Project no. 3: Developing a Custom Diffusion Schedule

(Leveraging EDM2 and the CelebA Dataset)

Prit Mhala
Stony Brook ID: 117437713
Stony Brook University
pmhala@cs.stonybrook.edu

Supervised by:
Prof. Dimitris Samaras (samaras@cs.stonybrook.edu)
Ph.D. Student Mentors:
Alexandros Graikos (agraikos@cs.stonybrook.edu), Srikar Yellapragada (srikary@cs.stonybrook.edu), Kostas Triaridis (kostas@cs.stonybrook.edu)

1. Introduction

Diffusion models rely heavily on a predefined noise schedule that determines how Gaussian noise is gradually added and removed throughout the generative process, and prior work has shown that this choice significantly influences sample quality, stability, and convergence [3]. Although several schedules such as Karras ρ, linear, cosine, logarithmic, and other heuristic variants have been proposed, there is still no clear consensus on how to choose an optimal schedule across different datasets or model configurations [1]. The primary objective of this project is to study this open question by benchmarking multiple diffusion schedules using the EDM2 architecture, which introduces improved preconditioning and training dynamics for diffusion models [2]. Using the CelebA dataset of human faces [4], downscaled to 64×64 resolution, I train separate EDM2-XS models under five different schedules and evaluate their performance using Fréchet Inception Distance (FID). Through this comparison, the goal is to develop an empirical understanding of how each schedule affects training behavior, reconstruction trajectories, and overall image synthesis quality.

2. Dataset Description

The CelebA dataset is a large-scale face dataset widely used in computer vision research for tasks such as attribute recognition, face detection, and generative modeling. It contains 202,599 cropped and aligned face images spanning 10,177 unique identities, offering significant diversity in pose, illumination, expression, and background. Along with high-quality images, the dataset includes well-defined train–validation–test splits and several auxiliary annotation files such as evaluation partitions, landmark coordinates, bounding boxes, and attribute labels. This structured organization makes CelebA straightforward to preprocess and highly suitable for experimentation in both discriminative and generative tasks. Due to its size, quality, and versatility, CelebA has become a standard benchmark for evaluating deep learning models in face-related applications [4].

For this project, CelebA is an ideal choice because it provides a controlled yet diverse dataset to analyze how different diffusion noise schedules influence model behavior and sample quality. The initial focuses on unconditional image generation, the dataset's consistent face structure allows the EDM2 model to learn meaningful representations without incorporating conditioning signals. CelebA is particularly well suited for the second phase of the project, where the focus shifts to conditional diffusion, because it offers three rich sources of conditioning information: (1) five

visual landmarks including the eyes, nose, and mouth, and (2) 40 binary facial attributes describing appearance characteristics. These structured annotations create an ideal setup for studying how conditioning strength and type interact with the choice of diffusion noise schedule.

3. EDM2 Architecture Overview

EDM2 is a recent diffusion model architecture introduced by NVIDIA Research that focuses on improving the training dynamics and sample quality of diffusion models through principled noise preconditioning and stable ODE-based sampling [1]. It extends the original EDM framework [3] by introducing refined σ-data preconditioning, improved handling of noise levels, and a Heun 2nd-order sampler that enables more consistent convergence across different noise schedules. For this project, EDM2 is an ideal choice because it provides a stable and well-engineered baseline that allows the effect of the noise schedule to be studied in isolation. The architecture is also optimized for low-resolution datasets such as CelebA 64×64, making experimentation computationally feasible while still yielding meaningful differences in FID across schedules. The official PyTorch implementation released by NVIDIA [2] offers modular hooks that make it straightforward to patch and modify the noise schedule, which is essential for the systematic comparisons in this study.

For computational efficiency, this project uses the **EDM2-XS (Extra Small) configuration**, a lightweight version of the model containing approximately **125 million parameters**, designed specifically for fast experimentation under limited GPU resources. The XS variant preserves the core EDM2 components such as residual blocks, attention layers, and σ-conditioning while reducing depth and channel width to fit within the memory limits of commonly available GPUs like the **NVIDIA T4**. Since all experiments were conducted on single or dual T4 GPUs, the XS-64 architecture provides an effective trade-off between training time and generative quality, enabling multiple noise schedules to be benchmarked within practical time and resource constraints.

4. Methodology

4.a Downscaling of CelebA Dataset to 64 × 64

To train EDM2 efficiently on CelebA, the original 178×218 images were downscaled to 64×64 resolution. Each image was then center-cropped to a square 178×178 region to remove background imbalance and ensure uniform alignment across samples. The cropped image was downscaled to 64×64 using the LANCZOS filter, which preserves sharpness and minimizes aliasing artifacts, and saved at high JPEG quality to avoid unnecessary information loss. To verify that the resolution reduction did not introduce significant distortion, a subset of 100 random images was evaluated using PSNR and SSIM, yielding average scores of 30.29 dB and 0.8968 respectively, demonstrating good structural fidelity after downsampling.

