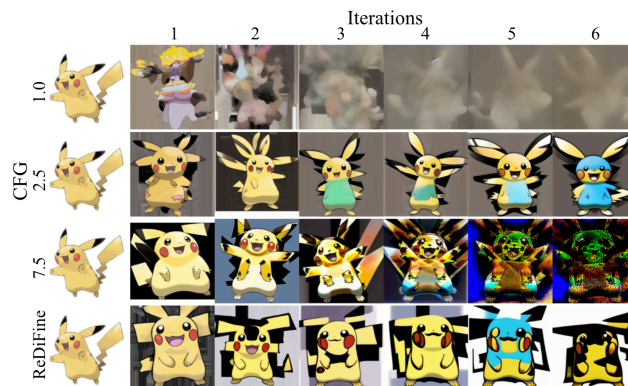


Figure 32: Chain of Diffusion where a single model is continually finetuned across multiple iterations. Model collapse is observed consistently at CFG scales of 1.0 and 7.5, while the baseline with a CFG scale of 2.5 and ReDiFine effectively mitigate model collapse. This demonstrates that model collapse is a universal phenomenon across different settings, and the effect of ReDiFine is effective robustly.

1458 E.5 QUANTITATIVE RESULTS

1459
1460 In addition to visual inspections for images generated using ReDiFine, we compare the quantitative
1461 results of ReDiFine to baselines with different CFG scales across FID, CLIP score, and Recall.
1462 Furthermore, we compute the average sample-wise feature distance (SFD) between a pair of images
1463 corresponding to the text prompt as the fidelity metric applicable for text-to-image generation,

$$1464 \text{SFD}_k = \frac{1}{N} \sum d(f(x_{k,i}), f(x_{0,i})) \quad (17)$$

1465
1466 to evaluate each iteration k . SFD overcomes the problem of FID being sensitive to the number of
1467 images to compare.

1468 DiNOv2 features are used to compute FID, Recall, and SFD. We follow Kynkäänniemi et al. (2019)
1469 to compute Recall and set the number of neighbors for computing Recall 5. The results, shown in
1470 Figure 33, demonstrate that ReDiFine achieves performance comparable to the optimal CFG scales
1471 (2.5 for Pokemon and Kumapi, 1.5 for CelebA-1k and Butterfly) across different datasets and metrics.
1472

1473 F ABLATION STUDY

1474 This section provides an ablation study to understand how condition drop finetuning and CFG
1475 scheduling contribute to the success of ReDiFine.
1476
1477

1478 F.1 CONDITION DROP FINETUNING

1479
1480 We conducted an ablation study to understand how the probability of dropping text embedding during
1481 finetuning affects the image quality in the Chain of Diffusion. We examine 0.1, 0.2, and 0.4 as
1482 Stable Diffusion is trained using 0.1 or 0.2. For both Pokemon and CelebA-1k datasets, a probability
1483 of 0.2 works the best, as shown in Figure 34 and Figure 35, respectively. Interestingly, condition
1484 drop finetuning helps to mitigate the color saturation problem, but its effect decreases with a higher
1485 probability. For both of these datasets, condition drop finetuning can mitigate image degradation to
1486 some degree, but still, there is a large quality degradation that needs to be improved.
1487

1488 F.2 CFG SCHEDULING

1489
1490 We also evaluated how different CFG scale decreasing strategies impact image degradation in the
1491 Chain of Diffusion. We experimented with two different exponential decay rates and compared
1492 them with a linear decreasing strategy. Figure 36 demonstrates that CFG scheduling is effective
1493 for Pokemon dataset, generating high-quality images comparable to those generated by ReDiFine.
1494 However, as shown in Figure 37, it fails to enhance image quality on CelebA-1k dataset. This
1495 highlights the necessity of condition drop finetuning for achieving universal improvements in the
1496 Chain of Diffusion across various datasets.
1497

1498 G ANALYSIS

1499
1500 In this section, we present a series of analyses of images generated through the Chain of Diffusion.
1501 Specifically, we examine the distribution of latent values and the differences between conditional and
1502 unconditional scores. Additionally, we analyze the power spectra of the images using 2D Fourier
1503 transforms and explore fingerprints through forensic analysis (Corvi et al., 2023a;b).

