Deep generative models for engineering design

Fan, J.

Chapter 2

Generation of Plausible Designs

In generative engineering design, the task of DGM is to generate structural meaningful designs, which refers to the plausibility criterion in the field of DGM. This chapter studies the diffusion model (the state-of-the-art DGM backbone) for structural designs and proposes a solution that prioritizes the plausibility when using DGMs in generating designs, aiming to addressing the research question 1: *How to prioritize plausibility in generation with DGMs?* The content of this chapter has been published in the paper [167].

(a) Implausible bicycles generated by EDM [71]

(b) Bicycles generated by our PoDM [167]

Figure 2.1: Generated bicycle designs. Our work aims to minimize the proportion of implausible designs generated by focusing on a certain range of noise levels.

2.1 Introduction

Diffusion-based generative models have been reported to surpass GANs [49, 133, 72, 165] in various image synthesis tasks [61, 41, 71]. Despite the success of diffusion-based models, several issues exist to address when generating design structures. For instance, denoising diffusion probabilistic models (DDPM) [61] can generate structure images with high visual quality; however, it suffers from a slow generation speed due to the usage of an excessive number of denoising steps [79, 156, 71]. As a remedy, the denoising diffusion implicit model (DDIM) [156] greatly reduces the denoising steps, which, however, compromises the visual quality. Recently, the influential work "Elucidating the Design Space of Diffusion-Based Generative Models" (which introduced the method termed EDM) [71] incorporated a probability distribution to sample noise levels during the denoising process. This approach enhances generation speed while maintaining satisfactory visual quality, particularly excelling in rendering image details such as human hair curls and skin pores. However, when assessing the plausibility of the generated design images, we observe that the EDM often produces structurally implausible images, see Figure 2.1a.

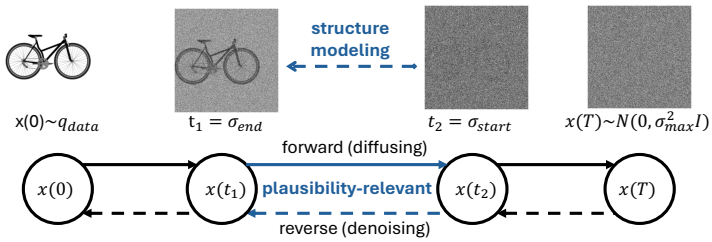


Figure 2.2: Forward and reverse processes of diffusion models in directed graphs, with a highlight on the plausibility-relevant range of noise levels, i.e., $[\sigma_{\text{end}}, \sigma_{\text{start}}]$.

For the denoising process, recent research has explored the impact of the noise schedule on the properties of generated images [114, 158, 159, 71]. For example, while lower noise levels can enhance the quality of image details [71], higher noise levels can affect the diversity of generated results [158]. In this chapter, we discover that in using diffusion models, there exists a plausibility-relevant range of noise levels that predominantly affect the plausibility of the images (see Figure 2.4). More importantly, this range can be determined by the evolution of pixel-value distributions in the forward diffusion process (see Section 2.2.2). We visualize the plausibility-relevant range and its relevance to structural generation in Figure 2.2. To determine the plausibility-

relevant range, we simulate the forward diffusion process on real structural designs and trace the distribution of pixel values as the noise level increases. We observe that the disappearance of the structural signal has a clear corresponding phase in the development of pixel-value distributions.

Taking this observation, we modify the training and generation procedures of EDM to prioritize sampling the noise levels in the plausibility-relevant range (see Figure 2.3a for an illustration), resulting in a new method, plausibility-oriented diffusion model (PoDM). We experimentally test PoDM on three datasets, the BIKED dataset [137] (see some generated examples in Figure 2.1b), Seeing3DChairs [7] and Shoes [193], in terms of the following metrics: (1) Fréchet inception distance (FID) [58] for visual quality. FID measures how close the generated data distribution is to the real data distribution by comparing their feature representations extracted from a pretrained network: lower FID scores indicate more realistic and higher-quality generations. FID has served as the golden standard metric in generative modeling since it exists. We defer the detailed introduction of FID in Chapter 3; (2) plausible design rate (PDR), the proportion of plausible designs for 1 000 evaluated images (see Section 2.3.2 for details); and (3) frames per second (FPS) for generation speed. On the BIKED data, our PoDM outperforms EDM on PDR: 93.5% (PoDM) vs. 83.4% (EDM) and on FID: 4.87 (PoDM) vs. 7.84 (EDM), while achieves comparable FPS with EDM. Compared to DDPM, PoDM has a comparable PDR value (PDR: 94%, FID: 11.77) but is ca. 15× faster in terms of FPS.

Lastly, we further test the performance of PoDM in incorporating modern image-editing methods, e.g., inpainting, interpolation via latent space and point-based dragging, and hereby manipulating structure.

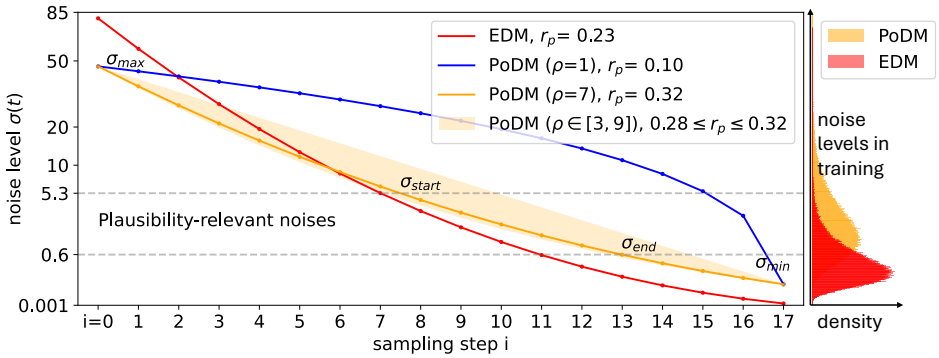
2.2 Plausibility-oriented Diffusion Modeling

2.2.1 Background

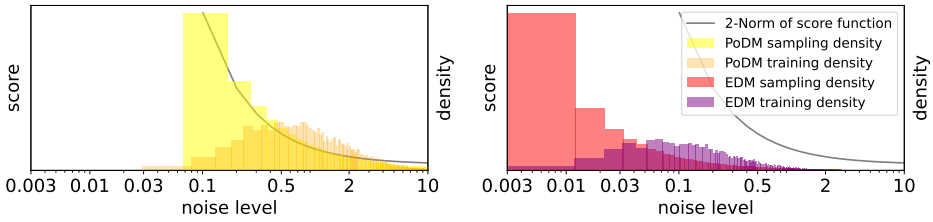
We built up our contribution based on the stochastic differential equation (SDE) model of the diffusion process [159, 71]. Given a data point $\boldsymbol{x} \in \mathbb{R}^d$, we corrupt it with the following forward Itô SDE [116]:

$$d\boldsymbol{x} = f(\boldsymbol{x}, t)dt + g(t)dB_t, \tag{2.1}$$

where $f\colon \mathbb{R}^d \times [0, T] \to \mathbb{R}^d$ is the drift vector, $g\colon [0, T] \to \mathbb{R}$ is the dispersion coefficient, and $B_t \in \mathbb{R}^d$ is the standard Brownian motion. Notably, f and g, are



