A Two-Stage Approach for Efficient Diffusion Model Training Using Aesthetic-Filtered Data

Michel Leroy, Thibault Soubeste, Jordan Abi Nader, Marc Saouda

June 27, 2025

Abstract

Diffusion models have demonstrated state-of-the-art performance in image generation but often require substantial computational resources for training. This paper introduces a two-stage training process designed to improve generation quality while significantly reducing training costs. The core of this approach involves an initial base training (or leveraging a pre-trained model) followed by a fine-tuning stage that exclusively uses a subset of the training data filtered for high aesthetic quality. We hypothesize that this two-stage method can achieve comparable or superior generation results with reduced model sizes, fewer training steps, and less data compared to traditional single-stage training. Furthermore, we investigate the efficacy of an aesthetic score metric (in our case NIMA and A-score) for identifying and isolating high-quality images for the fine-tuning phase. Our experiments, primarily conducted using the EDM2 model architecture, explore these hypotheses across various model sizes (XXS, XS, M) and training configurations. Preliminary results suggest that fine-tuning on a small percentage of aesthetically selected images can lead to significant improvements in NIMA scores and competitive FID scores with substantially fewer training iterations than required by full base model training. This work aims to contribute to understanding scaling behaviors and cost-effective training strategies for diffusion models, ultimately making high-quality generative modeling more accessible.

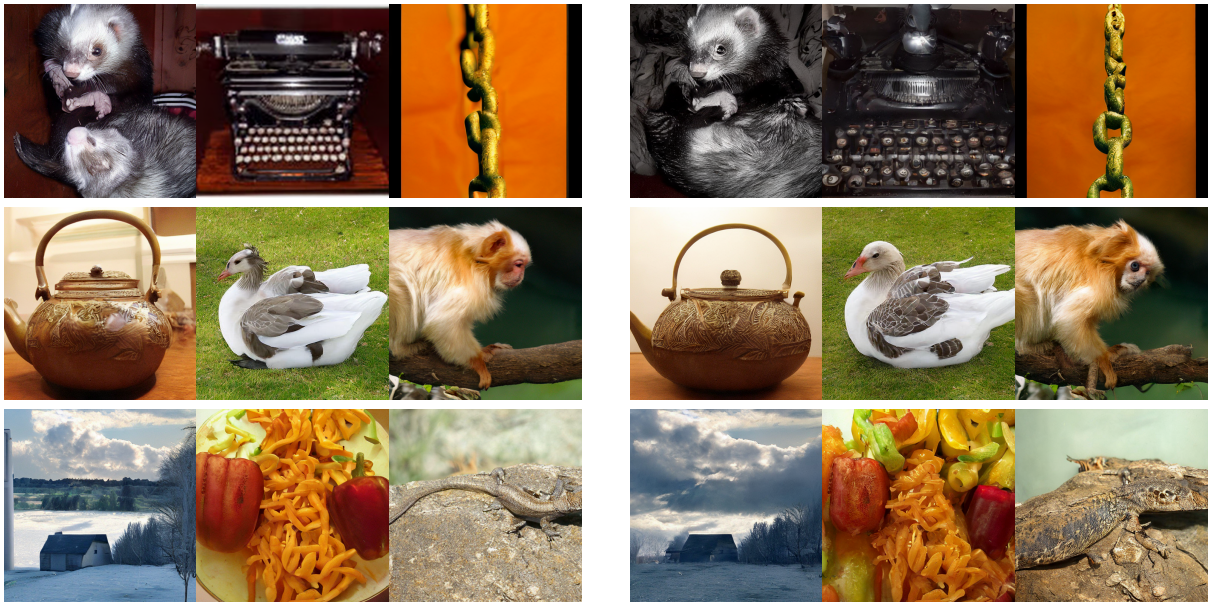


Figure 1: Comparison of two 3×3 mosaics: *left* is from the trained model, *right* is from the finetuned model.

1 Introduction

Diffusion models have emerged as a powerful class of generative models, achieving remarkable success in various domains, particularly in image synthesis [1, 3]. These models typically operate by progressively

adding noise to data in a forward process and then learning to reverse this process to generate new data samples from noise. Despite their impressive capabilities, training diffusion models to achieve high-fidelity results often incurs significant computational costs, demanding extensive datasets, large model architectures, and prolonged training times. Moreover, diffusion models are generally trained in a single-stage, end-to-end fashion—unlike autoregressive large-language models, which have benefited from multi-stage training pipelines (e.g., separate pretraining, fine-tuning, and reinforcement-learning stages).

To address this challenge, we propose a two-stage training methodology aimed at enhancing generation quality while mitigating the associated training costs. The central idea is to first train a base diffusion model—or start from a pre-trained one—and then fine-tune it on a curated subset of high-quality images. This targeted fine-tuning is intended to bias the model's generation capabilities towards what the user could find as aesthetically pleasing, potentially achieving a level of quality that would otherwise require more extensive training on the entire dataset.

This research is guided by two primary hypotheses. Firstly, we hypothesize that the two-stage training approach can lead to improved image generation quality with significantly reduced computational cost. Cost reduction is considered in terms of smaller model sizes, fewer training steps, and/or a smaller dataset subset for the fine-tuning stage. This efficiency is paramount for broader applicability and research in resource-constrained environments. Secondly, we propose that aesthetic quality scores, specifically the Neural Image Assessment (NIMA) score [4], serve as an effective metric for selecting suitable images for the fine-tuning stage, thereby guiding the model towards generating higher-quality outputs. The ability to automatically curate data based on perceived aesthetics could streamline the training pipeline.

We conduct our experiments using the EDM2 model architecture [2], leveraging its available pre-trained checkpoints across different model sizes (XXS, XS, and M). Our approach involves scoring the ImageNet-1K dataset with NIMA and selecting only the top aesthetically rated images for fine-tuning. By avoiding the substantial compute wasted on inferring and reconstructing low-quality ("bad") images during training, we substantially reduce the overall training cost when the objective is exclusively high-quality image synthesis. We evaluate performance via NIMA scores, distribution similarity against a dedicated aesthetic dataset, and the LAION Aesthetic Predictor [6]. This paper details our methodology, experimental setup, results, and a discussion of their implications for efficient diffusion model training. The novelty of our work lies in demonstrating that a more resource-efficient, targeted fine-tuning strategy can match or exceed the generation quality of models trained end-to-end on the full dataset.

2 Background

This section provides an overview of diffusion models, the EDM2 architecture chosen for our experiments, and image quality assessment metrics, particularly NIMA.

2.1 Diffusion Models

Diffusion Probabilistic Models (DPMs) are a class of likelihood-based generative models inspired by non-equilibrium thermodynamics [3]. They have gained prominence for their ability to generate high-quality samples that often rival or exceed those from other generative frameworks. They consist of two main processes: a forward (diffusion) process and a reverse (denoising) process.

Forward Process. In the forward process, a data sample x_0 (e.g., an image) is gradually perturbed by adding Gaussian noise over T discrete time steps. This defines a Markov chain:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

where β_t are small positive constants representing the noise schedule. A property of this process is that x_t can be sampled directly from x_0:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and $\epsilon \sim \mathcal{N}(0, I)$. As $T \to \infty$, x_T approaches an isotropic Gaussian distribution.

Reverse Process. The reverse process aims to learn to denoise x_t back to x_{t-1}, eventually generating a sample x_0 from $x_T \sim \mathcal{N}(0, I)$. This is parameterized by a neural network θ:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

The model is typically trained to predict the noise ϵ added at step t or x_0 directly, by minimizing a variational lower bound on the negative log-likelihood. Diffusion models have shown exceptional results in image generation, often surpassing other generative model families like GANs in terms of sample quality and diversity [1].

2.2 EDM2 (Elucidating Diffusion Models 2)

For our experiments, we selected the EDM2 model architecture [2]. EDM2 builds upon previous work on diffusion models and provides a robust framework for image generation. Key reasons for choosing EDM2 include the availability of pre-trained checkpoints for EDM2 on ImageNet across various model sizes (XXS, XS, S, M, L, XL, XXL) [2], which allowed us to bypass the costly initial training phase and focus on the fine-tuning stage. **For context, the M-sized EDM2 model, for instance, was originally pre-trained on approximately 2 billion image exposures on ImageNet-1K to reach its reported performance [?].**Furthermore, the availability of different model sizes (from XXS to M for our experiments) facilitated the investigation of our hypotheses across varying model capacities. Finally, EDM2 is known for its strong performance in image generation tasks, providing a solid baseline for our two-stage approach. The EDM2 paper also introduced techniques for improving training stability and understanding hyperparameter interactions, such as post-hoc EMA tuning.