4.b Noise Schedules

Karras ρ Schedule The Karras schedule [3] focuses sampling effort at low noise levels, where denoising is most difficult. It uses a power-law interpolation between $\sigma_{\min}$ and $\sigma_{\max}$, controlled by the exponent ρ (typically 7).

$$\sigma(t) = \left[\sigma_{\max}^{1/\rho} + t\left(\sigma_{\min}^{1/\rho} - \sigma_{\max}^{1/\rho}\right)\right]^{\rho}$$

This adaptive curve improves stability and detail recovery near the clean-image regime.

Linear Schedule The linear noise schedule increases variance uniformly with timestep, as in the original DDPM formulation (Ho et al., 2020) [5] .

$$\beta_t = \beta_{\min} + t(\beta_{\max} - \beta_{\min})$$

It produces even noise progression but may under-allocate steps at low noise levels where denoising is harder.

Cosine Schedule The cosine schedule (Nichol & Dhariwal, 2021) [6] smooths the variance progression using a cosine curve, improving robustness at both low and high noise levels.

$$\hat{\alpha}_t = \frac{\cos^2\left(\frac{t/T+s}{1+s}\frac{\pi}{2}\right)}{\cos^2\left(\frac{s}{1+s}\frac{\pi}{2}\right)}$$

This results in stable behavior and typically fewer artifacts during late denoising.

Logarithmic Schedule The logarithmic schedule performs linear interpolation in log-space, producing exponential decay of noise in real space.

$$\sigma(t) = \exp(\log \sigma_{\min} + t(\log \sigma_{\max} - \log \sigma_{\min}))$$

This offers uniform spacing in the log domain and smooth progression across diffusion scales.

Quadratic Schedule The quadratic schedule increases noise following a t^2 curve, leading to slow initial growth and faster increase near the end.

$$\sigma(t) = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min})\, t^2$$

It provides a gentle start and a controlled rise in noise, serving as a balanced middle ground between linear and more aggressive curves.

4.c Unconditional setup

For the Karras ρ schedule, the official EDM2 framework was cloned into the Kaggle environment and used without modification, retaining the default `EDM2Loss` formulation that samples noise according to the Karras power-law distribution. All models were trained on the CelebA 64×64 training set using the `edm2-img64-xs` preset, with a batch size of 64 across two NVIDIA T4 GPUs, for a total of 2 million training images (2 Mi). The final checkpoint, `network-snapshot-0002097-0.100.pkl`, was used for evaluation. Model performance was assessed by generating 1,000 samples at sampling steps $\{8, 16, 24, 32\}$ and computing FID against 1,000 CelebA test images embedded once using Inception-V3.

For the Linear, Cosine, Logarithmic, and Quadratic schedules, the overall training and evaluation pipeline was kept identical, with the only change being the noise sampling rule. Specifically, the `EDM2Loss.__call__` method was patched so that each forward pass sampled σ from the corresponding schedule instead of the default log-normal Karras distribution. No other hyperparameters were altered, ensuring strict comparability across schedules.

To maintain consistency between training and inference, the EDM2 ODE sampler was similarly patched at evaluation time to follow the same schedule-specific σ progression used during training. For example, under the Linear schedule, noise levels were sampled as $\sigma(t) = \sigma_{\max} - (\sigma_{\max} - \sigma_{\min})t$

with $t \sim \mathcal{U}(0,1)$ during training, and decreased linearly from $\sigma_{\max}$ to $\sigma_{\min}$ during sampling. The standard second-order Heun solver was retained for all experiments. All models were evaluated using 32 sampling steps, ensuring that observed differences in sample quality and FID arise solely from the choice of noise schedule.

4.d Conditional setup using 40 binary attributes

To support attribute-conditioned diffusion, the EDM2 training pipeline was extended with a custom dataset class, `CelebAAttributesDataset`, which replaces the default image-only loader when attribute conditioning is enabled. Each image is paired with a 40-dimensional binary attribute vector derived from the official CelebA annotations, with original $\{-1, +1\}$ labels converted to $\{0, 1\}$ for numerical stability. These attribute vectors are provided directly as conditioning labels to the EDM2 network during training, while all architectural components, loss formulation, and hyperparameters remain identical to the unconditional setup.

For evaluation, a fixed subset of about 20,000 images from the CelebA test split was used, with corresponding attribute vectors and filenames stored to ensure deterministic sampling. During inference, each noise sample was conditioned on its associated attribute vector and passed to the EDM2 sampler using 32 denoising steps and the same noise schedule as training. This consistent setup ensures that differences in visual quality or attribute controllability can be attributed solely to the conditioning mechanism rather than variations in sampling depth or noise parameterization.