(a) Plausibility-relevant noise range and noise schedules



(b) Magnitude of the score function and density of noise schedules

Figure 2.3: (a) We showcase the plausibility-relevant noise range $[\sigma_{\text{end}}, \sigma_{\text{start}}]$ (*dashed interval*) computed on the BIKED dataset with the techniques proposed in Section 2.2.2. For the denoising process, our PoDM method takes an exponentially decaying schedule (*blue and orange curves*) with a parameter ρ controlling the decay rate. PoDM method determines the minimal/maximal noise levels of the schedule (σ_{min} and σ_{max}, respectively) based on $[\sigma_{\text{end}}, \sigma_{\text{start}}]$ (Equation (2.8)). As for the noise levels sampled in the training process, our PoDM method uses a log-normal density concentrating on $[\sigma_{\text{end}}, \sigma_{\text{start}}]$ (*the orange histogram shown vertically*), in contrast to the fixed distribution used in EDM (*the red histogram*). (b) For PoDM and EDM, we depict the magnitude of the score function (Equation (2.2)) at different noise levels, which is compared to the density of the noise levels in the training and denoising steps. The noise density in the denoising process of PoDM matches closely with the score function.

pre-determined by the user and have no trainable parameters. The corresponding reverse-time/backward SDE is [4]:

$$d\boldsymbol{x} = [f(\boldsymbol{x}, \tau) - g(\tau)^2 \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}, \tau)]d\tau + g(\tau)d\mathbf{B}_\tau, \tag{2.2}$$

where τ goes from T to 0, $p(\boldsymbol{x}, \tau)$ is the probability density of $\boldsymbol{x}$ at τ in the forward process, and $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}, \tau)$ is known as the score function. The flow of probability mass in Equation (2.2) can be equivalently described by an ordinary differential equation (ODE) [100, 159, 71]:

$$\frac{d\boldsymbol{x}}{d\tau} = f(\boldsymbol{x}, \tau) - g(\tau)^2 \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}, \tau). \tag{2.3}$$

We follow the choice of the drift and dispersion terms in [71] (a.k.a. EDM):

$$f(\boldsymbol{x}, \tau) = 0, \quad g(\tau) = \sqrt{2\sigma(\tau)}, \quad \sigma(\tau) = \tau,$$

and the score function is approximated by $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}, \tau) = (D_\theta(\boldsymbol{x}; \sigma(\tau)) - \boldsymbol{x})/\sigma(\tau)^2$, where D_θ is a neural network trained on samples drawn from the forward SDE (see [71] for details on the loss function). Due to the above choice, σ and τ are interchangeable henceforth. To solve/sample from Equation (2.3), an N time-step discretization is used with the following noise schedule: $\sigma_N = 0, \forall i \in [0..N-1]$:

$$\sigma_i = \left(\sigma_{\max}^{\frac{1}{\rho}} + \frac{i}{N-1}(\sigma_{\min}^{\frac{1}{\rho}} - \sigma_{\max}^{\frac{1}{\rho}})\right)^\rho, \tag{2.4}$$

where $\sigma_0 = \sigma_{\max}$ and $\sigma_{N-1} = \sigma_{\min}$. EMD recommends the setting: $\sigma_{\min} = 0.002, \sigma_{\max} = 80, \rho = 7$. The exponent ρ affects how much the steps near $\sigma_{\min}$ are shortened at the cost of longer step length near $\sigma_{\max}$. In the stochastic sampling procedure, we denote by $\boldsymbol{x}_i$ the data point obtained at σ_i. We first increase the noise level slightly and perturb $\boldsymbol{x}_i$:

$$\boldsymbol{x}_i' = \boldsymbol{x}_i + \sqrt{\hat{\sigma}_i^2 - \sigma_i^2}\mathcal{N}(\mathbf{0}, S_{\text{noise}}^2\mathbf{I}), \tag{2.5}$$

$$\hat{\sigma}_i = \sigma_i \left(1 + \mathbb{1}_{[S_{\min},\, S_{\max}]}(\sigma_i) \min\left(S_{\text{churn}}/N, \sqrt{2} - 1\right)\right), \tag{2.6}$$

where S_{churn} controls the degree of randomness in sampling: $S_{\text{churn}} = 0$ realizes deterministic generation. Afterwards, we apply the reverse-time ODE (Equation (2.3)) with $\boldsymbol{x}_i'$ from $\hat{\sigma}_i$ to σ_{i+1}. The default settings of stochastic sampling are: $S_{\text{churn}} = 40$, $S_{\min} = 0.05$, $S_{\min} = 50$, $S_{\text{noise}} = 1.003$. The training data of $D_\theta(\boldsymbol{x}; \sigma(t))$ are

Table 2.1: Grid search results of the parameter ρ of the noise schedule on the BIKED data w.r.t. three performance metrics. The *column* r_p measures the proportion of noise levels falling into the plausibility-relevant range, which is strongly correlated with the performance metrics.

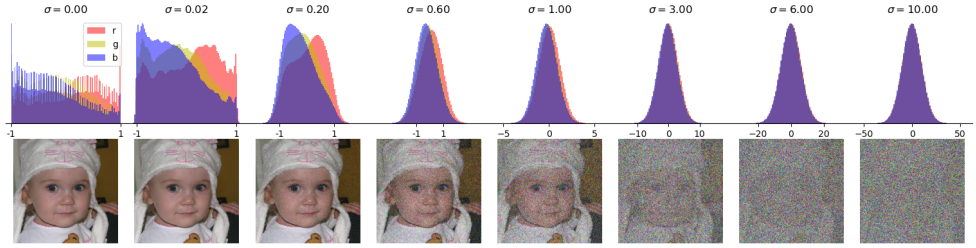
ρ	r_p	FID↓	DPS↑	PDR↑
1	0.10	12.67	4.70	82.8%
3	0.28	5.64	4.81	88.6%
5	0.31	5.25	4.88	89.6%
7	0.32	**4.87**	**4.90**	**93.5%**
9	0.32	5.18	4.87	91.5%

sampled from Equation (2.1) with a log-normal distribution: $\ln(\sigma) \sim \mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2)$. In [71], the following empirical setting is suggested: $P_{\text{mean}} = -1.2, P_{\text{std}} = 1.2$.

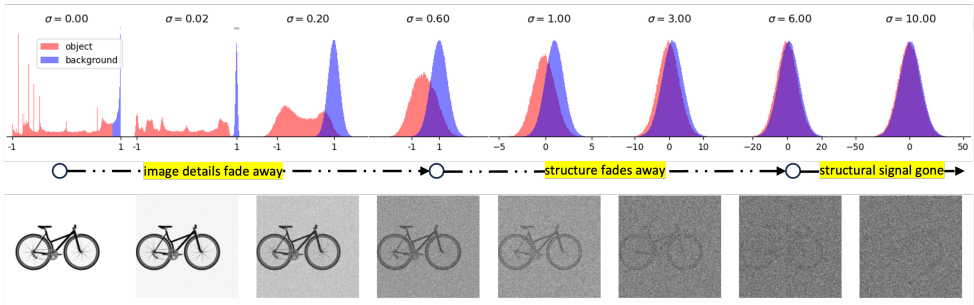
2.2.2 Noise Range Relevant to Plausibility

We conjecture that in the forward/backward diffusion process, there exist a range of noises that plays a crucial role in object construction. We validate it by running the forward process with fine-grained noise levels: starting with a source image $\boldsymbol{x}(0)$ (the pixel values are standardized to $[-1, 1]$ before adding the Gaussian noise), we keep adding small Gaussian noises thereto until its pixel-value distribution becomes indistinguishable from a Gaussian: $\boldsymbol{x}(t) = \boldsymbol{x}(0) + 0.1t \times \mathcal{N}(\mathbf{0}, \mathbf{I})$. In Figure 2.4b, we show the pixel-value distribution at intermediate time steps computed from 100 images sampled from BIKED [137]. The BIKED images consist of a single bicycle object and a monotone background. We depict the pixel-value distribution separately for the object and the background (the red and blue histograms, respectively). We observe three phases: (1) from the beginning to the first time when the pixel-value distribution of the bicycle converges to a Gaussian, i.e., $\sigma \in [0, 0.6]$. The pixel-value distribution of the bicycle is substantially different from that of the background in this phase; (2) the bicycle structure starts to fade away while the pixel-value distribution thereof overlaps more with the background, i.e., for $\sigma \in [0.6, 6.0]$; (3) the bicycle structure almost disappears for $\sigma > 6$. Empirically, we assume that the noise range ($\sigma \in [0.6, 6]$ in Figure 2.4) in which the bicycle structure fades away determines the plausibility of the generated structural design. We denote this noise range as $[\sigma_{\text{end}}, \sigma_{\text{start}}]$ and propose two techniques to determine this interval.