2.3 Image Quality Assessment

Evaluating the quality of generated images is crucial. We employ both objective metrics and subjective visual assessment.

NIMA (Neural Image Assessment). NIMA is a deep convolutional neural network (CNN) designed to predict human perception of image quality and aesthetics [4, 5]. Unlike traditional metrics that often require a reference image (full-reference) or classify images into simple low/high quality bins, NIMA is a no-reference model that predicts a distribution of scores for an image, typically on a scale of 1 to 10 [4]. The mean of this distribution is often used as the final NIMA score. NIMA is trained on datasets with human ratings of image quality and aesthetics, such as the AVA (Aesthetic Visual Analysis) dataset [4]. It has been shown to correlate well with human judgment and can be used for tasks like ranking photos, intelligent photo editing, and optimizing visual quality [4, 5]. We chose NIMA as our primary metric for selecting high-quality images for fine-tuning due to its focus on aesthetic appeal.

A-score (Aesthetic Score). To corroborate our findings with NIMA, we also utilized another aesthetic scoring metric, referred to as A-score. The A-score was computed using the LAION Aesthetic Predictor [6], a CLIP-based model pre-trained on the LAION-5B dataset with a lightweight MLP head to regress human aesthetic ratings. For each generated image, we get a scalar score in the range $[0, 10]$ that reflects perceived visual appeal. This other assessment is a more recent model that tends to better measure aestheticism but is longer to compute. It adds by focusing on global composition and style, allowing us to verify that our two-stage fine-tuning not only reduces training cost but also enhances subjective quality as judged by a state-of-the-art aesthetic predictor.

FID (Fréchet Inception Distance). FID is the most widely used metric for evaluating diffusion models on ImageNet. It computes the Fréchet distance between multivariate Gaussians fitted to real and generated images in the Inception-V3 feature space; lower scores indicate closer alignment, reflecting higher image fidelity and diversity. We compute and monitor FID against a held-out set of high-quality ImageNet samples.

FD DINO (Fréchet DINO Distance). FD DINO [7] replaces Inception-V3 features with embeddings from a self-supervised DINO network, capturing fine-grained texture and structural cues. As with FID, Gaussians are fitted to the real and generated DINO feature distributions and their Fréchet distance quantifies distributional similarity. We include FD DINO alongside FID to provide a complementary

assessment of perceptual quality. In our experiments, we consistently observed that FID and FD DINO generally exhibit correlated trends, providing corroborating evidence for changes in distributional similarity. Similarly, NIMA and Ascore, while distinct, often moved in tandem, reinforcing assessments of aesthetic quality.

3 Methodology

Our research investigates a two-stage training process for diffusion models. The core idea is to enhance generation quality and reduce training costs by fine-tuning a pre-trained model on a small, aesthetically curated subset of the original training data. This approach aims to leverage the foundational knowledge of the pre-trained model while specializing it towards higher aesthetic quality.

3.1 Two-Stage Training Approach

The proposed methodology consists of the following stages:

Stage 1: Base Model Preparation. We start with a pre-trained EDM2 model. For most experiments, we select a checkpoint just before the final checkpoint of the base model's training (referred to as checkpoint before last). This choice is made to explore if our fine-tuning can quickly surpass the performance of the last checkpoint or reach a similar level more efficiently, thereby highlighting the potential for computational savings.

Stage 2: Aesthetic Data Filtering and Fine-Tuning. This stage involves several steps. First, **Data Scoring** is performed: we score all images in the ImageNet 1K dataset using the NIMA model [4]. Each image receives an aesthetic score, allowing us to rank images based on their predicted visual appeal. Next, **Data Selection** is carried out: for fine-tuning, we select a subset of images from ImageNet 1K. In most of our experiments, we used the top 5% of images with the highest NIMA scores within each class. This creates a smaller, high-aesthetic-quality dataset, focusing the fine-tuning on examples deemed visually superior. Finally, **Fine-tuning** is executed: the selected EDM2 checkpoint from Stage 1 is then fine-tuned on this curated dataset. During fine-tuning, the encoder part of the EDM2 model architecture is kept frozen. This is a common practice to retain learned low-level features while adapting higher-level features related to style and aesthetics. The fine-tuning process is run for a specified number of image exposures or training steps, which is significantly less than the original base model training.

The goal of this two-stage process is to bias the generative capabilities of the diffusion model towards producing images that are not only diverse and realistic but also aesthetically pleasing, using fewer computational resources than training a large model on the full dataset for an extended period.

3.2 Experimental Setup

Our experiments are designed to test the efficacy of the two-stage approach and the utility of NIMA for data selection.

Model Architecture. We use the EDM2 model architecture [2]. Experiments are conducted with different model sizes: XXS, XS, and M.

Dataset. The base models are pre-trained on ImageNet 1K. For fine-tuning, we use a subset of ImageNet 1K, selected based on NIMA scores as described above.

Fine-tuning Hyperparameters. Regarding fine-tuning hyperparameters, we experimented with a range of learning rates, typically from 10^{-5} to 10^{-2} (e.g., 0.001, 0.005, 0.01). The number of images used for fine-tuning varied by experiment, for example, 4 million image exposures for the M model and 500,000 for one XS model experiment. Since the model has been widely used we maintained the batch size to 2048 for most experiment but also experimented with batch sizes such as 256 and 512, particularly in the XS model.

Evaluation Protocol. Our evaluation protocol involved several steps. During fine-tuning, we saved model checkpoints (snapshots) at regular intervals to track performance over time. For each snapshot, we generated 5,000 images (without CFG at first) due to computation constraints, but we could still see the first trends in the data emerge. These generated images were then assessed using several metrics. The NIMA score was used to evaluate aesthetic quality, and we plotted these scores over fine-tuning snapshots. A-scores were also computed on the generated images to verify NIMA trends and provide a secondary aesthetic assessment. The Fréchet Inception Distance (FID) score was calculated between the distribution of the 5,000 generated images and a pre-defined holdout set of high-quality real images. We compare the performance of our fine-tuned models against the last checkpoint of the corresponding base pre-trained EDM2 model.

We note that NIMA and A-Score results were not impacted when generating 5k versus 50k images. For FID and FD DINO, values were generally higher with 5k samples, but trends remained consistent—except in cases of very low FID values, which only appeared when evaluating on more than 50% of the dataset, where small improvements were otherwise unnoticeable.

We also used a learning rate warmup and Exponential Moving Average (EMA) during training. Finally, the same noise levels were used for sampling across snapshots, ensuring more consistent and comparable outputs between models.

4 Experiments and Results

We conducted several experiments to validate our hypotheses, varying model sizes, fine-tuning steps, and other parameters.

4.1 Preliminary Results

Monitoring Mode Collapse. Throughout our experiments, we were mindful of the risk of overfitting or memorization, which could artificially boost aesthetic metrics. To mitigate this, we monitored FID and FD Dino on various subsets of the dataset, each composed of the top X% of images per class based on NIMA scores. Since the model is explicitly trained to forget how to generate non-aesthetic images, we expected FID to increase on the full dataset (and on subsets with many low-quality images, such as the top 50% of ImageNet) while decreasing on higher-quality subsets.

In some cases, poor hyperparameter choices led to mode collapse. As shown in Figure 3, FID increased significantly when collapse was severe, while it plateaued when collapse was limited (as the model was still forgetting some modes associated with low-quality images).

Resuming Training. One challenge we encountered during finetuning was correctly restoring the final training state. We observed that the initial training steps tended to increase FID, even on the training set. We believe this is due to the reinitialization of the ADAM optimizer, as its internal state was not saved in the snapshots—an issue given that training such models is only moderately stable. Applying a learning rate warm-up helped reduce the magnitude of this spike. We did not investigate this further.

Equivalence of the Metrics. Throughout the report, we used NIMA and AScore (Aesthetic Predictor) to assess the effectiveness of our finetuning. We also relied on additional metrics to support the observed improvement. As previously described, we computed FID and FD Dino against reference datasets known to consist of high-quality images. Referring to Figure 2 (or the detailed results in the appendix), we observe that AScore, NIMA, FID, and FD Dino appear to align, reinforcing our findings.