1504 G.1 LATENT ANALYSIS

1505
1506 Figure 38a illustrates how the distribution of latent values evolves across different iterations. The
1507 histograms show the final latent vectors before decoding into pixel space, comparing various CFG
1508 scales and ReDiFine. For a CFG of 1.0, the latent distribution rapidly converges into a Gaussian-
1509 like shape, with its variance shrinking over iterations. This behavior is consistent with previous
1510 work (Bertrand et al., 2023; Alemohammad et al., 2023; Dohmatob et al., 2024b), which theoretically
1511 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

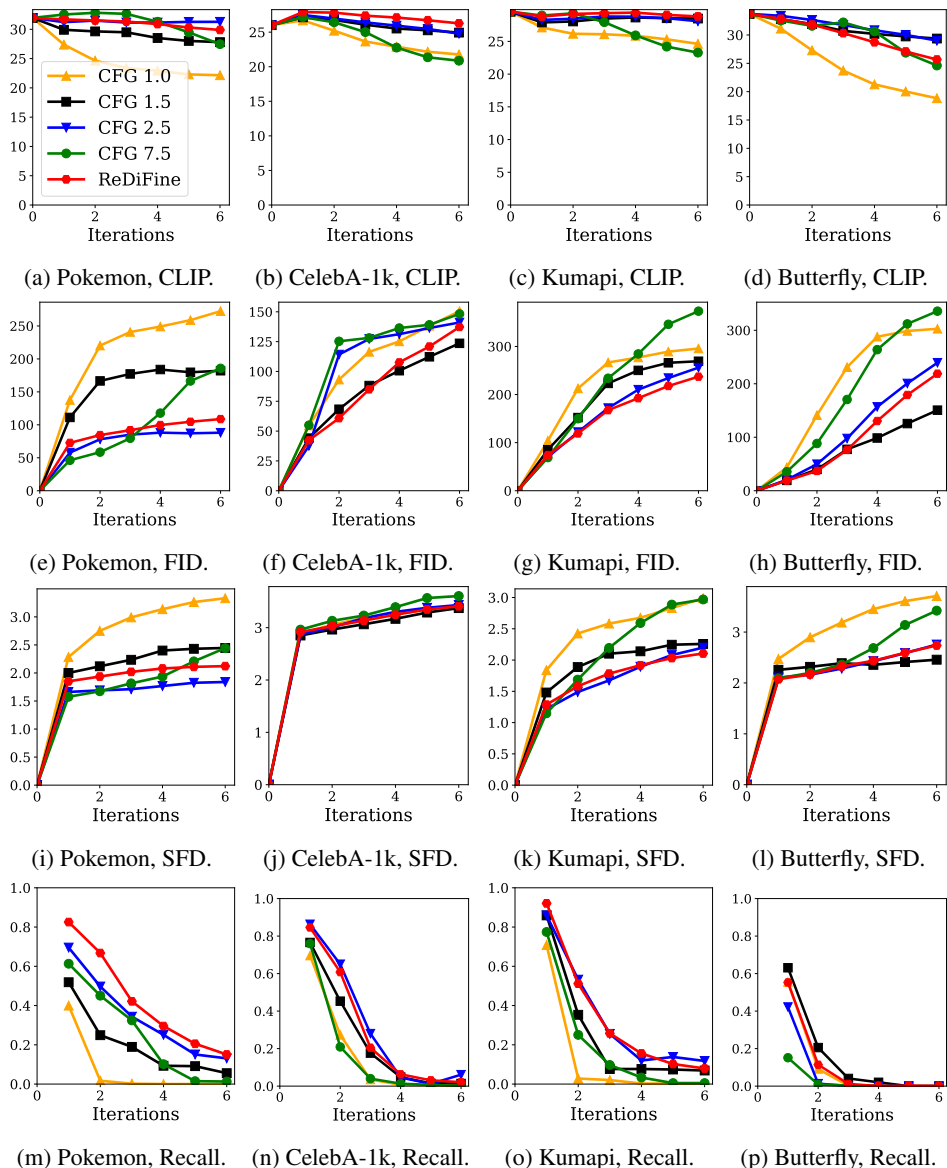
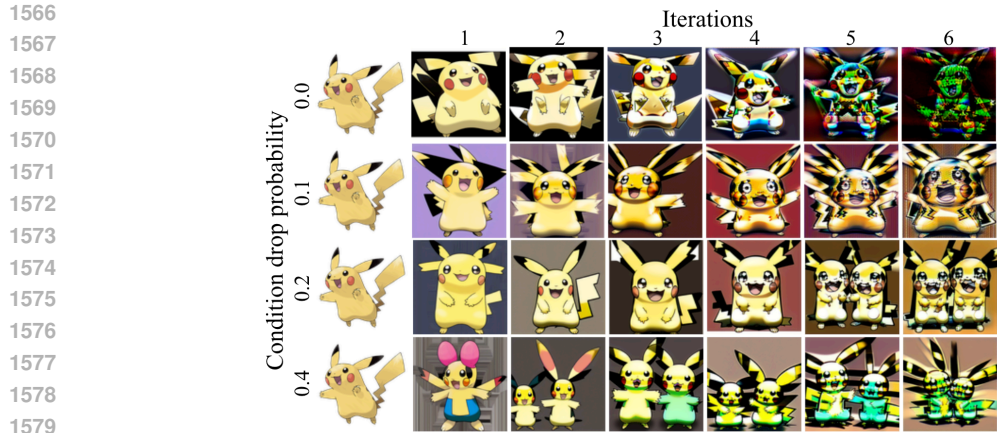