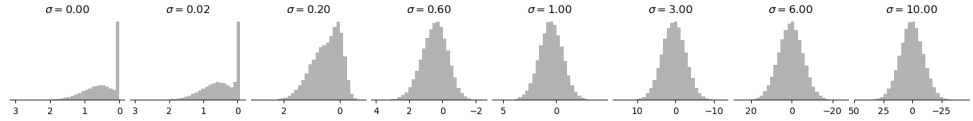
Technique 1 *Choose σ_{end} to be the largest noise level at which the object pixel values are not normally distributed according to the Shapiro-Wilk test with a significance level*



(a) FFHQ [72] (resolution: 64×64)

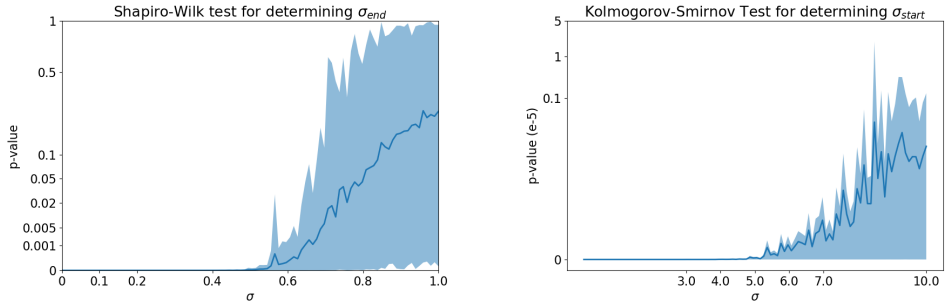


(b) BIKED [137] (resolution: 256×256)



(c) Distribution of latent features extracted from the BIKED data [137] with a convolutional autoencoder.

Figure 2.4: Evolution of the pixel-value distribution in perturbation experiments for (a) FFHQ, a real-world data set, and (b) the BIKED data, a set of design structures. In (c), we show the evolution of the latent-value distribution after encoding the BIKED data into a 64-dimensional space with a convolutional autoencoder.



(a) Shapiro-Wilk test to determine σ_{end} **(b)** KS test to determine σ_{start}

Figure 2.5: On the BIKED data, (a) we illustrate **Technique 1** by plotting the p-value of the Shapiro-Wilk test computed from 100 randomly picked images, which are perturbed by different noise levels; (b) for **Technique 2**, we show the Kolmogorov-Smirnov (KS) test between the bicycle's and background's pixel-value distribution as a function of the noise levels. In both plots, the variability is depicted as the *shaded region*.

of 0.01.

In a sampling process, σ_{end} is the noise level at which the generation is structurally finalized and object pixel values remain normally distributed. Denoising with noise levels of $\sigma \leqslant \sigma_{\text{end}}$ performs a refinement, during which the backward process approximates the object pixel values from a normal distribution to a local data distribution. To estimate σ_{end}, we propose to use the Shapiro-Wilk test [149] to track the distribution of the object's pixel values during the perturbation test. As displayed in Figure 2.5a, the measured p-value increases with the noise level. In practice, specific dataset might exist, where the object pixel-value distribution is already Gaussian distributed, here we set a minimum limitation of σ_{end} to be 0.08 (converting the default parameter values of EDM to our parameter system, we obtain the value $\sigma_{\text{end}} = 0.08$).

Technique 2: *Choose σ_{start} to be the noise level at which pixel-value distributions of object and background are sufficiently close, with the p-value of a Kolmogorov-Smirnov test begins to diverge.*

In the synthesis of structural design images, σ_{start} is the noise level at which the structural formation begins, i.e., pixels of objects begin to distinguish themselves from pixels of the background. To measure such a difference, we first approximate the pixel-value distributions of the object and background with Gaussians, respectively, and then conduct the Kolmogorov-Smirnov test between the two Gaussian approximations. As shown in Figure 2.5b, σ_{start} is taken when the curve of p-value measured in the

Kolmogorov-Smirnov test begins to diverge.

Our work gives an insight into defining the plausibility-relevant range of noise levels so that the training and sampling effort can prioritize this range. By implementing our two techniques, for BIKED images, we determine σ_{end} to be 0.6 and σ_{start} to be 5.3. This observation can also be seen in real-world images, such as those from FFHQ [72]. Since it is difficult to automatically separate faces from backgrounds, we first track the distribution of pixel values in each RGB channel without distinguish pixels of faces and of backgrounds: as seen in Figure 2.4a, the distribution of pixel values in each color channel starts from an irregular distribution, gradually converges to a Gaussian distribution (at about $\sigma = 0.60$), and then overlaps each other at a certain noise scale (at about $\sigma = 0.60$). To showcase that the above observation can also be seen in the latent diffusion models [139], we first encode 100 BIKED images with a convolutional autoencoder trained on BIKED into a 64-dimensional latent space and then show the evolution of the latent-value distribution (see Figure 2.4). We observe a pretty similar trend in the latent-value distribution as with the pixel-value distribution. Note that the plausibility of design images, i.e., BIKED [137] and Seeing3DChairs [7], is easier to assess, the background can be automatically separated from the main object, and therefore the plausibility-relevant noise range can be more accurately estimated. Thus, this work will focus on structural designs to demonstrate the efficiency of our proposal.

Analysis on more datasets In Figure 2.6 and Figure 2.7, we additionally plot the pixel-value distribution during the perturbation experiment on datasets, e.g., the Seeing3DChairs [7] and Shoes [193].

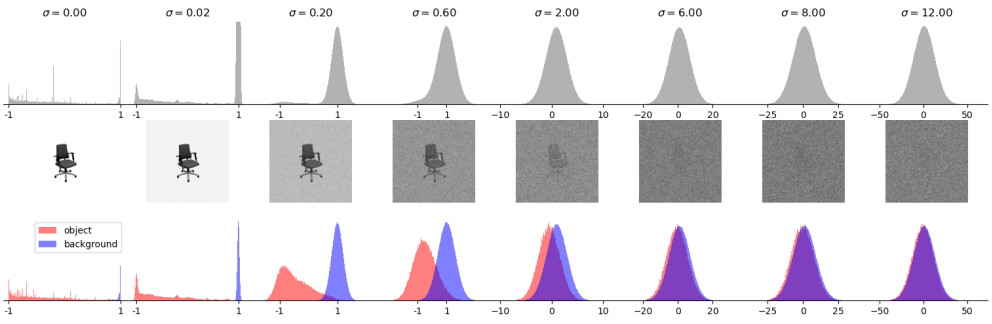


Figure 2.6: Evolution of the pixel-value distribution in perturbation experiments for Chair designs. In the top row, all pixel values are plotted as histograms in gray, while in the next row, the pixels are divided into background pixels and target pixels and then plotted as histograms.