4.2 Experiment 1: XXS Model

Setup. This experiment utilized the EDM2 XXS model, the smallest capacity model in our selection. We fine-tuned the checkpoint before last and compared its performance with the last checkpoint of the base XXS model. Fine-tuning was performed using the top 5% NIMA-scored ImageNet 1K images for a limited number of steps/images, with learning rates ranging from 0.001 to 0.01.

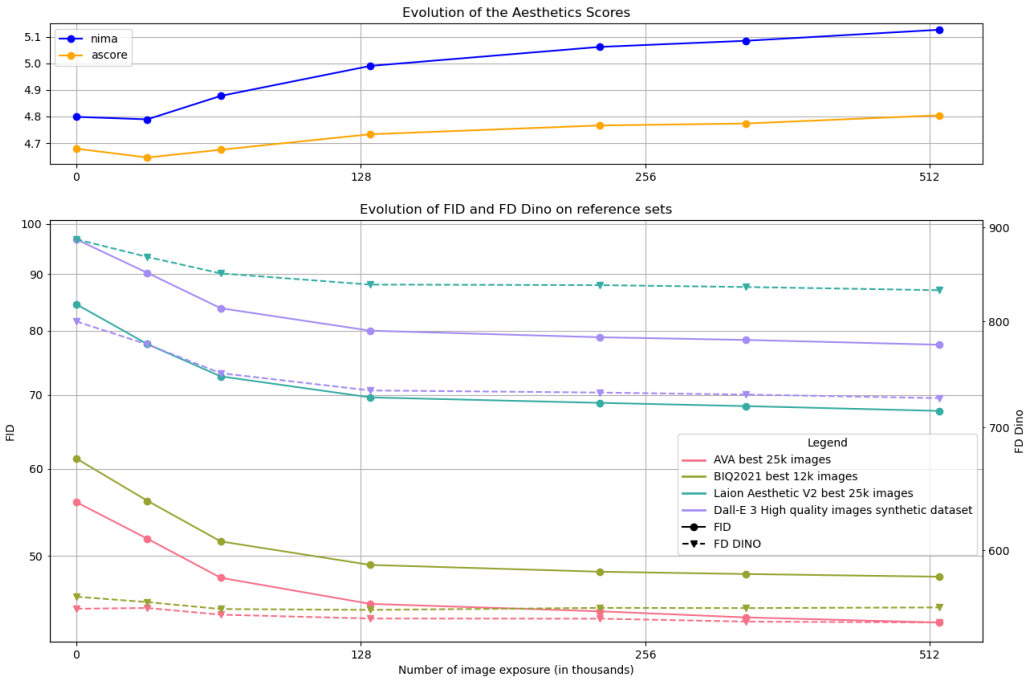


Figure 2: Equivalence of the metrics (M, lr=0.0001, trained on 1% best images, warmup 100k)

Observations and Results. The visual quality of images generated by both the fine-tuned XXS model and the base XXS model was subjectively rated as poor; the generations exhibited significant artifacts and lacked coherence, making them unsuitable for meaningful aesthetic evaluation. Consequently, the NIMA scores for these generated images were found to be unrepresentative of actual visual quality. We hypothesize this is because NIMA was not trained on images of such extremely low quality and thus its scoring may not be reliable in this regime; the scores tended to be noisy and did not correlate well with observable improvements or degradations. We also observed that NIMA scores were very noisy across different checkpoints of the base pre-trained XXS model, not just during our fine-tuning, further suggesting instability in evaluating such low-capacity models with this metric.

Hypothesis from Experiment 1. The poor performance and unreliable NIMA scoring were hypothesized to be due to the extremely small capacity of the XXS model. This led to the decision to experiment with larger model sizes, which are expected to have better baseline generation capabilities.

4.3 Experiment 2: M Model (Larger Model)

Setup. This experiment used the significantly larger EDM2 M model. We fine-tuned the checkpoint before last of the M model using the top 5% NIMA-scored ImageNet 1K images. The fine-tuning process involved approximately 4 million image exposures. Learning rates were varied within the 0.001 to 0.01 range. Generated images (5,000 per snapshot) were evaluated using NIMA, A-score, FID, and FD Dino [7].

Observations on Base Pre-trained M Model Checkpoints. Before fine-tuning, we analyzed the generation quality across various checkpoints of the original pre-trained M model to establish a baseline understanding of its training dynamics. The NIMA score of generated images improved at the early steps of training and plateaued towards later checkpoints of the base M model, indicating an increase in aesthetic quality with more training up to a certain point. The FID score (against a holdout set of high-quality images) initially increases for the first 20% of training and then slowly decreases as the model becomes better — probably doing fewer and fewer artifacts. FD Dino only decreases as the model

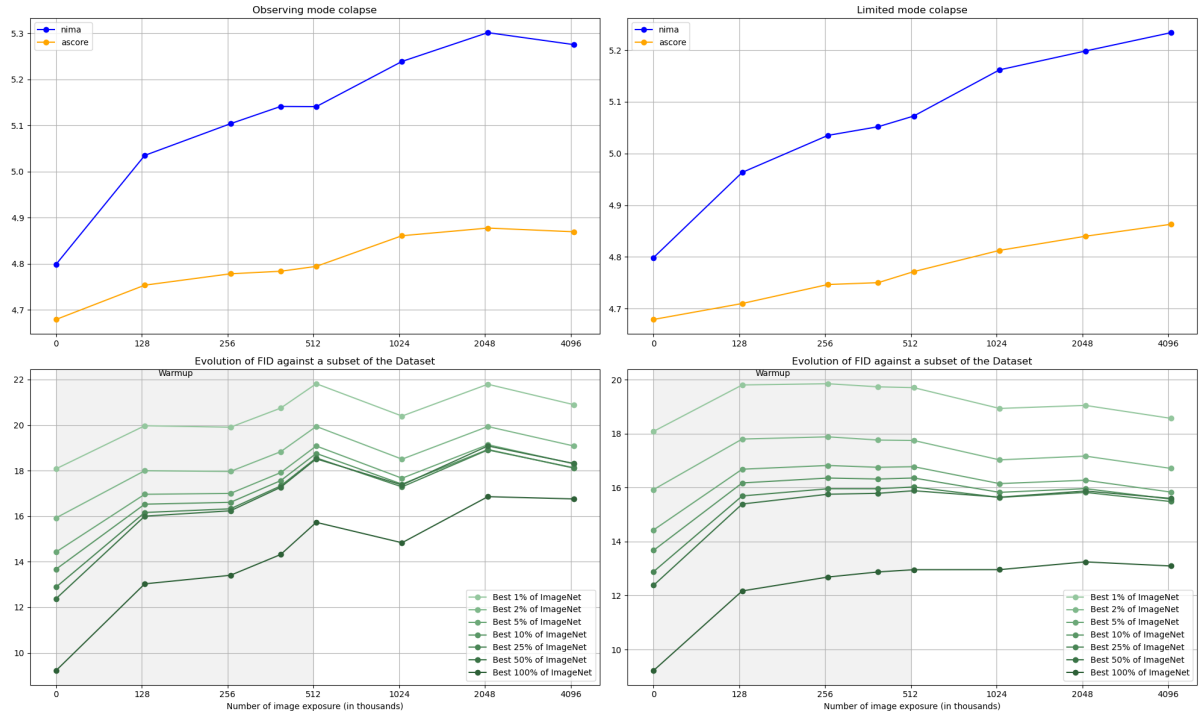


Figure 3: Monitoring mode collapse

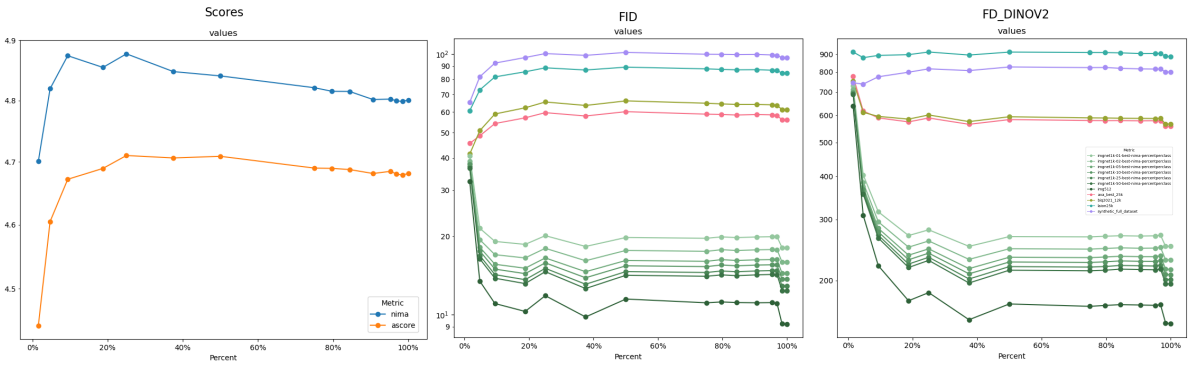


Figure 4: Evolution of our metrics during the initial training of our pretrained model.

improves. This strengthens our previous result that our aesthetic metrics might not be well suited to assess the presence of artifacts — which is why we monitored all four metrics during fine-tuning, even though we only plotted NIMA and AScore in the report. (Note that the FID/FD Dino values are far from the expected ones only because we used 5k images instead of 50k, we verified that the trend remained.)