Figure 33: Quantitative results of ReDiFine and baselines (different CFG scales).

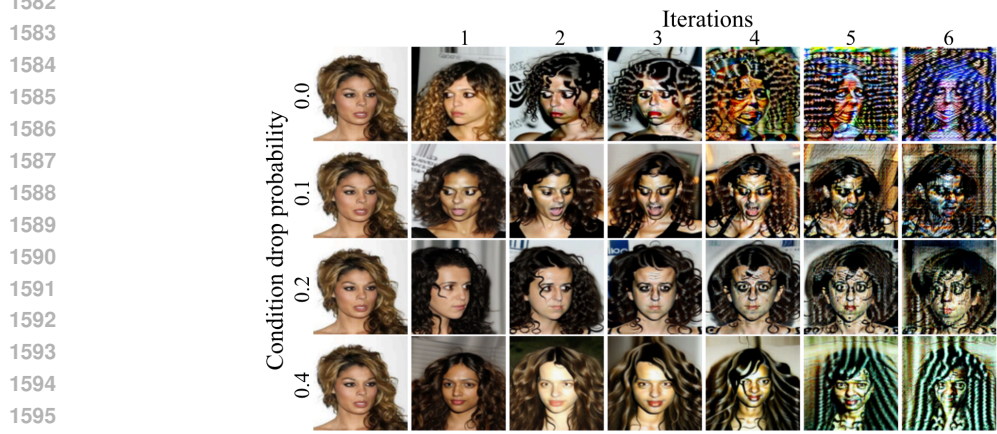
leads to blurrier, more homogeneous outputs in pixel space. Conversely, at a CFG scale of 7.5, the latent distribution develops longer tails and tends toward a more uniform spread across space. A CFG scale of 2.5, which demonstrates the best reusability among the three, better preserves the latent distribution 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

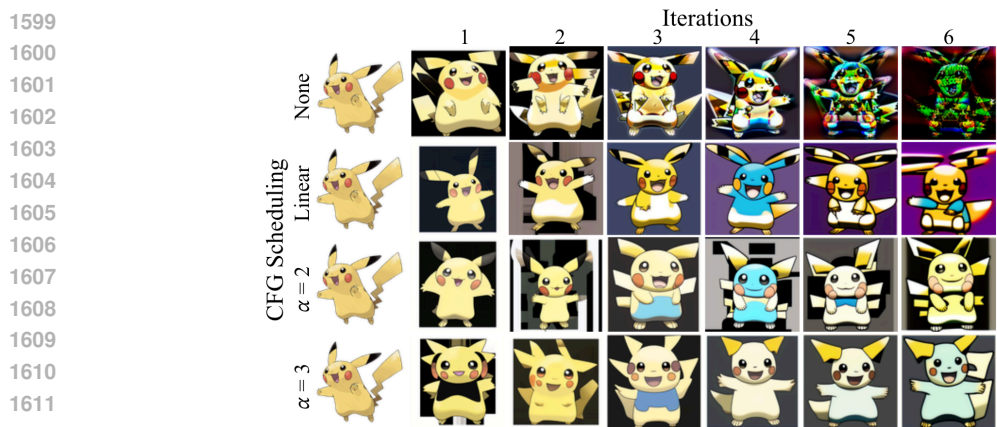
Next, we plot the evolution of the average norm of Diff ($= \text{Cond Score} - \text{Uncond Score}$) across diffusion steps for different iterations in Figure 38b. In the first iteration, the highest Diff value is observed for CFG 1.0, followed by CFG 2.5 and CFG 7.5. This behavior can be interpreted 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



1580 Figure 34: Chain of Diffusion with condition drop finetuning on Pokemon dataset.



1597 Figure 35: Chain of Diffusion with condition drop finetuning on CelebA-1k dataset.



1613 Figure 36: Chain of Diffusion with CFG scheduling on Pokemon dataset.

1614
1615
1616 entire diffusion steps, creating a significant gap compared to CFG 2.5 and 1.0. We conjecture that
1617 this accumulation of Diff is the responsible for the high-frequency degradation in images generated
1618 with CFG 7.5. In contrast, the Diff value for CFG 1.0 remains relatively stable or even decreases
1619 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