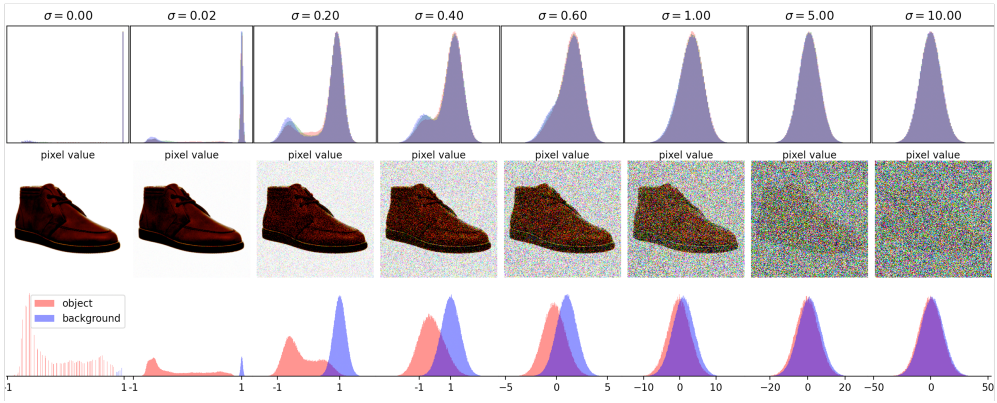


Figure 2.7: Evolution of the pixel-value distribution in perturbation experiments for Shoes designs. In the top row, we plot the pixel values of the three R-G-B channels as the corresponding colors; while in the bottom row, we disregard the color channels and divide them into pixel values (background) and pixel values (object). This plot shows that design images with color channels still follow our observation.

With color images in the FFHQ [72] data set, we track the pixel-value distribution for each color channel with object (human face) and the background separated in two distributions. In Figure 2.8, we plot the evolution of the pixel-value distribution in perturbation experiments. It can be clearly seen that the evolutionary process of FFHQ images is the same as the evolutionary process of the BIKED images, where the three phases can be observed.

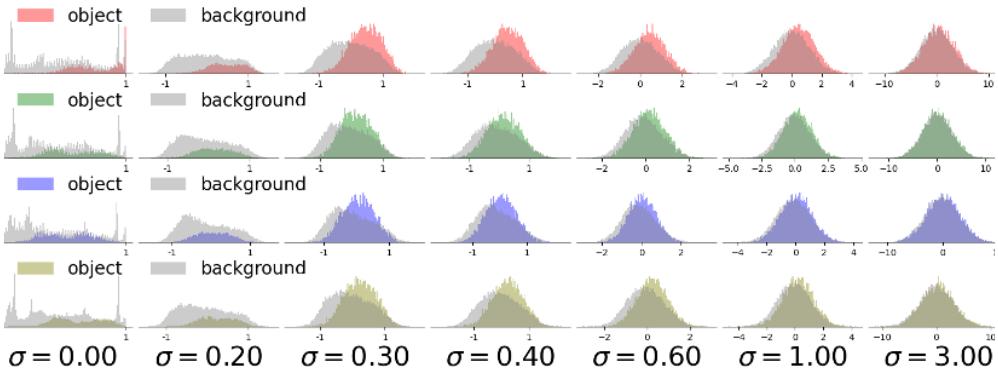


Figure 2.8: Evolution of the pixel-value distribution in perturbation experiments for FFHQ [72] (resolution: 64×64). Here, we manually separate the face from the background pixels. Rows from up to down: the R, G, B color channels and the grayscale.

2.2.3 Training and Generation Procedures

We modify the training and generation procedures so that our diffusion model can concentrate on the plausibility-relevant range of noise levels.

Noise density in training For the structure images, the generation of the structure takes place mostly in the noise range $[\sigma_{\text{end}}, \sigma_{\text{start}}]$ while the noise levels that are too small or large have marginal effects on the plausibility of the final outcome. Hence, it is sensible to sample more noise levels in this interval from the forward SDE. We propose the following log-normal distribution to sample noise levels:

$$\ln(\sigma) \sim \mathcal{N}(\mu,\, \zeta^2),\ \mu = \frac{\ln(\sigma_{\text{start}}) + \ln(\sigma_{\text{end}})}{2},\ \zeta = \frac{\ln(\sigma_{\text{start}}) - \ln(\sigma_{\text{end}})}{2}, \tag{2.7}$$

which implies $\Pr(\sigma \in [\sigma_{\text{end}}, \sigma_{\text{start}}]) \approx 68\%$. In this method, the majority of the noise levels are drawn in $[\sigma_{\text{end}}, \sigma_{\text{start}}]$ while we have ca. 32% probability to sample noise levels at the beginning and the end of the forward process.

Noise schedule for image generation In the backward diffusion process (Equation (2.3)), there are two important factors w.r.t. the noise levels: (1) the noise range $[\sigma_{\min}, \sigma_{\max}]$ in which we apply the ODE (Equation (2.3)) and (2) the decaying noise schedule. For the former, we determine the range based on the training noise density as follows: Equation (2.7) implies that the score function $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}, \sigma)$ is trained on noise levels drawn almost in $[\mu-3\zeta, \mu+3\zeta]$, i.e., $\Pr(\log(\sigma) \in [\mu-3\zeta, \mu+3\zeta]) \approx 99.7\%$. Therefore, when sampling new images, applying the reverse-time ODE out of $[\mu-3\zeta, \mu+3\zeta]$ requires the score function to extrapolate, which we have no guarantee about its accuracy. Hence, we set

$$\log \sigma_{\min} = \mu - 3\zeta = 2\log \sigma_{\text{end}} - \log \sigma_{\text{start}}\ , \tag{2.8}$$

$$\log \sigma_{\max} = \mu + 3\zeta = 2\log \sigma_{\text{start}} - \log \sigma_{\text{end}}\ . \tag{2.9}$$

For the latter, we follow the exponential decay in Equation (2.4), where, in addition, we tune the hyperparameter ρ for the BIKED dataset. In Table 2.1, we summarize the tuning results from a simple grid search, where $\rho = 7$ is the best setting. Also, we observe that the performance metrics (e.g., FID, DPS, and PDR) are quite sensitive to ρ, suggesting that tuning this parameter is necessary across different structural image data. Moreover, we calculate the proportion of the noise levels $\{\sigma_{N-1}, \ldots, \sigma_0\}$ (determined by Equation (2.4)) falling into the plausibility-relevant range $[\sigma_{\text{end}}, \sigma_{\text{start}}]$, which we call the prioritization density r_p. It measures how much training effort is

targeted at the structure modeling. In Figure 2.3 and Table 2.1, we show r_p with varying hyperparameter ρ, and we observe that the performance metrics are positively related to it.

We demonstrate a theoretical insight into the noise schedule in Figure 2.3b. On the BIKED data, we depict the norm of score function $\nabla_{\vec{x}} \log p(\vec{x}, \sigma(t))$ over noise levels. Comparing it with the histograms of the noise schedule, we see that PoDM's noise scheduling in sampling/denoising is concordant with the score function, meaning that finer steps of simulating the backward ODE/SDE (see Equation (2.3)) are taken where the norm of the score function is large. We argue that it is sensible to do so since the score function is the major drift term of the backward ODE/SDE.