4.e Conditional setup using facial landmark attributes

To enable geometric conditioning, the EDM2 training pipeline was extended with a custom dataset class, `CelebALandmarkDataset`, which conditions generation on five facial landmarks represented as ten continuous (x, y) coordinates corresponding to the eyes, nose, and mouth. Each image was paired with its landmark vector loaded from a preprocessed CSV file, and only images with valid landmark annotations were included during training. These landmark vectors were provided directly as conditioning labels to the EDM2 network, while all other architectural components, loss formulation, and training hyperparameters were kept identical to the unconditional setup.

For evaluation, landmark vectors were prepared for the entire CelebA test split to ensure deterministic and reproducible sampling. During inference, each noise sample was conditioned on its corresponding additional 10-dimensional landmark vector and generated using the EDM2 sampler with 32 denoising steps and the same noise schedule as training. This consistent setup allows the effect of spatial landmark conditioning to be analyzed independently of changes in model architecture or diffusion dynamics.

5. Results

a. Results on Unconditional Setup

The performance of each noise schedule was first evaluated in a controlled setting where the trained models generated 1,000 samples using four different sampling step counts (8, 16, 24, 32). Table 1 summarizes the corresponding FID scores. Across all schedules, the FID decreases with increasing step count, reflecting the expected improvement in ODE solver accuracy as more integration steps are used. Among the schedules, the Linear, Logarithmic, and Quadratic variants achieved slightly lower FIDs than Karras ρ and Cosine at nearly all step counts, although the overall spread remained modest at higher step values (approximately 32.1–32.3).

A second evaluation was performed on the full CelebA test set of 19,962 images. Here, the Linear schedule achieved the best overall FID (10.7467), followed closely by Logarithmic (10.7609) and Quadratic (10.7644). Cosine (10.8015) and Karras ρ (10.8144) yielded marginally higher FIDs. These results indicate that, for CelebA 64×64, schedules that provide smoother or more uniform progression in noise magnitude tend to produce slightly better generative performance than curvature-heavy schedules such as Cosine or the power-law Karras formulation.

Table 1: FID values for 1,000 generated samples evaluated at different sampling step counts.

Schedule	**8 Steps**	**16 Steps**	**24 Steps**	**32 Steps**
Linear	46.1609	33.4561	32.4373	32.1901
Karras ρ	46.2117	33.5683	32.4998	32.2529
Cosine	46.2130	33.5477	32.4988	32.2639
Logarithmic	46.2031	33.5049	32.4734	32.2019
Quadratic	46.1658	33.5179	32.4838	32.2639

Table 2: FID results computed on the full CelebA test set (19,962 images).

Schedule	**FID**
Linear	10.7467
Logarithmic	10.7609
Quadratic	10.7644
Cosine	10.8015
Karras ρ	10.8144

Figures 1–5 show qualitative samples generated at 32 sampling steps for all five noise schedules. The overall structure and identity consistency are similar across schedules, but subtle differences in sharpness and contrast can be observed. In particular, the Linear, Logarithmic, and Quadratic schedules (Figures 2–4) tend to produce slightly crisper textures compared to the softer outputs of Karras-ρ and Cosine (Figures 1 and 5), matching the FID trends reported earlier.

(a) 00025 (b) 00026 (c) 00027 (d) 00028 (e) 00029

Figure 1: Generated samples at 32 steps using the Karras-ρ noise schedule

(a) 00025 (b) 00026 (c) 00027 (d) 00028 (e) 00029

Figure 2: Generated samples at 32 steps using the Linear noise schedule

Figures 6–8 illustrate the denoising trajectories produced by different noise schedules at increasing sampling depths. At 4 sampling steps (Figure 6), all schedules struggle to recover meaningful facial structure, with outputs dominated by high-frequency noise. Early differences are still observable: the Karras-ρ and Logarithmic schedules begin to reveal coarse facial features slightly

(a) 00025 (b) 00026 (c) 00027 (d) 00028 (e) 00029

Figure 3: Generated samples at 32 steps using the Logarithmic noise schedule

(a) 00025 (b) 00026 (c) 00027 (d) 00028 (e) 00029

Figure 4: Generated samples at 32 steps using the Quadratic noise schedule

(a) 00025 (b) 00026 (c) 00027 (d) 00028 (e) 00029

Figure 5: Generated samples at 32 steps using the Cosine noise schedule

earlier than Linear and Quadratic, while the Cosine schedule remains heavily noise-dominated at this low step count.

As the number of steps increases to 10 (Figure 7), structural details such as facial contours, eye placement, and skin tone become more apparent across most schedules. Logarithmic and Karras-ρ show faster perceptual convergence, producing recognizable faces at intermediate noise levels, whereas Linear and Quadratic exhibit a more gradual transition from noise to structure. The Cosine schedule continues to lag in early detail recovery, although it maintains smooth transitions near lower noise levels.