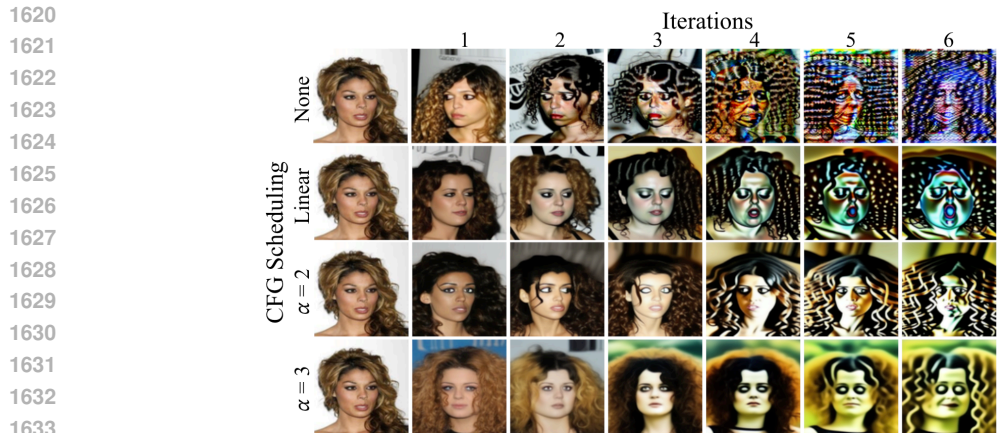
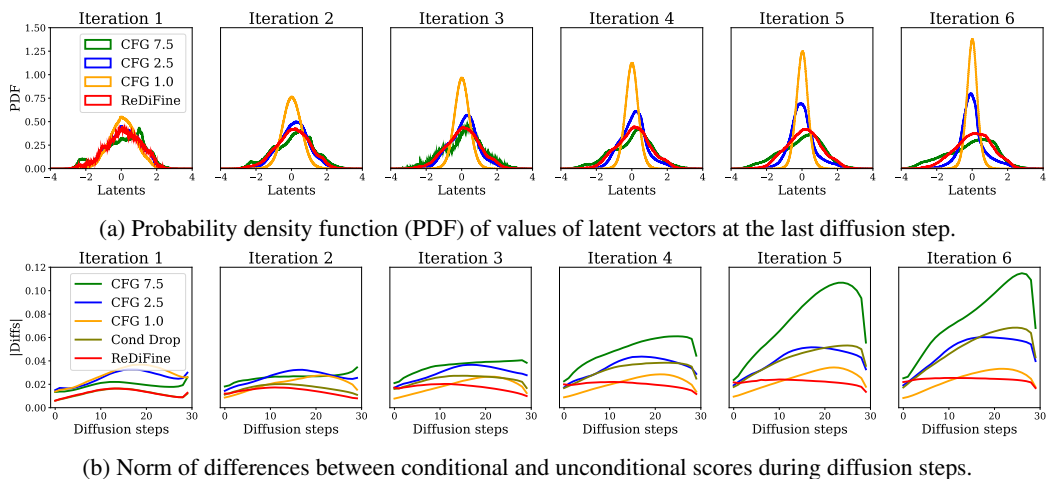


Figure 37: Chain of Diffusion with CFG scheduling on CelebA-1k dataset.



(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 38: 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.

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 39 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.