2.3 Evaluation and Results

In this section, we compare PoDM with several cutting-edge models: the foundational Denoising Diffusion Probabilistic Models (DDPM) by Ho et al. [61], the faster-sampling variant Denoising Diffusion Implicit Models (DDIM) by Song et al. [156], the highly-tuned design-space model Elucidating the Design Space of Diffusion-Based Generative Models (EDM) by Karras et al. [71], and also the non-diffusion generative model Adversarial Latent Autoencoder with Self-Attention for Structural Image Synthesis (SA-ALAE) [165]. We selected these models because they together provide a broad spectrum of generative modeling approaches, sampling schemes, and architectural innovations: DDPM serves as the basic diffusion baseline; DDIM introduces a practical improvement in sampling efficiency; EDM represents the state of the art in diffusion-model design choices (including noise scheduling, preconditioning and sampling strategies); and SA-ALAE offers a useful structural-design-specific comparison beyond diffusion models, targeting complex engineering images.

By benchmarking PoDM against these four models, we are able to situate our contributions with respect to both the canonical diffusion approach (DDPM) and more advanced variants (DDIM, EDM) that already address sampling speed, noise schedule or other design choices, as well as against a structurally-focused adversarial model (SA-ALAE) from the engineering-design domain. Such a comparison allows us to evaluate how much PoDM's novel noise-scheduling strategy adds (i) over the vanilla diffusion chain, (ii) relative to diffusion models that already optimize schedule or sampling protocol, and (iii) in the specific context of plausible structural image synthesis typified by engineering blueprints rather than natural images.

2.3.1 Training Configurations

Our work utilizes the model architecture from the DDIM [156] repository for all diffusion-based models, which follows the U-Net proposed by Ho et al. [61]. More precisely, the implemented model has six feature map resolutions from 256×256 to 4×4, one residual block for each upsampling/downsampling, and an attention layer at the feature map resolution of 16×16. For sampling with DDPM and DDIM, we use the same trained model with default training settings, i.e., timesteps of 1 000 and linear schedule of β with $\beta_0 = 10^{-4}, \beta_T = 0.02$. For EDM, we remove EDM's preconditions, since they did not bring much enhancement to the results according to their experiments, and implement their noise schedules for both training and sampling with default parameters, i.e., $\sigma_{\text{min}} = 2 \times 10^{-3}, \sigma_{\text{max}} = 80, \rho = 7, P_{\text{mean}} = -1.2, P_{\text{std}} = 1.2$. For our PoDM, we determine $\sigma_{\text{start}} = 5.3$ and $\sigma_{\text{end}} = 0.6$ by analyzing the BIKED dataset and inheriting the loss function from EDM. For stochastic sampling in both EDM and our PoDM, we allow the "churn" modification (Equation (2.5)) for all sampling steps, i.e., $S_{\text{min}} = 0$, $S_{\text{max}} = \infty$, and set the S_{churn} to 5. For DDIM, we use 50 as the number of sampling steps, whereas for both EDM and our PoDM, the number of sampling steps is set to 18. The set "Standardized Images" from BIKED Dataset [137] contains 4 512 grayscale pixel-based images with original shape of $1\,536 \times 710$. We pad them with background pixels to a square form with the shape of $1\,536 \times 1\,536$. Then, we reshape these images into a resolution of 256×256 with the scale of $[-1, 1]$ in order to maintain the height-width ratio and ease the complexity in generation. From the whole dataset, we randomly select 100 images for validation, 1 000 images for testing, and the rest of the images for training. We run training on four NVIDIA DGX-2's Tesla V100 GPUs with a batch size of 32 and a learning rate of 5×10^{-5}. Model parameters are saved every 1 000 steps. If the loss converges, we keep training until 100 000 steps and then stop it when the denoising loss does not decrease for 20 epochs. For each model, we select the best-performing model within the saved checkpoints in the last 20 000 steps.

2.3.2 Results

In our work, we evaluate the generative models regarding sampling speed, visual quality, and design plausibility. For the sampling speed, we simply record the sampling time for generating 5 000 images and calculate the sampling speed in FPS (frames per second) for each model. For visual quality, we further use the 5 000 images generated and calculate the FID [58] between the test images and the generated images. The

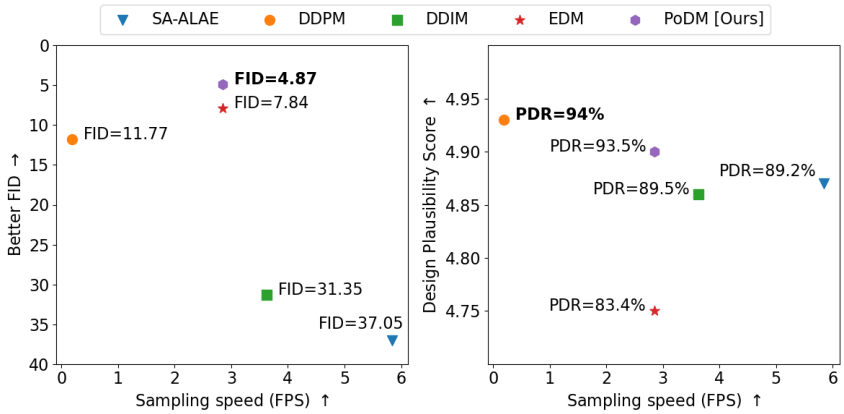


Figure 2.9: On the BIKED data, we show two performance views of the five considered generative models: FID vs. FPS and design plausibility score (DPS) vs. FPS.

measured FIDs are displayed in Figure 2.9.

To quantitatively evaluate the plausibility of generated designs, we implement a human evaluation method, in which the human evaluator bypasses visual qualities (e.g., blurriness and background noise) and scores the represented design in terms of plausibility. We refer to the evaluation score as the design plausibility score (DPS). In this work, the generated bicycle designs are evaluated using a five-point scoring system based on the following criteria:

- No missing fundamental part;
- No floating material or extra part;
- Every part is complete;
- Parts are connected;
- Rational positioning.

For generative model considered, we randomly select 1 000 samples from the 5 000 generated bicycle images. We shuffle all selected images and keep tracking their DPS in a manner that associates each image's score with its corresponding model. This experiment aims to prevent potential biases in the evaluation of the generated images by individual target models and to sustain a uniform evaluation standard across all images. We record the measured DPSs in Figure 2.10 and an average DPS for each model in Figure 2.9. Besides, we calculate the (PDR), which is the proportion of plausible designs, i.e., designs with DPS of 5, in 1 000 generated images.

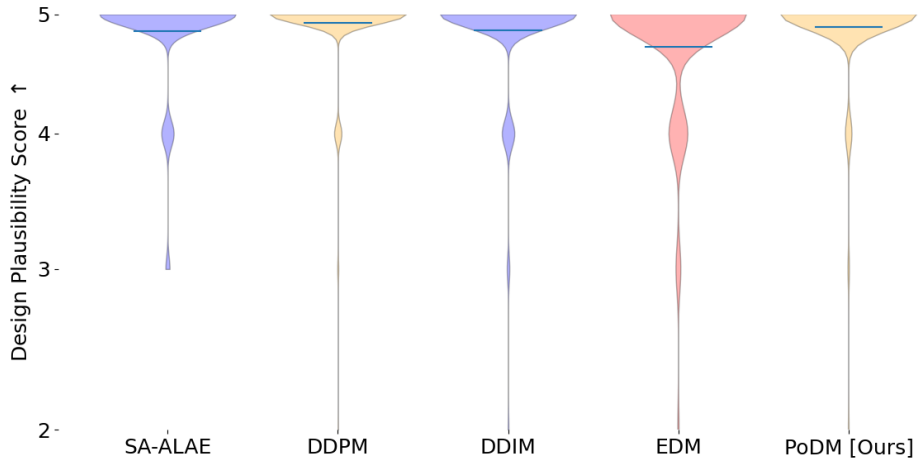


Figure 2.10: On the BIKED data, we show the detailed empirical distribution of the design plausibility scores measured for each model.