At 32 sampling steps (Figure 8), all schedules converge to visually coherent and high-quality samples, with only subtle differences in sharpness and texture. Linear, Logarithmic, and Quadratic schedules produce slightly crisper facial details, while Karras-ρ and Cosine yield marginally smoother appearances. These qualitative trends align with the quantitative FID results, confirming that increased sampling depth reduces schedule-dependent variance and that noise scheduling primarily influences convergence speed rather than final sample fidelity.

b. Results on Conditional Setup with 40 attributes

Figures 9–11 show qualitative results for attribute-conditioned diffusion using 40-dimensional CelebA attribute vectors under the Karras-ρ noise schedule. In the minimal attribute setting (Figure 9), conditioning was derived from image ID 183075.jpg, which activates only a single attribute (*No Beard*). As a result, the denoising trajectory closely resembles the unconditional case, with facial structure emerging gradually and limited semantic variation observed at lower σ values.

In contrast, the maximal attribute configuration (Figure 10) uses image ID 184620.jpg, activating 20 attributes including *Arched Eyebrows*, *Black Hair*, *Heavy Makeup*, *Smiling*, and *Young*. This dense semantic conditioning leads to stronger and earlier attribute driven structure during denoising, with facial appearance becoming more consistent across samples at intermediate noise

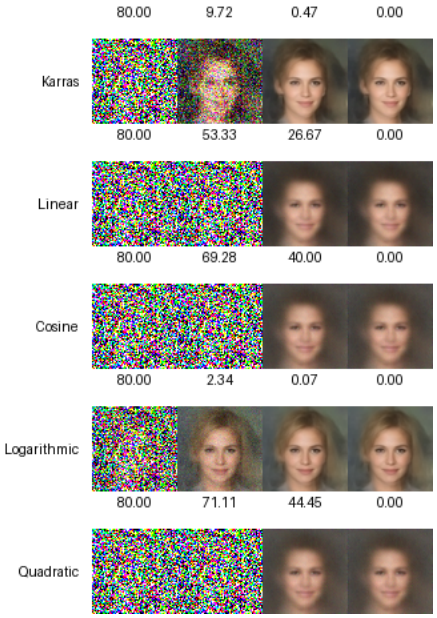

Figure 6: Unconditional samples generated using different noise schedules with 4 sampling steps.

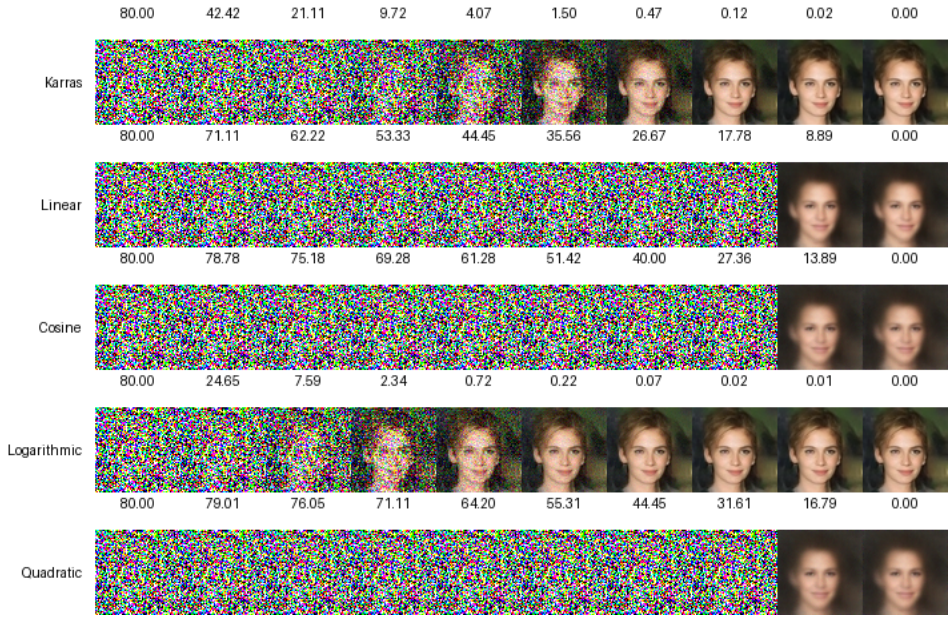

Figure 7: Unconditional samples generated using different noise schedules with 10 sampling steps.

levels.

Figure 11 demonstrates selective conditioning using image ID 182797.jpg, where a semantically focused subset of 15 attributes—most notably *Bald* and *Mustache*—is activated. In this case, the model reliably enforces the specified attributes while allowing other facial details to vary naturally. Overall, these results confirm that EDM2 effectively integrates high dimensional attribute conditioning, achieving controllable generation with an overall FID of 20.2637 on the CelebA test set.