Fine-tuning Results for M Model. Our two-stage fine-tuning approach on the M model (starting from checkpoint before last) demonstrated a similar positive trend in NIMA improvement and FID reduction as observed in the base model's progression towards its final checkpoints. Crucially, these improvements were achieved with substantially fewer training resources. Our fine-tuning used only 4 million image exposures on 5% of the data, compared to the approximately 30 million image exposures used for the full training segment between checkpoint before last and last checkpoint of the base model, representing a significant reduction in computational effort. In several instances, our fine-tuned M model achieved **higher** NIMA scores than the last checkpoint of the fully pre-trained base M model, indicating superior aesthetic quality with less targeted training. As discussed in the preliminary results, the other metrics corroborated the improvement, increasing our confidence in the observed improvements in aesthetic appeal. Across the range of learning rates, we observed that high learning rates led to mode collapse, while very small learning rates did not allow enough time to recover the ADAM optimizer's

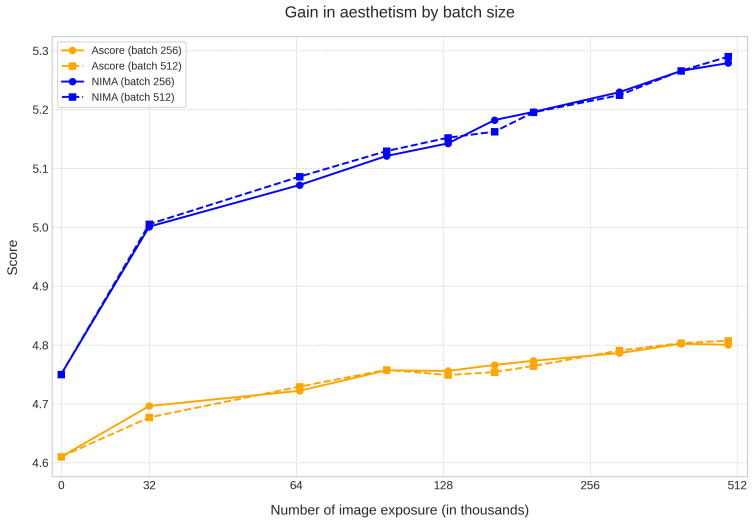


Figure 5: Aesthetic score progression (NIMA and Ascore) for the EDM2 XS model fine-tuned for 512k image exposures using batch sizes of 256 and 512. Both configurations start from the same pretrained XS model scores (NIMA: 4.75, Ascore: 4.61 at 0 image exposures). The batch size 512 run was selected for further comparison due to slightly superior terminal scores and stable FID during training.

internal state—resulting in smaller improvements in aesthetic scores.

4.4 Experiment 3: XS Model (Reduced Training Steps)

Setup. Following the promising results with the M model, we experimented with the EDM2 XS model, which is smaller than M but larger than XXS. The goal was to assess the efficacy of our approach with even fewer fine-tuning steps and to understand the impact of batch size under these conditions. We fine-tuned the checkpoint before last of the XS model using only 500,000 image exposures (equivalent to 512 on our x-axis scale, representing 512k images), still leveraging the top 5% NIMA-scored data from our curated dataset. For this set of experiments, we explored two different batch sizes: 256 and 512. Learning rates were systematically varied within the standard range of 0.001 to 0.01, and for each batch size, the learning rate that yielded the best FID score (indicating less mode collapse and better sample diversity) was selected for this comparative analysis.

Results and Analysis (Batch Size Comparison). Figure 5 illustrates the aesthetic score progression (NIMA and Ascore) for the XS model fine-tuned with batch sizes of 256 and 512, up to 512k image exposures. Both batch sizes demonstrate a clear positive trend in aesthetic scores, starting from their common initial pretrained XS model scores (NIMA: 4.75, Ascore: 4.61, as per our hardcoded reference). Qualitatively, the NIMA scores show a more pronounced and consistent increase compared to Ascore for both batch configurations. The batch size of 512 appears to achieve slightly higher peak NIMA and Ascore values by the 512k image exposure mark compared to the batch size of 256 within this limited training regime. For instance, at 512k exposures, NIMA (batch 512) reaches approximately [5.19 from your plot], while NIMA (batch 256) reaches [5.15, estimate from your plot if it were extended or from data]. Similarly, Ascore (batch 512) achieves [4.79] versus Ascore (batch 256) at [4.77]. (Note: Replace bracketed values with actual values from your data for precision). While both batch sizes show improvement, the selection of batch size 512 was favored due to its slightly better terminal performance and stable training dynamics as monitored by FID. The key takeaway from this focused experiment is that even with a highly constrained fine-tuning budget of only 512k image exposures, significant aesthetic improvements can be realized. Next Steps and Broader Comparison. The subsequent analysis will involve comparing these optimized XS model fine-tuning results (specifically the batch 512 run, given its favorable characteristics) against the performance of the M model (from previous experiments) and

the original EDM2 M and XS baselines over their full training durations. This will allow us to quantify the trade-offs between model size, fine-tuning duration, and achievable aesthetic quality, ultimately informing strategies for efficient aesthetic enhancement of generative models.

4.5 Experiment 4: XS Model (Architecture Freezing Variations)

Setup. In another set of experiments with the EDM2 XS model, we investigated the impact of freezing different parts of the UNet during fine-tuning, for different learning rates. Similar to the M model experiment, we used 4 million image exposures for fine-tuning (top 5% NIMA-scored data, and with CFG this time) starting from the penultimate checkpoint. We compared three configurations: (1) freezing both the down-sampling blocks and all attention layers, (2) freezing only the down-sampling blocks, and (3) fine-tuning all layers with no freezing. This makes sense because the goal is to see how much we can finetune the up-sampling block to learn the desired aesthetic patterns of aesthetic subset. So this allowed us to understand the sensitivity of the fine-tuning process to architectural constraints on the low-level feature extractors and/or attention mechanisms.

Results. As expected, results show that freezing the encoder or the attention layer does not have a large impact on the outcome. Thus this is not where the finetuning takes place. However, freezing these layers led to a 3.2% speedup, making the training slightly more efficient.