On the BIKED data, we first show the overall performance values in Figure 2.9: our PoDM achieves a compelling FID of **4.87**, a satisfactory DPS of **4.90**, and a high plausible design rate of **93.5%**. DDPM [61] requires the longest sampling time of 5.26 seconds for each image but performs decently well in terms of image quality, i.e., FID of 11.77, and design plausibility, i.e., DPS of 4.93 with only 6.0% implausible outcomes. As shown in Figure 2.9, EDM [71] can significantly improve the sampling speed to 2.85 FPS and even enhance the visual quality to a FID of 7.84. However, EDM performs poorly in design plausibility compared to PoDM, i.e., DPS of 4.75 and a plausible design rate of 83.4%. As seen, DDIM and EDM demonstrate a trade-off between visual quality and plausibility of generated images, whereas the DDPM leverages extremely slow sampling speed to perform decently in both aspects. We need to address that although it seems that DDPM's DPS value (= 4.93) is slightly higher than PoDM's (= 4.90), there is actually no statistical difference between them (based on a Mann–Whitney U test). Hence, we state that our PoDM method can achieve the same design plausibility, a better FID, and a much faster generation/sampling speed than DDPM.

In Figure 2.16, Figure 2.17, Figure 2.18, Figure 2.19, Figure 2.20, we plot a certain number of randomly generated bicycle designs for each model trained on the BIKED dataset, respectively. Here, we provide the results for the qualitative evaluation.

We additionally train PoDM and EDM on Seeing3Dchair [7] images of resolution

(a) Chairs generated by EDM with scores: DPS-4.13, DPR-51.5%, FID-33.48

(b) Chairs generated by our **PoDM** with scores: DPS-**4.65**, DPR-**74.5%**, FID-32.15

Figure 2.11: Generated chair designs with limited training epochs

(128×128) with only 100 epochs. After training, we showcase the generated designs for a visual comparison in Figure 2.11. Visually, PoDM generates much more plausible chairs than EDM. Measured on 1k generated images, the mean design plausibility score (DPS) and design plausibility rate (DPR: DPS=5) are: PoDM (DPS 4.65, DPR 74.5%, FID 32.15); EDM (DPS 4.13, DPR 51.5%, FID 33.48). Overall, we state that our method can *significantly improve the speed of generating structural designs while maintaining their plausibility.*

2.3.3 Alignment Test

It is not surprising that the most-used automatic metrics in the generative modeling community do not align well with human judgments and fail to capture the plausibility of generated designs. To demonstrate this, we visualize the correlation between human evaluation results and metrics results in Figure 2.12.

In our procedure we collect human plausibility ratings for generated designs and compare them with automatic metric scores: Fréchet Inception Distance (FID), which measures the distance between the distribution of generated images and the distribution of reference real images based on deep feature statistics; the Structural Similarity Index (SSIM), which assesses luminance, contrast and structural similarity between two images; and the Learned Perceptual Image Patch Similarity (LPIPS), which computes deep-feature distances learned to match human perceptual judgments. What we observe is only a weak or moderate alignment: designs judged more plausible by humans do not consistently receive better metric scores. This highlights a risk of relying solely on these metrics when structural plausibility is the focus of evaluation.

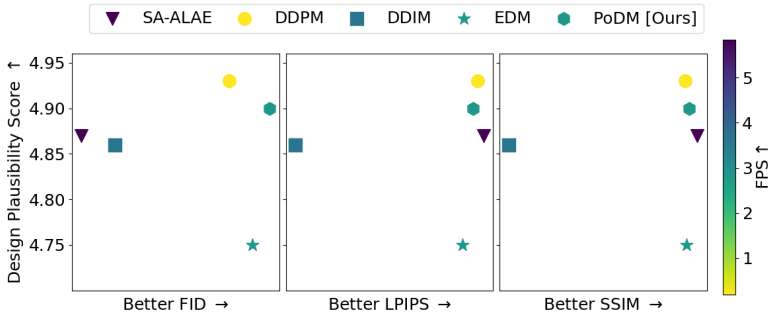
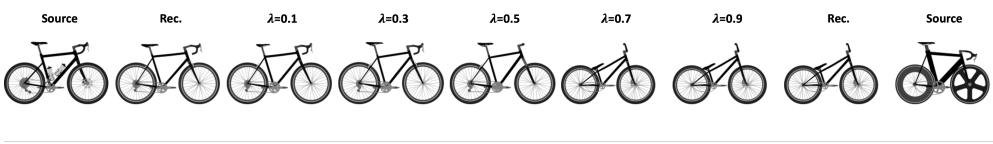


Figure 2.12: Alignment test. Here we test the alignment between human evaluation results (design plausibility score) and various metric results. The plot shows that non of them have a strong correlation to the human evaluation.

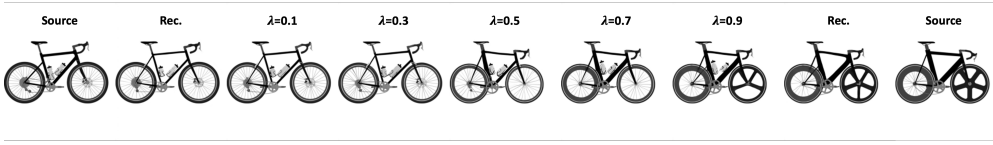
2.4 Controllable Generation and Design Editing

In this section, we test the PoDM's understanding of structural design space by applying cutting-edge image editing methods, e.g., interpolation via latent space, point-based dragging and inpainting, on bicycle designs.

Interpolation via latent space Interpolation via latent space can be quite useful



(a) Interpolation with latent space set at noise level $\sigma_{\max}$



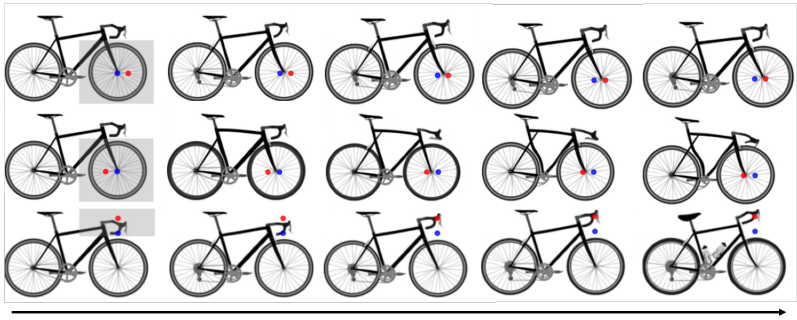
(b) Interpolation with latent space set at noise level σ_{start}

Figure 2.13: PoDM-driven structural interpolation via latent space set at various perturbation steps. In (a), the reconstruction has a poor accuracy, and interpolation fails to produce intermediate structures. In (b), the interpolation displays a smooth transformation between two source structures.

in exploring structural design space. After encoding a source data $\boldsymbol{x}(0)$ to pure noise $\boldsymbol{x}(T)$ via the forward process, the diffusion model is supposed to decode $\boldsymbol{x}(T)$ back to $\boldsymbol{x}(0)$ by utilizing a corresponding ODE [159]. However, in our implementation shown in Figure 2.13a, PoDM-motivated reconstruction has poor accuracy, which might be caused by the prioritizing strategy. We argue that it is unnecessary to conduct the forward process completely, instead, perturbed images at noise level σ_{start} retain good reversibility. Taking images at noise level σ_{start} as latent code allows well-performing reconstruction and interpolation, as shown in Figure 2.13b.