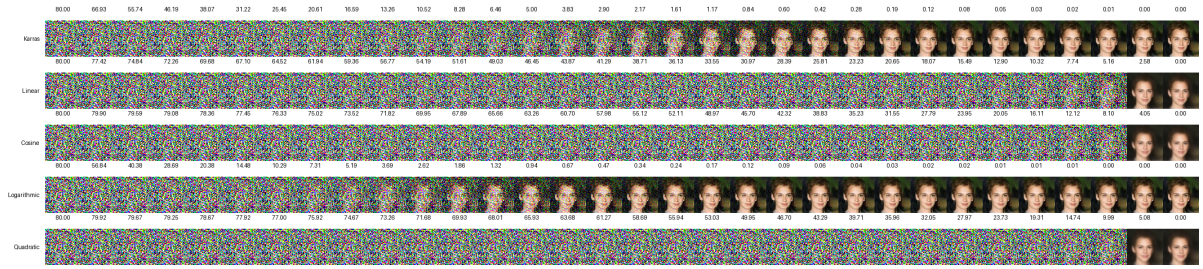

Figure 8: Unconditional samples generated using different noise schedules with 32 sampling steps.

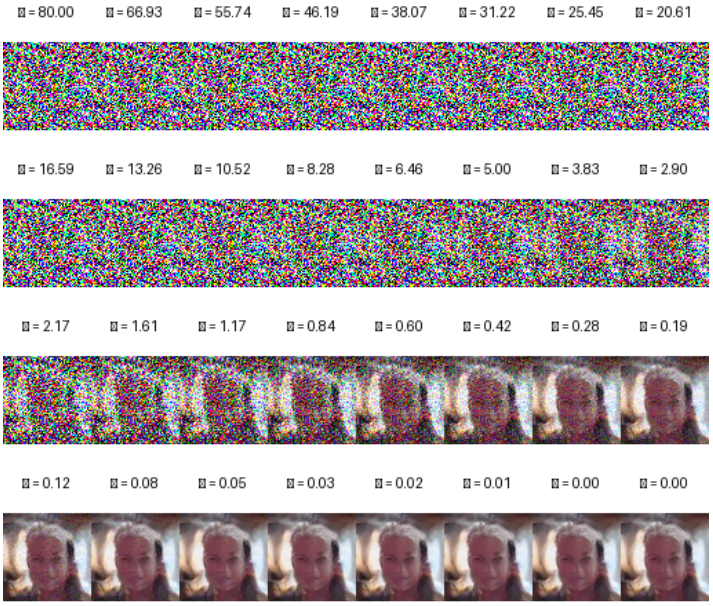

Figure 9: Conditional generation using minimal CelebA attribute configuration (40 attributes).

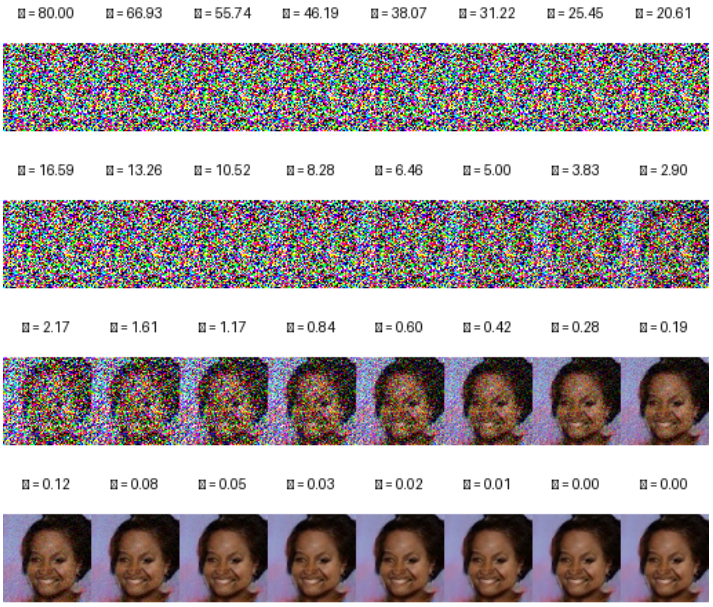

Figure 10: Conditional generation using maximal CelebA attribute configuration (40 attributes).