1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727

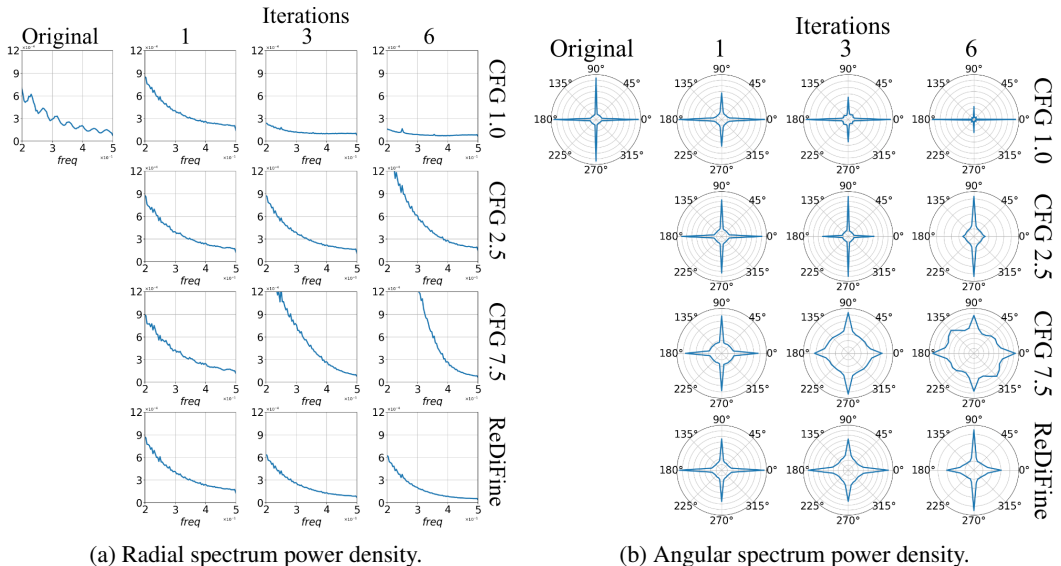


Figure 39: Power spectrum density of the original training set and synthetic sets for Pokemon dataset. **Images generated by ReDiFine maintain power density distribution during Chain of Diffusion while baselines fail. Even CFG scale 2.5 cannot maintain the distribution for the last iteration.** (a) Radial 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.

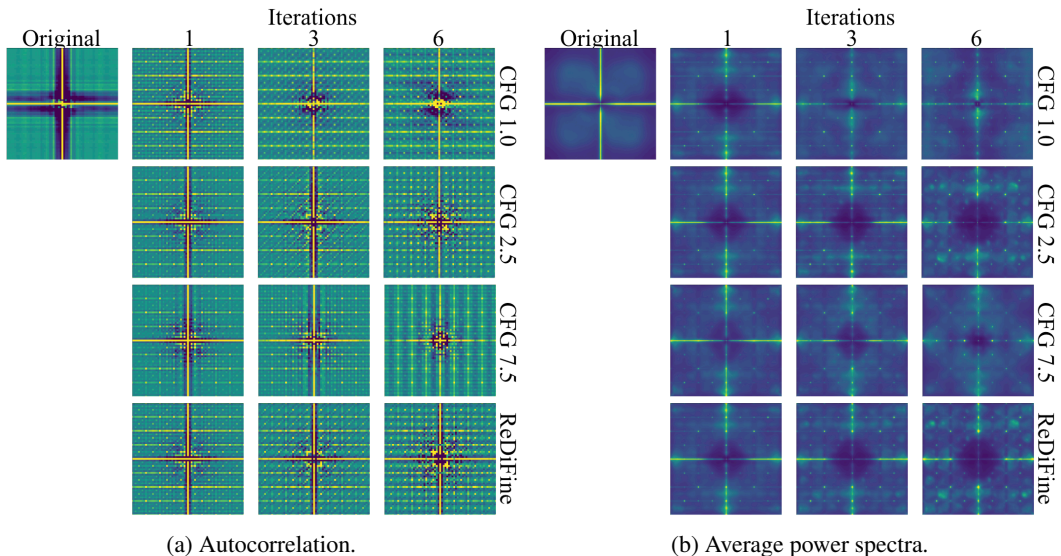


Figure 40: **The fingerprints of the original training set and synthetic sets show clear differences, and ReDiFine produces fingerprints similar to CFG scale 2.5.** (a) Autocorrelation of image fingerprints. Horizontal and vertical lines gradually disappear for CFG scales 1.0 and 7.5 while they are maintained 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.

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

1728 potential to be used for synthetic image detection. We analyze fingerprints of synthetic images for
1729 different CFG scales and iterations, and compare them to fingerprints of the original training set.
1730 Both autocorrelation and average power spectra show clear differences between the original training
1731 set and synthetic images, as shown in Figure 40. Moreover, how the fingerprints of synthetic images
1732 evolve throughout the Chain of Diffusion differ for ReDiFine and different CFG scales. Specifically,
1733 fingerprints of synthetic images from ReDiFine are similar to those of images from CFG scale 2.5,
1734 while other CFG scales (1.0 and 7.5) make fingerprints different from the first iteration as iterations
1735 proceed. Horizontal and vertical lines in autocorrelation gradually disappear and central regions in
1736 power spectra vary for further iterations. Also, the varying central regions in power spectra imply that
1737 low frequency features increase and decrease for CFG scale 1.0 and 7.5, respectively, aligning with
1738 visual inspections. Generating images with fingerprints similar to those of the original real images
1739 can be an interesting future direction to reduce the degradation in the Chain of Diffusion. Pokemon
1740 dataset is used for fingerprint analysis.

1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781