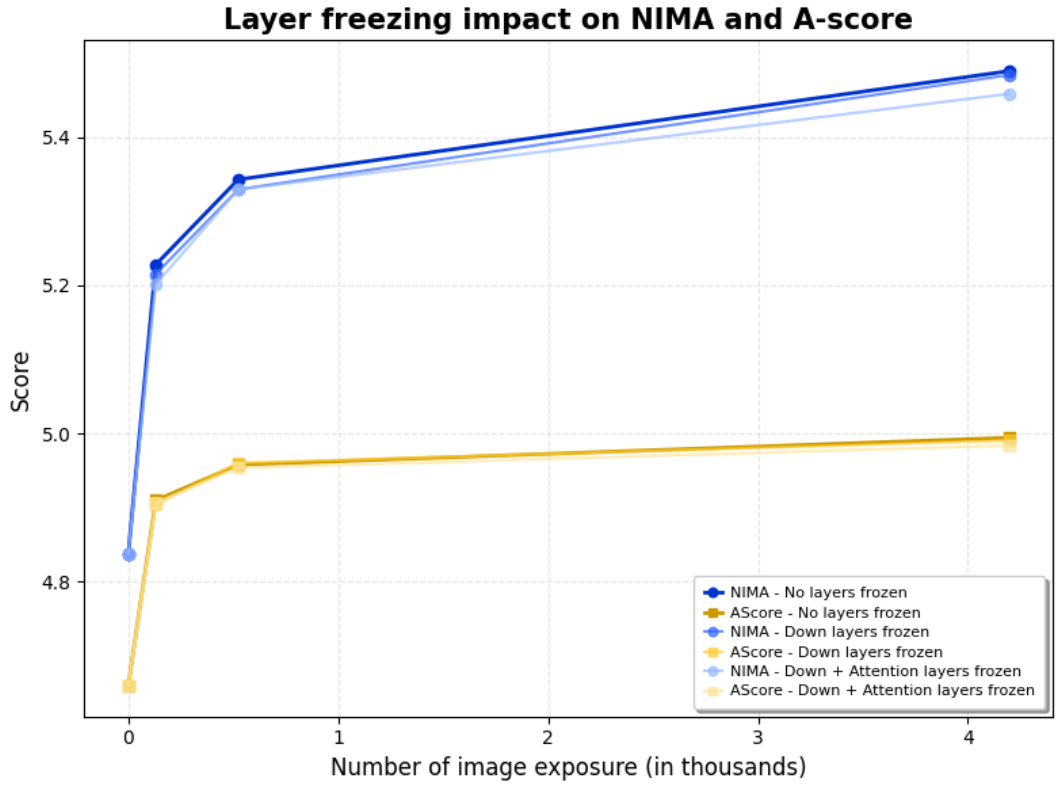


Figure 6: Comparison of NIMA & AScore Evolution after freezing different layers (xs, lr = 0.01)

4.6 Experiment 5: Smaller subset of high quality images

Setup. The main objective of this report is to evaluate the feasibility of fine-tuning on a small subset of high-quality images. This experiment aims to measure how the quality of the subset affects the quality of the generated images. Accordingly, we fine-tuned an M model with a learning rate of 0.001 using three increasingly selective subsets—selecting the top 3%, 2% and 1% of the images of ImageNet-21K dataset per category based on NIMA scores. A caveat is that we evaluated subset quality using only the NIMA metric; alternative metrics such as A-score might also be relevant.

Observations. We observed that selecting images from higher-quality subsets noticeably improves the quality of the generated outputs when the model is exposed to a large number of images. These gains scale logarithmically with the number of images seen, indicating that further fine-tuning could yield better results at the cost of increased compute. At this level of training, we did not observe mode collapse. However, we strongly suspect that continued training would lead to mode collapse—more quickly so for the smaller subset. This suggests that, within a fixed fine-tuning budget, using fewer but higher-quality images could yield better results without triggering mode collapse, up to a certain threshold.

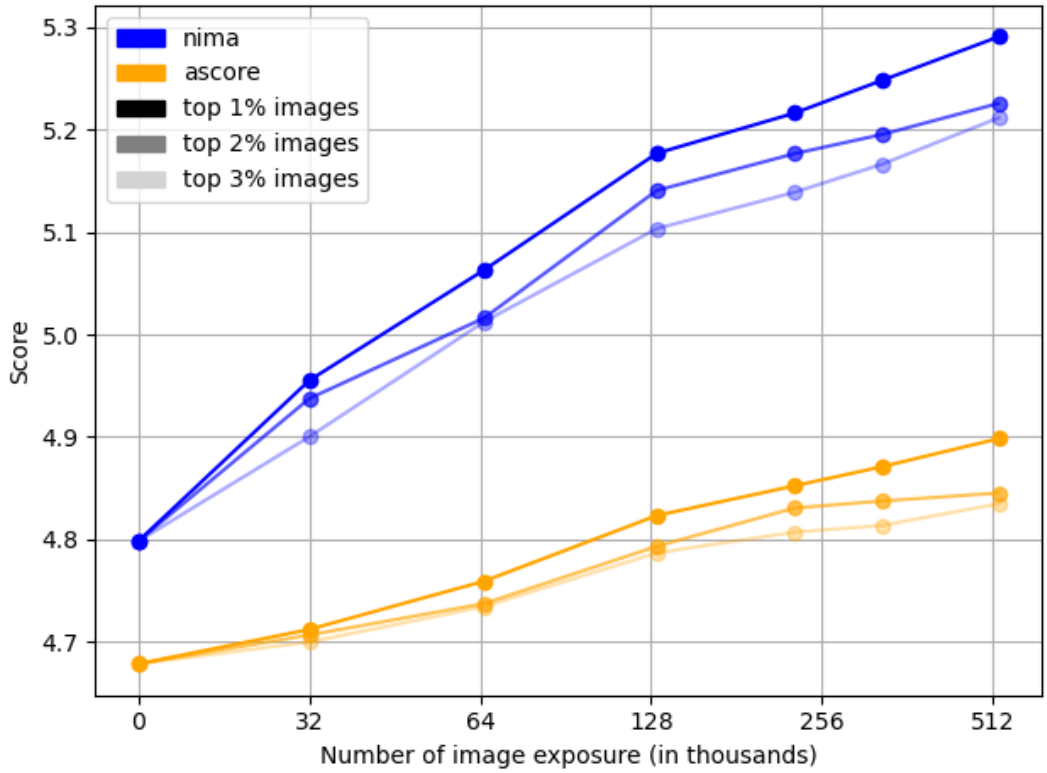
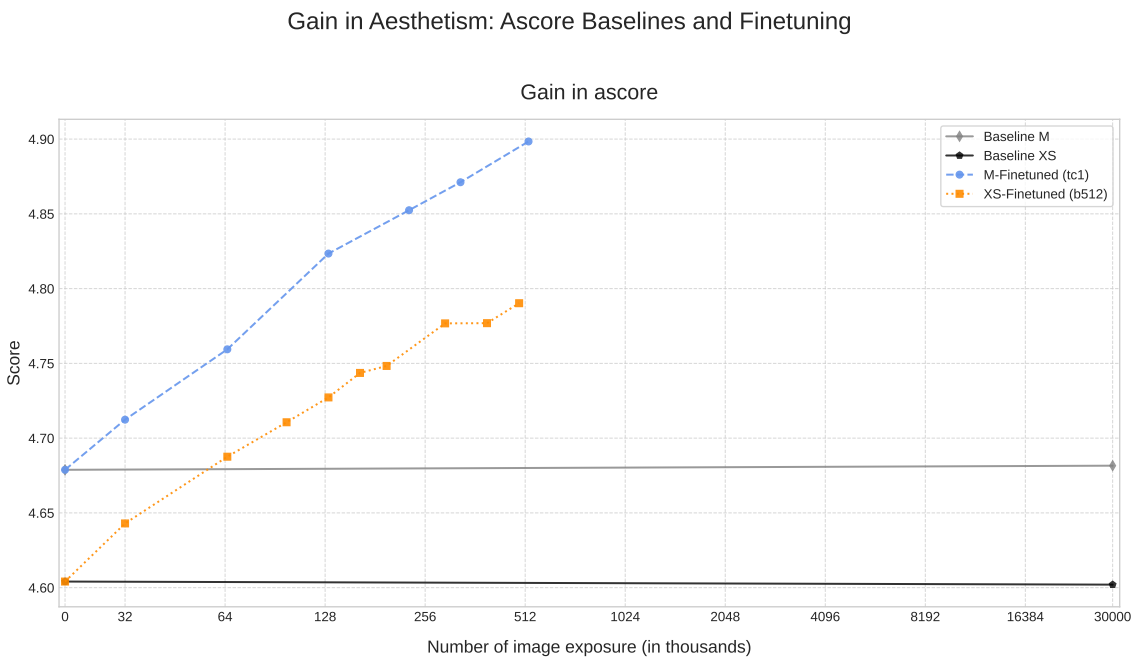
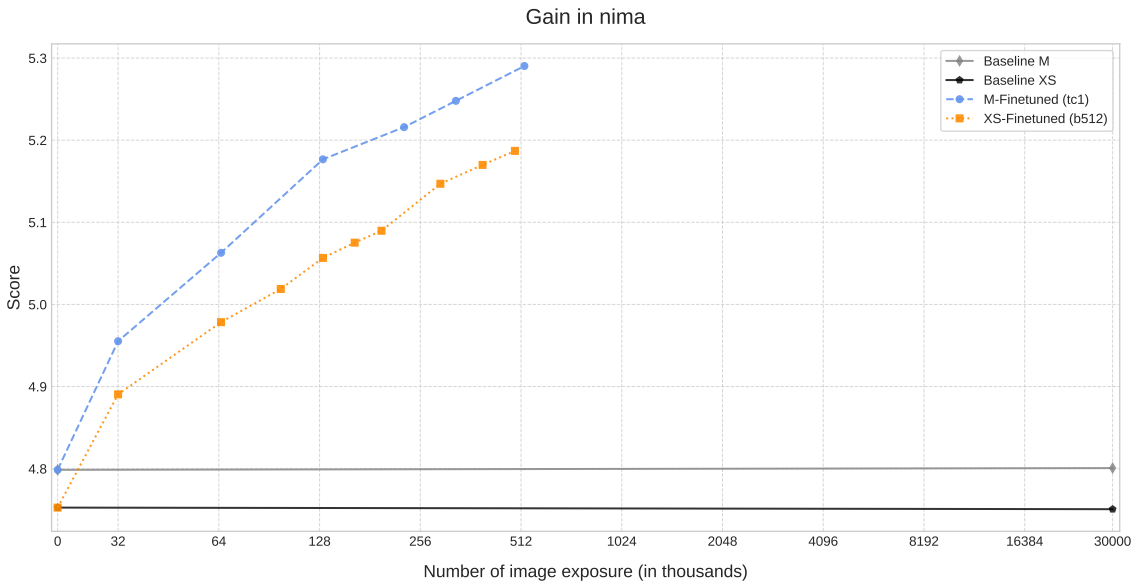