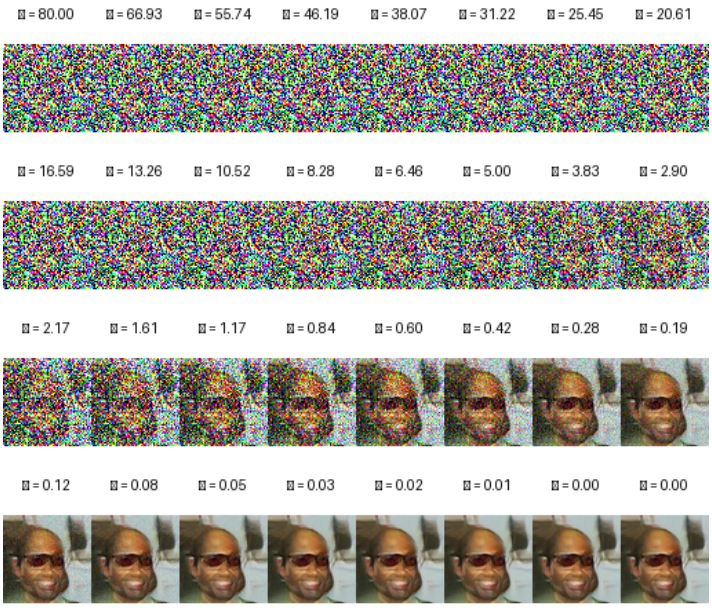

Figure 11: Conditional generation using selected attributes (bald and mustache).

c. Results on Conditional Setup with facial landmark attributes

Figures 12–14 present qualitative results for landmark-conditioned diffusion, where facial geometry is controlled using five landmark points (eyes, nose, and mouth). In the symmetric landmark configuration (Figure 12), conditioning was derived from image ID 182643.jpg, where the left and right eyes are nearly horizontally aligned. This symmetry leads to stable frontal face generation, with balanced facial structure emerging consistently as noise decreases.

The minimal landmark configuration (Figure 13) corresponds to image ID 193355.jpg, which exhibits the smallest inter eye distance among the test samples. In this case, the generated faces show compressed facial geometry during intermediate denoising stages, demonstrating that reduced landmark extent directly constrains spatial structure while still allowing identity level variation.

Conversely, the maximal landmark configuration (Figure 14) uses image ID 198369.jpg, characterized by the largest inter eye distance. This setting produces faces with larger spatial extent and more pronounced geometric variation, where landmark conditioning strongly influences early stage layout and finer texture details are resolved at lower noise levels. Overall, landmark conditioning provides effective geometric control over the generative process, achieving an overall FID of 28.3906 on the CelebA test set.

6. Conclusion

This project presented a systematic empirical study of diffusion noise schedules within the EDM2-XS framework, using the CelebA 64×64 dataset under both unconditional and conditional generation settings. Five noise schedules Karras ρ, Linear, Cosine, Logarithmic, and Quadratic were evaluated using a strictly controlled training and inference pipeline, ensuring that observed differences arose solely from the choice of noise schedule. In the unconditional setting, all schedules exhibited consistent convergence behavior, with Fréchet Inception Distance (FID) improving monotonically as the number of sampling steps increased. Among the schedules, Linear, Logarithmic, and Quadratic consistently achieved slightly lower FID scores than Karras ρ and Cosine, both in the 1,000-sample evaluation and on the full CelebA test set, where the Linear schedule attained

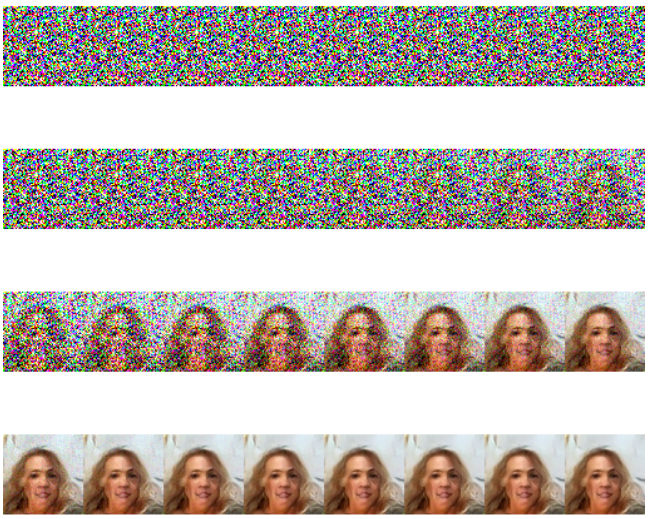

Figure 12: Landmark-conditioned generation using symmetric facial landmark configuration.