Point-based dragging As a novel image editing method, point-based dragging [122, 150] can precisely and iteratively "drag" the handle point to a target point and the remaining parts of the image will be correspondingly updated to maintain the realism. We implement DragDiffussion [150] on BIKED images and plot the results in Figure 2.14a. To the best of our knowledge, our work is the first to apply point-based dragging on structural design.

Extending the target design geometry is common in the day-to-day work of engineering, but it is still very time-consuming because engineers need to manually modify each related part to edit the geometry of the final design. Meanwhile, our work has shown that after the DGM is trained with historical design data and the geometric dependencies among parts are captured accordingly, the DGM can perform this dragging task. In Figure 2.15a, we showcase the advantages of using DGMs in geometric editing tasks. Inspired by this, we assume that with the same optimization pipeline,



(a) PoDM-driven structure editing with DragDiffusion [150]

(b) PoDM-driven inpainting with RePaint [97]

Figure 2.14: PoDM-driven structural interpolation via latent space set at various perturbation steps. In (a), from left to right, the handle point is iteratively dragged from the initial handle point (*blue*) towards the selected target point (*red*). In (b), the part enclosed by a *dashed line* is redesigned with RePaint.

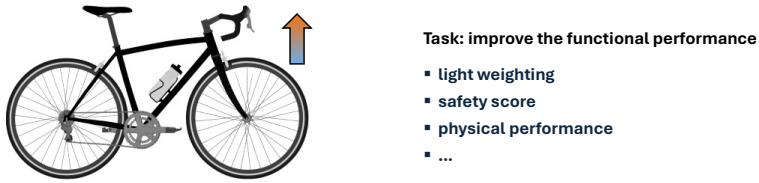
DGMs can also modify the target design to achieve better functional performances, which will be detailed described in Section 6.1.

Inpainting In an inpainting task, the generative model is tasked to generate the inpainting area to match the known part. A DDPM-based inpainting mechanism, RePaint [97], has achieved the state-of-the-art performance on diffusion-based inpainting tasks by utilizing the known part as guidance at each step. We adapt RePaint to our PoDM and test it on BIKED images. The inpainting results are shown in Figure 2.14b.

2.5 Conclusion

When generating engineering designs with DGM, the primary task is to ensure that the generated design is reasonable, whereas the current DGM cannot achieve this well. We observe that the performance of the diffusion-based generative models exhibits a trade-

(a) Intuitive design editing task



(b) Function-oriented design editing task

Figure 2.15: Scenarios of engineering day-to-day operation tasks. Inspired by the geometric operation task shown in (a), we assume that with the same optimization pipeline DGMs can also edit the design to achieve a better functional performance.

off among visual quality, the plausibility of generated images, and sampling time. We assume that there is a range of noise levels, that is responsible for the plausibility of the outcome, especially in generating structures. Following this observation, we propose a plausible-oriented diffusion model (PoDM) that leverages a novel noise schedule to prioritize this range of noise levels in both training and sampling procedures. We observe that the well-known EDM has a poor performance in generating plausible structures. Our PoDM method significantly improves the plausibility of generated images over EDM and also achieves a satisfactory plausibility score comparable to DDPM but with a much-reduced generation time. Additionally, we demonstrate with convincing results that the improvement in the plausibility thanks to the prioritization of the determined noise range. Further implementations of PoDM-driven image editing tools showcase PoDM's ability to semantically manipulate complex structural designs, paving the way for future work in the field of generative design.

This chapter is inspired by, but not limited to, engineering design generation. We believe that our observations and determinations of the phases in the diffusion process are equally applicable to images from natural scenes and, therefore, beneficial for all diffusion-based synthesis tasks. In addition, we hope that our work will inspire more research on the tool for automatically evaluating the plausibility of generated images and the relevance between noise level and generated features.

Limitation The designs generated in this chapter are assessed through both qualitative and quantitative evaluations. However, we have found that the existing metrics do not align well with human judges' assessments in terms of plausibility, as detailed in Section 2.3.3. This misalignment raises concerns about the reliability of the current evaluation methods. To address this issue, we have developed a design plausibility score, which is manually evaluated as an assessment metric. Unfortunately, this process is not automated, making it time-consuming and less efficient. Therefore, it is crucial and urgent to create a novel metric that can effectively evaluate plausibility while aligning with human evaluation standards. Achieving this goal will enhance the accuracy and relevance of our assessments. In the next chapter, we will focus on developing such a metric for result plausibility, exploring innovative approaches that can bridge the gap between automated evaluations and human judgment.

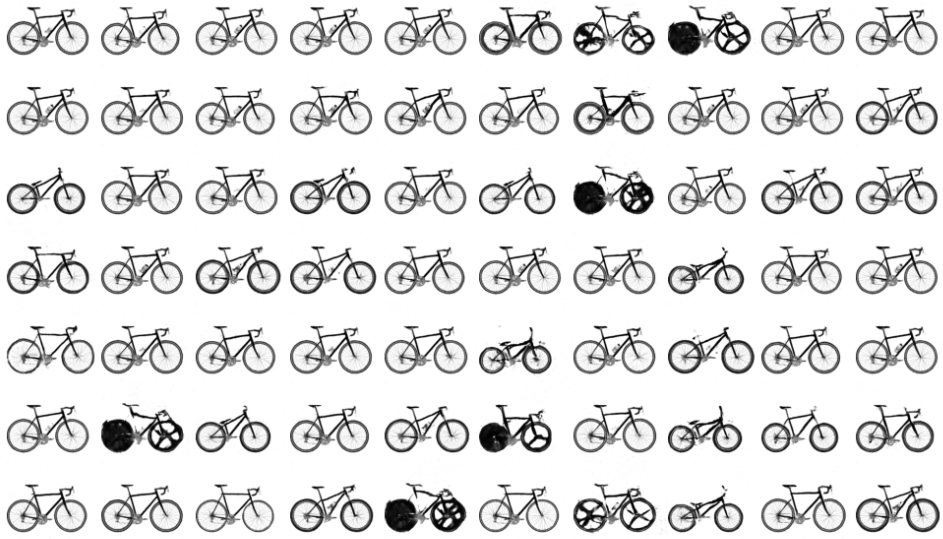


Figure 2.16: BIKED images randomly generated by SA-ALAE [165]. SA-ALAE shows an uneven performance over various classes bikes and a great portion of generated designs are of poor quality and present implausible structures.