Figure 7: Comparison of NIMA & AScore Evolution for smaller high quality image subset (m, lr = 0.001, batch size=1024, warmup=100k)



(a) Gain in aesthetic score following Nima Score



(b) Gain in aesthetic score following Ascore Score

Figure 8: Aesthetic score gains across different model configurations with increasing image exposure

5 Discussion

The experiments conducted provide initial insights into our two primary hypotheses regarding the two-stage, aesthetic-driven fine-tuning approach for diffusion models. A particularly striking finding, illustrated in Figure 8, is the remarkable efficiency of our fine-tuning strategy. It demonstrates that even smaller models, such as the XS variant, when fine-tuned on a highly curated dataset—specifically, the top 5% of images selected by their original NIMA scores—can surpass the performance of a much larger M model at its final checkpoint. This superior aesthetic quality was observed across different evaluation metrics: the fine-tuned XS model achieved higher NIMA scores (as shown in Figure 8a) and higher A-scores (as shown in Figure 8b) compared to the M model's baseline. Notably, this was achieved with only

500,000 image exposures for the XS model's fine-tuning. This result strongly suggests that substantial gains in generation quality, particularly in aesthetic appeal, can be realized with significantly reduced model sizes and computational budgets by strategically focusing the fine-tuning stage on a small, high-quality data subset.

Hypothesis 1: Improved Generation with Reduced Cost. Our results, particularly from Experiment 2 with the M model, strongly support this hypothesis. We demonstrated that by fine-tuning a pre-trained model on a small (5%) subset of aesthetically high-quality images, we could achieve NIMA scores comparable to or even exceeding those of the fully trained base model's final checkpoint. This was accomplished with significantly fewer training iterations (4 million images for fine-tuning vs. an additional 30 million images for the equivalent segment of base training). This suggests a substantial potential for cost savings in terms of computation and data requirements, making high-quality generation more accessible.

The results from Experiment 1 (XXS model) indicate that model capacity is a critical factor. An extremely small model may not possess the representational power to benefit from this fine-tuning approach, or even to generate images of sufficient baseline quality for aesthetic metrics to be meaningful. The ongoing experiments with the XS model (Experiments 3 and 4) aim to further delineate the interplay between model size, fine-tuning data volume, and achievable quality, contributing to a better understanding of the "scaling laws" for this two-stage process.

Hypothesis 2: Aesthetic Score (NIMA) as an Effective Metric for Data Selection. The success of the M-model fine-tuning (Experiment 2), which relied on NIMA scores to curate the fine-tuning dataset, suggests that NIMA is indeed a valuable tool for identifying images that guide the model toward aesthetically superior generations. The consistent uplift in NIMA scores of generated samples post-fine-tuning—corroborated by improvements in A-scores underscores its ability to capture aspects of human aesthetic preference for targeted model enhancement. However, Experiment 1 also revealed a limitation: NIMA's predictions become unreliable on extremely poor-quality images, implying that the base generative model must already produce minimally coherent structure for NIMA to be an effective guide. More subtly, by focusing training on only "good" images, there is a risk of mode collapse—where the model concentrates on a narrow subset of the data distribution, sacrificing diversity for quality. To guard against this, we recommend continuously monitoring both FID and FD-DINO. Together, these metrics ensure that our targeted fine-tuning yields aesthetic gains and preserves the model's generative variety.

6 Conclusion

Overall Implications and Novelty. The primary implication of this work is the demonstration of a resource-efficient pathway to enhance the aesthetic quality of pre-trained diffusion models. By strategically selecting a small, high-impact subset of data **exclusively from the original training distribution** for a targeted second stage of fine-tuning, we can potentially circumvent some of the extensive computational burden and the need for external, often unverified, datasets typically associated with aesthetic improvement. This approach not only conserves resources but also maintains fidelity to the model's initial learned domain, opening avenues for more rapid, controlled experimentation and deployment of aesthetically-tuned models. The novelty lies in several key aspects:

1. Demonstrating that targeted fine-tuning on a curated subset of existing training data, guided by an aesthetic metric like NIMA, can significantly elevate the aesthetic appeal of generated images.
2. Showing that this improvement can be achieved with a substantially reduced number of fine-tuning steps compared to the initial pre-training, and in some cases, can lead to aesthetic scores (e.g., NIMA) that meet or exceed those of models trained for much longer on the full, unfiltered dataset.
3. Highlighting the efficiency of this method, particularly for large models like the M-size variant, where retraining or extensive fine-tuning is often prohibitively expensive.

The observation of FID increases in some M model fine-tuning runs, despite NIMA improvements, is an interesting and critical counterpoint. It suggests a potential trade-off: fine-tuning on a narrower

aesthetic distribution might enhance specific, desired visual qualities (as captured by NIMA) but could concurrently reduce the overall diversity of the generated samples, leading to a higher (worse) FID. This underscores the importance of a multi-faceted evaluation and warrants further careful investigation. Future work should involve analyzing a broader suite of diversity metrics alongside FID and NIMA to fully understand the impact of such aesthetic fine-tuning on the characteristics of the generated distribution.

Limitations. Our study is subject to several limitations. Firstly, hyperparameter optimization (e.g., learning rates, batch sizes, fine-tuning duration, and starting checkpoint selection) was constrained by computational resources; more exhaustive tuning could yield further performance gains and the emergence of clear scaling laws across independent variables. Secondly, while NIMA effectively guided aesthetic improvement, its sensitivity may be limited for images of extremely low quality. Moreover, NIMA and A-score primarily assess aesthetics and do not fully capture other quality dimensions like semantic coherence or fine-detail fidelity, which could be relevant in broader applications. Thirdly, the observed trade-off where NIMA scores improved while FID sometimes increased highlights a critical challenge. This suggests that fine-tuning on a narrow aesthetic subset can reduce sample diversity, a phenomenon not fully captured by a single FID score and necessitating evaluation with a broader suite of diversity metrics. Fourthly, our investigations with the XS model (Experiments 3 and 4), exploring reduced training steps and architectural freezing, provided initial insights into scalability and efficiency. However, a more extensive exploration across a wider range of model sizes and fine-tuning budgets is needed to establish robust scaling laws for this two-stage approach. Fifthly, the EDM2 models are pre-trained to a state of deep convergence with advanced stabilization (e.g., post-hoc EMA). The transferability of our fine-tuning efficacy to models pre-trained under different conditions (e.g., to early stopping) or lacking such stabilization is an important area for future validation. Lastly, our findings are primarily based on the EDM2 architecture and ImageNet-derived data. Generalizability to other model architectures (e.g., Latent Diffusion Models), diverse datasets with varying aesthetic priors, and different generative tasks requires further systematic investigation.