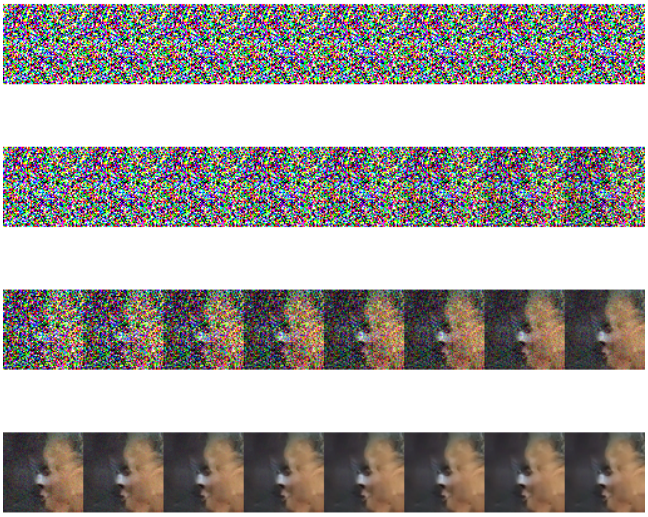

Figure 13: Landmark-conditioned generation using minimal facial extent landmarks.

the best FID of 10.7467. These results indicate that smoother or more uniform noise progressions can offer modest but consistent gains in generative quality, while EDM2 remains robust across a wide range of σ-parameterizations.

Beyond unconditional generation, the study was extended to conditional diffusion using two complementary forms of conditioning available in CelebA: high-dimensional semantic attributes and low-dimensional geometric landmarks. In the 40-attribute conditional setup, EDM2 demonstrated strong semantic controllability, achieving a full-dataset FID of 20.2637 while reliably enforcing both dense and sparse attribute configurations. Qualitative results showed that richer attribute vectors lead to earlier emergence of semantically meaningful structure during denoising, while sparse attribute combinations allow controlled manipulation of specific facial traits. Attribute-wise FID analysis further revealed a clear relationship between attribute frequency and generative performance, with rare attributes such as *Wearing Hat* and *Bald* exhibiting significantly higher FID scores than common attributes like *Young* or *Wearing Lipstick*, highlighting the impact

Figure 14: Landmark-conditioned generation using maximal facial extent landmarks.

of data imbalance on conditional diffusion quality.

In the landmark-conditioned setting, the model achieved an overall FID of 28.3906 and demonstrated strong geometric control over facial structure. Conditioning on five facial landmarks enabled precise manipulation of spatial layout, with symmetric, minimal, and maximal landmark configurations producing predictable changes in facial geometry across the denoising trajectory. These experiments showed that landmark conditioning primarily influences early-stage structure formation, while texture and identity-specific details are refined at later noise levels. Together, the conditional results illustrate that EDM2 can effectively integrate both semantic and geometric conditioning signals while preserving stable diffusion dynamics.

Table 3 reports the attribute-wise sample counts in the CelebA dataset along with the corresponding FID scores obtained under attribute-conditioned diffusion. A clear relationship between attribute frequency and generative quality can be observed. Attributes with large sample counts, such as *Young*, *Wearing Lipstick*, *Attractive*, and *No Beard*, consistently achieve lower FID values (approximately 18–21), indicating that the model is able to learn strong and stable conditional representations when sufficient training data is available. These frequent attributes provide dense supervision, enabling the diffusion model to reliably enforce semantic constraints during denoising.

In contrast, rare attributes exhibit substantially higher FID scores, reflecting degraded sample quality and weaker conditional control. Attributes such as *Wearing Hat* (FID 77.13), *Bald* (FID 67.32), *Gray Hair* (FID 52.56), and *Mustache* (FID 53.40) have very limited representation in the dataset, leading to insufficient coverage of appearance variations during training. As a result, the model struggles to accurately capture these semantics, producing noisier or less consistent samples. This trend highlights the sensitivity of attribute-conditioned diffusion models to class imbalance and demonstrates that conditional performance is strongly influenced by the availability of attribute-specific training examples rather than the conditioning mechanism itself.

Figure 15 illustrates the sampling ratios of each facial attribute across the training and test splits of the CelebA dataset. The distributions closely match for nearly all attributes, indicating that the train–test partitioning does not introduce significant skew or attribute-specific bias. This balanced split ensures that the observed differences in attribute-wise FID scores are not artifacts of mismatched evaluation distributions. Consequently, performance variations can be attributed to intrinsic data availability and model behavior rather than inconsistencies between training and

Table 3: Attribute-wise sample count and FID scores for CelebA attribute-conditioned generation.