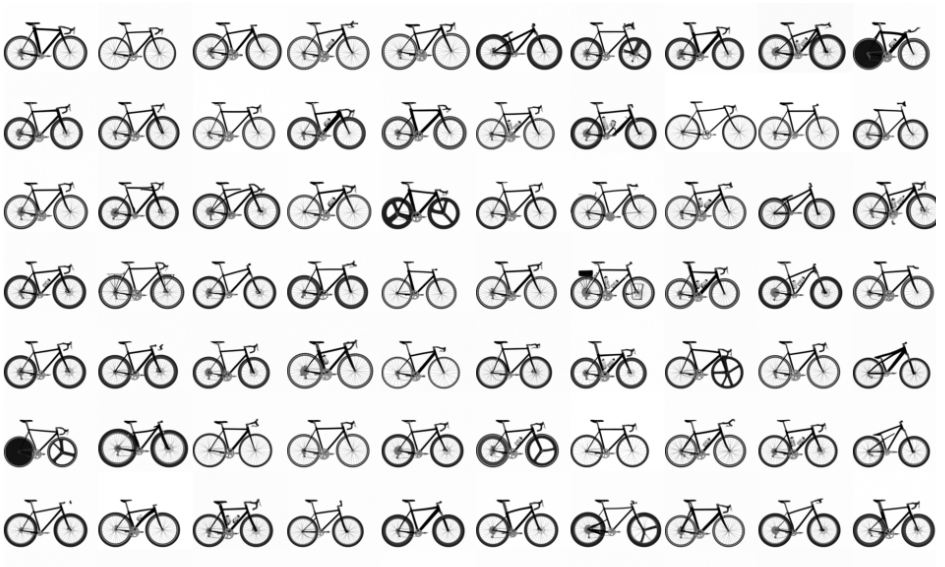


Figure 2.17: BIKED images randomly generated by DDPM [61]. DDPM presents a strong generative power in both visual quality and structural plausibility. However, DDPM requires always a tremendous number of denoising steps (i.e., 1 000), otherwise the results look like in Figure 2.18.

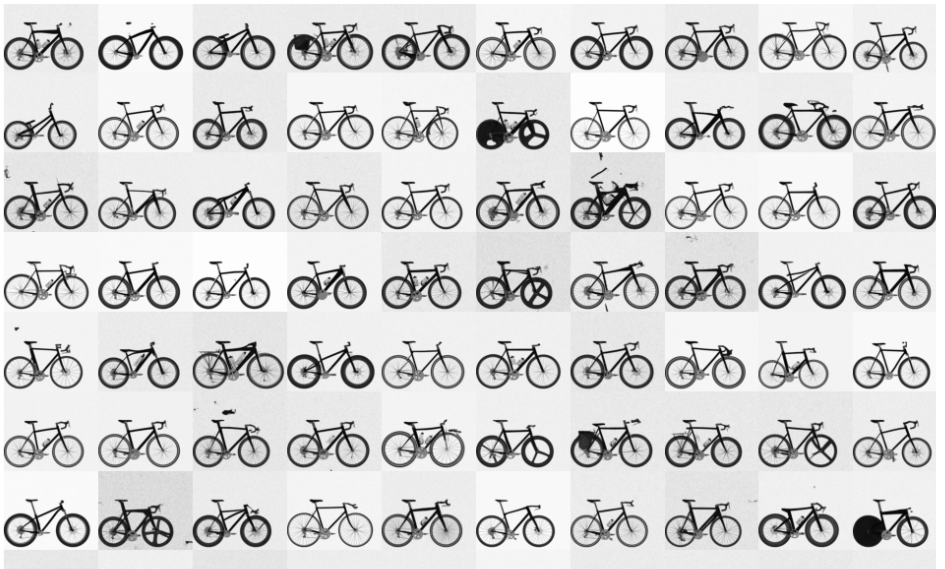


Figure 2.18: BIKED images randomly generated by DDIM [156]. DDIM leverages the same trained backbone model as DDPM, but attempts to break the Markov-chain of DDPM and to use a reduced number of denoising steps (i.e., 50). Hereby, the generated images present poor visual quality.

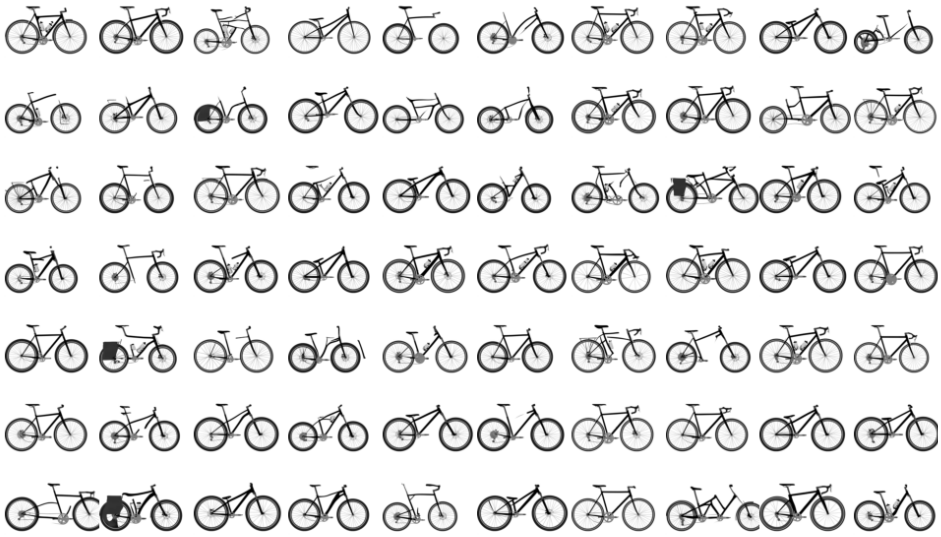


Figure 2.19: BIKED images randomly generated by EDM [71]. EDM is theoretically based on Score-matching models [159], but attempts to significantly reduce the sampling number by focusing on a pre-defined range of noise scales. Compared to DDIM, EDM has indeed deliver DDPM-like visual quality images, but we observe that it fails badly in design plausibility.

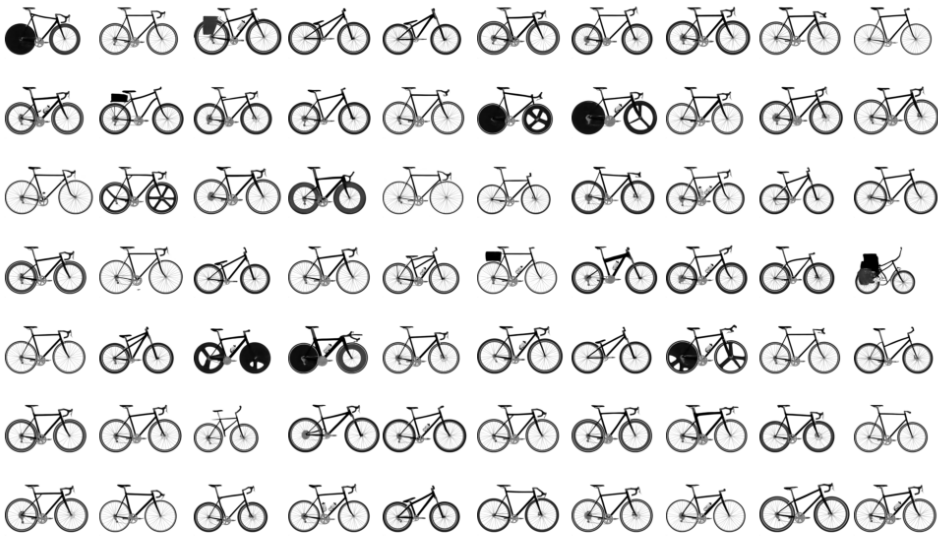


Figure 2.20: BIKED images randomly generated by PoDM. Based on the achievement of EDM, our work figures out a way of locating the focusing range of noise scales and hereby well-addresses the trio-trade-off among sampling time, visual quality and design plausibility.