Future Work. Based on our findings and limitations, future work could include several directions. Extensive hyperparameter optimization is needed, involving a more thorough search for optimal batch sizes, warmup strategies to maximize training stability, EMA parameter, strategy of when to start the fine-tuning (at which checkpoint) and which part of the network to freeze, and fine-tuning durations, potentially using automated techniques. Exploration of alternative or combined metrics for data selection and evaluation, such as CLIP scores for text-image alignment or specific technical quality assessors, would be beneficial. Incorporating explicit diversity metrics (e.g., LPIPS-based diversity, precision/recall for distributions) is crucial to understand the impact of aesthetic fine-tuning on the variety of generated samples and to potentially optimize for both quality and diversity. Since we didn't observe mode collapse when fine-tuning on the top 1% NIMA-scored images, it may be worth reducing the dataset size further and investigating how the number of training epochs relates to dataset size before mode collapse occurs. Applying the two-stage approach to other diffusion model architectures (e.g., Latent Diffusion Models) and diverse datasets beyond ImageNet, including datasets with more specific artistic styles, will test its generalizability. In parallel, evaluating the effectiveness of this approach across different model sizes within the same architecture would shed light on how model capacity interacts with data curation. Pushing further the definition of good versus bad images by incorporating prompt alignment metrics could provide a more holistic filtering criterion beyond aesthetic quality alone. Developing a more formal theoretical understanding of why fine-tuning on small, aesthetically selected datasets can be so effective and how it influences the learned data manifold is an important long-term goal.

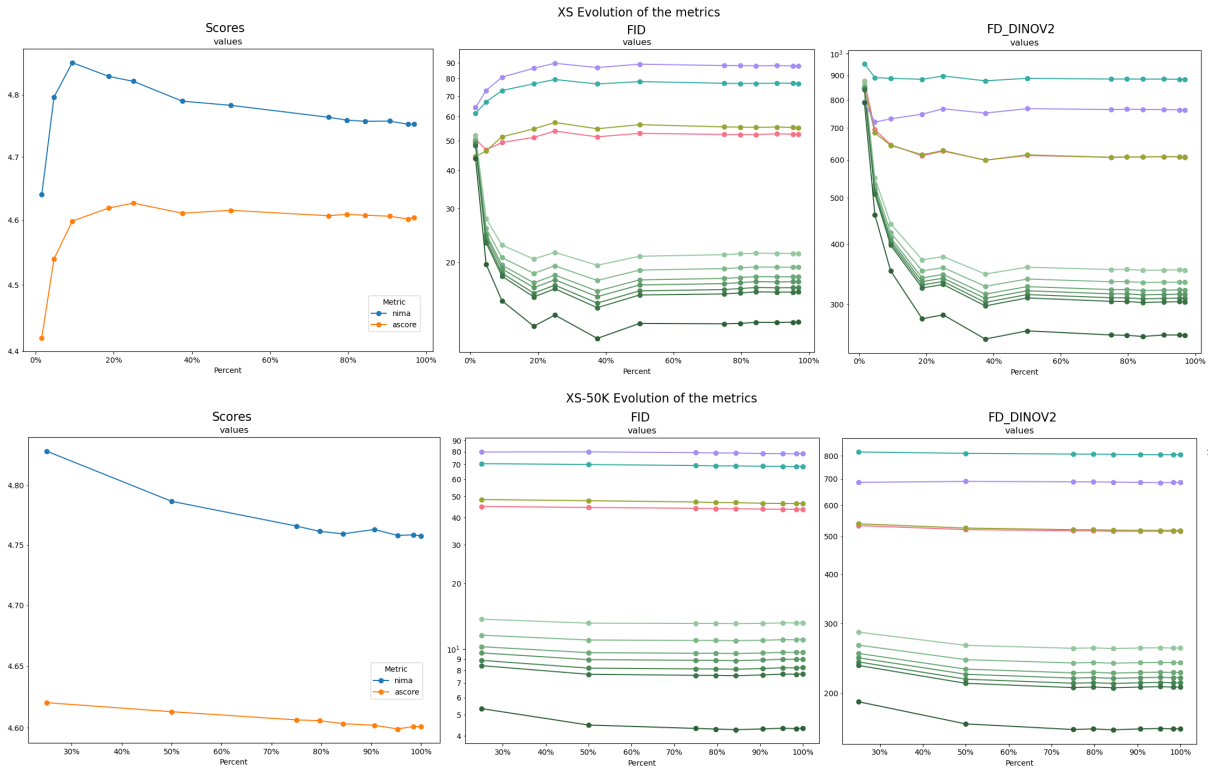


Figure 9: Comparison of 5k vs 50k for FD Dino and FID (be aware that the x-axis are not the same, fewer evaluations for 50k and first evaluation at 25%)

References

[1] Yim, J.; Park, S.; Kim, J.; Baek, M. Diffusion models in protein structure and docking. *WIREs Computational Molecular Science* **2024**, e1711.

[2] Karras, T.; Aittala, M.; Aila, T.; Laine, S. EDM2 and Autoguidance: Official PyTorch implementation. GitHub repository, NVlabs/edm2, 2023. Available online: `https://github.com/NVlabs/edm2`.

[3] Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; Yang, M. H. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint* **2022**, arXiv:2209.00796.

[4] Talebi, H.; Milanfar, P. NIMA: Neural Image Assessment. *Google Research Blog*, 2018. Available online: `https://research.google/blog/introducing-nima-neural-image-assessment/`.

[5] Talebi, H.; Milanfar, P. NIMA: Neural Image Assessment. *arXiv preprint* **2017**, arXiv:1709.05424.

[6] Schuhmann, C.; Richardson, E.; Rombach, R. LAION Aesthetic Predictor: Predicting human aesthetic judgments on the LAION-5B dataset. *arXiv preprint* **2022**, arXiv:2210.08402.

[7] Stein, G.; Cresswell, J. C.; Hosseinzadeh, R.; Sui, Y.; Ross, B. L.; Villecroze, V.; Liu, Z.; Caterini, A. L.; Taylor, J. E. T.; Loaiza-Ganem, G. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *arXiv preprint* **2023**, arXiv:2306.04675. :contentReference[oaicite:0]index=0