Attribute	Count	FID	Attribute	Count	FID
5_o_Clock_Shadow	1994	34.84	Arched_Eyebrows	5678	20.19
Attractive	9898	18.88	Bags_Under_Eyes	4045	27.74
Bald	423	67.32	Bangs	3109	26.03
Big_Lips	6528	21.51	Big_Nose	4232	27.65
Black_Hair	5422	24.33	Blond_Hair	2660	24.89
Blurry	1010	46.29	Brown_Hair	3587	22.99
Bushy_Eyebrows	2586	28.71	Chubby	1058	43.73
Double_Chin	913	43.83	Eyeglasses	1289	43.42
Goatee	915	48.07	Gray_Hair	636	52.56
Heavy_Makeup	8084	19.35	High_Cheekbones	9618	20.42
Male	7715	27.08	Mouth_Slightly_Open	9883	21.72
Mustache	772	53.40	Narrow_Eyes	2968	28.25
No_Beard	17041	19.95	Oval_Face	5901	22.80
Pale_Skin	840	41.73	Pointy_Nose	5704	21.51
Receding_Hairline	1694	35.36	Rosy_Cheeks	1432	29.48
Sideburns	926	47.78	Smiling	9987	20.64
Straight_Hair	4190	26.36	Wavy_Hair	7267	19.57
Wearing_Earrings	4125	23.52	Wearing_Hat	839	77.13
Wearing_Lipstick	10418	18.76	Wearing_Necklace	2753	25.19
Wearing_Necktie	1399	47.32	Young	15114	20.02

testing sets.

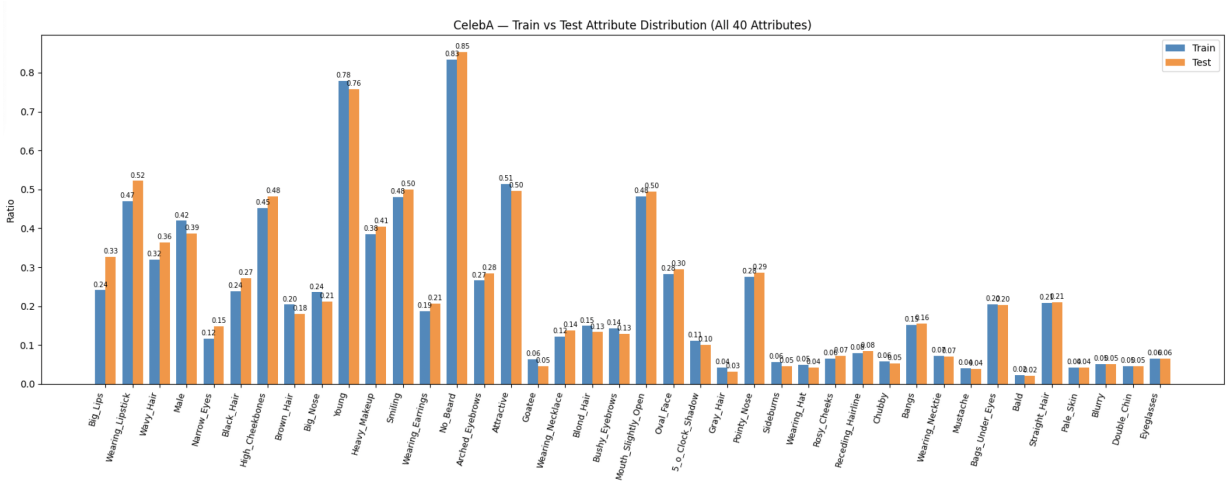

Figure 15: Sampling ratio per attribute bar graph for train and test data

7. Future Scope

The next phase of this project will extend the current analysis of noise schedules to the conditional diffusion setting. Specifically, both conditional configurations explored in this work conditioning on 40 binary facial attributes and conditioning on facial landmark coordinates will be implemented

and evaluated under the remaining noise schedules: Linear, Cosine, Logarithmic, and Quadratic. By applying identical training and evaluation protocols across all schedules, this extension will enable a direct comparison of how different noise parameterizations interact with semantic and geometric conditioning signals.

This future investigation will help determine whether the trends observed in the unconditional setting, where smoother noise schedules achieved marginally better FID scores, persist under conditional generation. It will provide deeper insight into how noise schedules influence conditional controllability, convergence speed, and robustness when conditioning complexity increases. Such results will contribute toward identifying schedule–conditioning combinations that offer the best trade-off between generative quality and control.

References

[1] Karras, T., Aittala, M., Aila, T., and Laine, S. (2024). *Analyzing and Improving the Training Dynamics of Diffusion Models (EDM2)*. CVPR 2024.

[2] NVIDIA Research. (2024). *Official PyTorch implementation of EDM2*. GitHub Repository.

[3] Karras, T., Aittala, M., Laine, S., and Aila, T. (2022). *Elucidating the Design Space of Diffusion Models (EDM)*.

[4] Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). *Deep Learning Face Attributes in the Wild (CelebA Dataset)*. ICCV 2015.

[5] Ho, J., Jain, A., and Abbeel, P. (2020). *Denoising Diffusion Probabilistic Models*. NeurIPS 2020. Available at: `https://arxiv.org/abs/2006.11239`.

[6] Nichol, A., and Dhariwal, P. (2021). *Improved Denoising Diffusion Probabilistic Models*. arXiv preprint. Available at: `https://arxiv.org/abs/2102.09672`.