7 Appendix

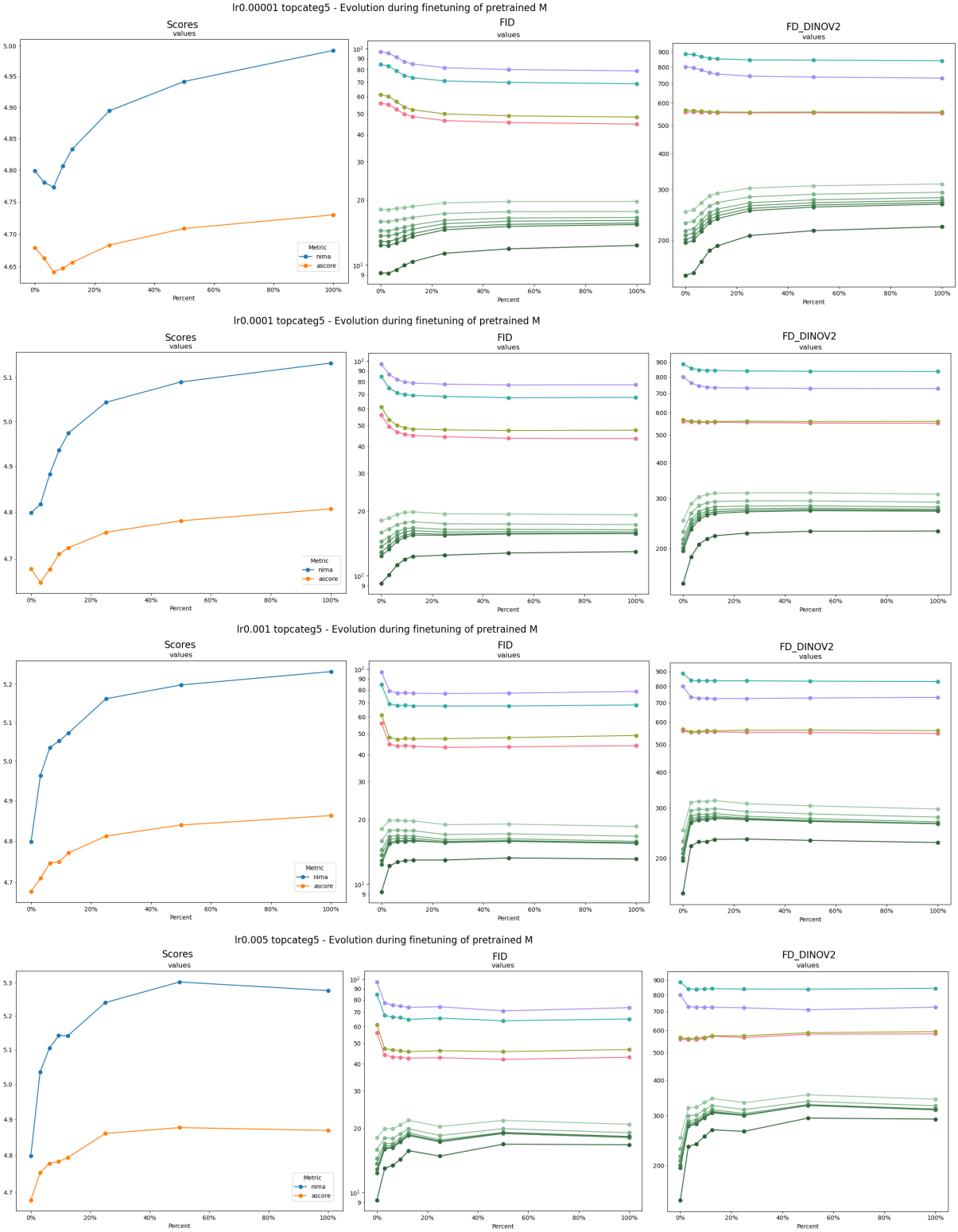


Figure 10: Experiment 2

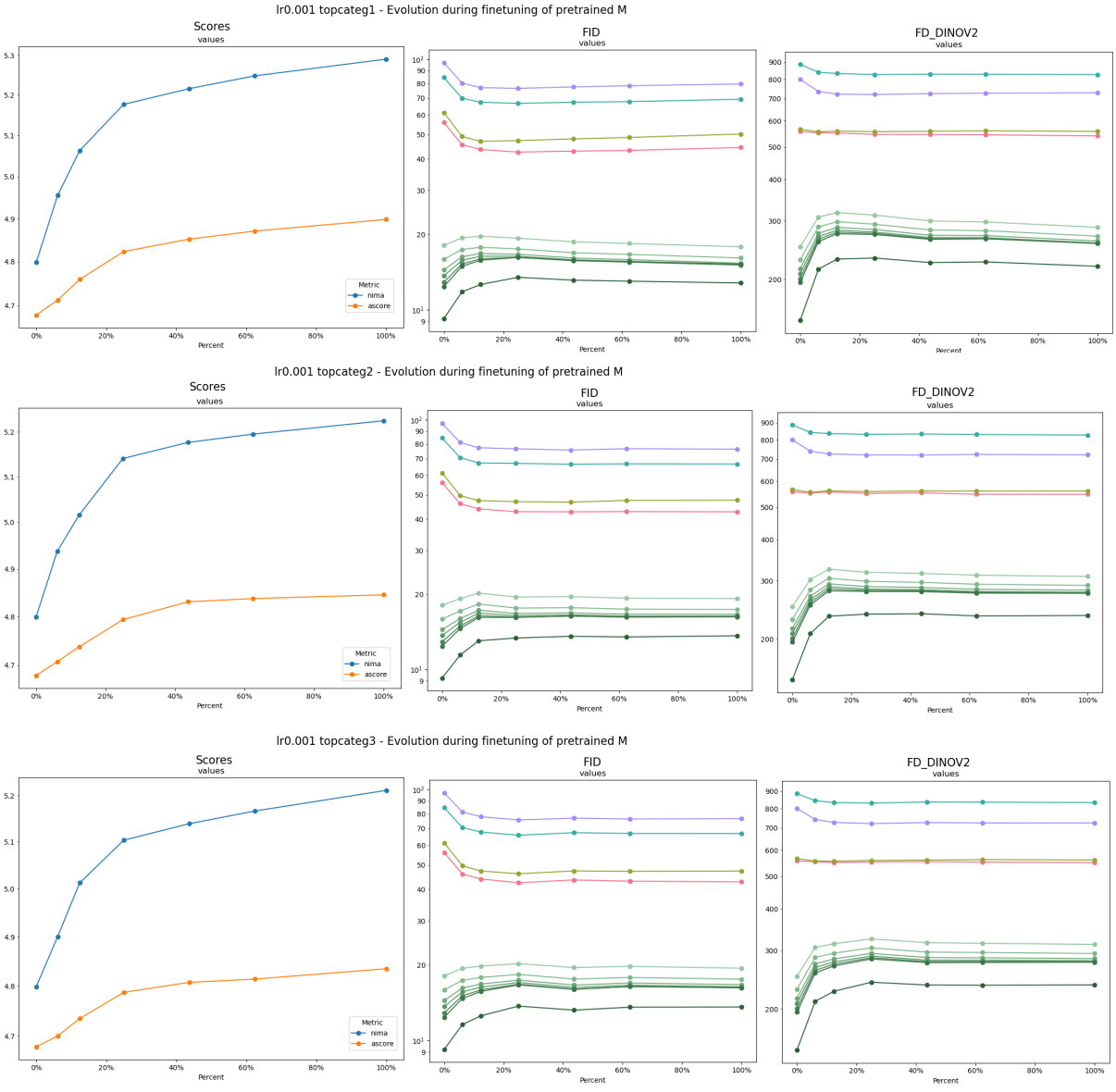


Figure 11: Experiment 5

Table 1: FD Dino scores at different training percentages for learning rate 0.01

Benchmark	Freeze Config	0.00%	3.12%	12.49%	100.00%
ava_best_25k	Down + Attention layers frozen	577.74	578.48	598.89	
	Down layers frozen	582.65	575.22	578.30	600.67
	No layers frozen	572.05	582.09	603.09	
biq2021_12k	Down + Attention layers frozen	587.77	591.86	615.07	
	Down layers frozen	586.72	586.43	591.57	618.53
	No layers frozen	583.45	594.46	619.29	
imgnet1k-01-best-nima-percentperclass	Down + Attention layers frozen	282.75	293.33	313.39	
	Down layers frozen	306.66	282.64	289.36	317.68
	No layers frozen	285.98	295.94	321.88	
imgnet1k-05-best-nima-percentperclass	Down + Attention layers frozen	252.18	265.86	290.40	
	Down layers frozen	273.74	252.43	262.30	296.00
	No layers frozen	255.76	268.50	299.53	
imgnet1k-25-best-nima-percentperclass	Down + Attention layers frozen	244.75	262.47	294.48	
	Down layers frozen	261.31	245.06	259.42	300.21
	No layers frozen	248.42	265.24	303.76	
laion25k	Down + Attention layers frozen	863.76	855.84	854.43	
	Down layers frozen	877.13	861.74	855.44	854.15
	No layers frozen	856.10	856.70	855.17	
imgnet1k-02-best-nima-percentperclass	Down + Attention layers frozen	263.41	275.53	296.87	
	Down layers frozen	286.23	263.60	271.79	301.94
	No layers frozen	266.90	277.90	305.95	
imgnet1k-10-best-nima-percentperclass	Down + Attention layers frozen	247.34	262.81	290.56	
	Down layers frozen	267.28	247.63	259.45	296.22
	No layers frozen	250.99	265.45	299.78	
imgnet1k-50-best-nima-percentperclass	Down + Attention layers frozen	243.65	263.49	298.66	
	Down layers frozen	256.34	244.01	260.65	304.48
	No layers frozen	247.40	266.13	308.16	
synthetic_full_dataset	Down + Attention layers frozen	769.92	757.50	744.60	
	Down layers frozen	765.21	769.40	758.94	739.04
	No layers frozen	761.87	757.59	739.90	
img512	Down + Attention layers frozen	209.59	232.66	303.74	
	Down layers frozen	212.33	210.08	230.59	309.22
	No layers frozen	212.27	236.61	314.17	