The Surprising Effectiveness of Skip-Tuning in Diffusion Sampling

Jiajun Ma [1 2 *] **Shuchen Xue** [3 4 *] **Tianyang Hu** [5] **Wenjia Wang** [1 2] **Zhaoqiang Liu** [6] **Zhenguo Li** [5]
Zhi-Ming Ma [3 4] **Kenji Kawaguchi** [7]

Abstract

With the incorporation of the UNet architecture, diffusion probabilistic models have become a dominant force in image generation tasks. One key design in UNet is the skip connections between the encoder and decoder blocks. Although skip connections have been shown to improve training stability and model performance, we point out that such shortcuts can be a limiting factor for the complexity of the transformation. As the sampling steps decrease, the generation process and the role of the UNet get closer to the push-forward transformations from Gaussian distribution to the target, posing a challenge for the network's complexity. To address this challenge, we propose Skip-Tuning, a simple yet surprisingly effective training-free tuning method on the skip connections. For instance, our method can achieve 100% FID improvement for pretrained EDM on ImageNet 64 with only 19 NFEs (1.75), breaking the limit of ODE samplers regardless of sampling steps. Surprisingly, the improvement persists when we increase the number of sampling steps and can even surpass the best result from EDM-2 (1.58) with only 39 NFEs (1.57). Comprehensive exploratory experiments are conducted to shed light on the surprising effectiveness of our Skip-Tuning. We observe that while Skip-Tuning increases the score-matching losses in the pixel space, the losses in the feature space are reduced, particularly at intermediate noise levels, which coincide with the most effective range accounting for image quality improvement.

1. Introduction

Over the past few years, Diffusion Probabilistic Models (DPMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020b) have garnered significant attention for their success in generative modeling, especially high-resolution images. A special trait of DPMs is that the training and sampling are usually decoupled. The training target is the multi-level score function of the noisy data, captured by the UNet in denoising score matching. Various sampling methods are developed based on differential equation solvers to generate new samples, enabling us to trade-off efficiency against quality (discretization error) by choosing the number of sampling steps. This leaves room for *post-training modifications* to the score net that may significantly improve the diffusion sampling process. Many works have been dedicated to efficient diffusion sampling with pre-trained DPMs with as few steps as possible, e.g., through improved differential equation solvers (Lu et al., 2022; Zhao et al., 2023; Xue et al., 2023), better time step selections (Xue et al., 2024), extra distillation training (Salimans & Ho, 2022; Song et al., 2023; Luo et al., 2023), etc. In this paper, we unveil an important yet missing angle to improving diffusion sampling by looking into the *network architecture*.

The concept of DPM (Sohl-Dickstein et al., 2015) long predates their empirical success. Despite the elegant mathematical formulation, the empirical performance has been lacking until the adoption of the UNet architecture for denoising score matching (Song & Ermon, 2019; Ho et al., 2020). One distinctive feature of the UNet design is the skip connection between the encoder and decoder blocks, which was originally designed for image segmentation (Ronneberger et al., 2015). Nevertheless, numerous works have since demonstrated its effectiveness in DPMs, and after various architectural modifications, such skip designs are still mainstream. When experimenting with the transformer architecture, Bao et al. (2023) conducted comprehensive investigations that the long skip connections can be helpful for diffusion training. However, such skip connections may not be an ideal design choice for few-shot diffusion sampling. As the sampling steps decrease, the generation process or role of the UNet gets closer to the push-forward transformations from Gaussian distribution to the target, which essentially contradicts the goal of score matching.

[*] Equal contribution [1] The Hong Kong University of Science and Technology [2] Hong Kong University of Science and Technology (Guangzhou) [3] University of Chinese Academy of Sciences [4] Academy of Mathematics and Systems Science [5] Huawei Noah's Ark Lab [6] University of Electronic Science and Technology of China [7] National University of Singapore. Correspondence to: Tianyang Hu <hutianyang.up@outlook.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Pushing data-agnostic Gaussian distributions towards highly complicated and multi-modal data distributions is extremely challenging for the network's expressivity (Xiao et al., 2018; Hu et al., 2023). From this perspective, skip connections, especially low-level ones, may restrict the UNet's capacity since they provide shortcuts from the encoder to the decoder.

To address the challenge, we propose Skip-Tuning, a simple and training-free modification to the strength of the residual connections for improved few-step diffusion sampling. Through extensive experiments, we found that our Skip-Tuning not only significantly improves the image quality in the few-shot case, but also is universally helpful for more sampling steps. Surprisingly, we can break the limit of ODE samplers in only 10 NFEs with EDM (Karras et al., 2022) on ImageNet (Deng et al., 2009) and beat the heavily optimized EDM-2 (Karras et al., 2023) with only 39 NFEs. Our method generalizes well across a wide range of DPMs with various architectures, e.g., LDM (Rombach et al., 2022) and UViT (Bao et al., 2023). Comprehensive exploratory experiments are conducted to shed light on the surprising effectiveness of our Skip-Tuning. Our findings indicate that although the original denoising score matching losses increase with Skip-Tuning, the counterparts in the feature space decrease, especially for intermediate noise values (sampling stages). The effective range coincides with that for image quality improvement, as identified by our exhaustive window search. Extensive experiments on fine-tuning with feature-space score-matching are conducted, showing significantly worse performance compared with Skip-Tuning. Besides FID, we also experimented with other metrics for generation quality, e.g., Inception Score, Precision & Recall, and Maximum Mean Discrepancy (MMD) (Jayasumana et al., 2023). For instance, an investigation of the inversion process shows that Skip-Tuned UNet can result in more Gaussian inversed noise in terms of MMD with various kernels.

This work contributes to a better understanding of the UNet skip connections in diffusion sampling by showcasing a simple but surprisingly useful training-free tuning method for improved sample quality. The proposed Skip-Tuning is orthogonal to existing diffusion samplers and can be incorporated to fully unlock the potential of DPMs.

2. Preliminary

Diffusion probabilistic models. DPMs (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020b; Kingma et al., 2021) add noise to data through the SDE

$$\mathrm{d}\mathbf{x}_t = f(t)\mathbf{x}_t\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t,$$

where $\mathbf{w}_t \in \mathbb{R}^D$ represents the standard Wiener process. For any $t \in [0, T]$, the distribution of $\mathbf{x}_t$ conditioned on $\mathbf{x}_0$ is a Gaussian distribution, i.e., $\mathbf{x}_t|\mathbf{x}_0 \sim \mathcal{N}(\alpha_t\mathbf{x}_0, \sigma_t^2\mathbf{I})$. The functions α_t and σ_t are chosen such that $\mathbf{x}_T$ closely approximate a zero-mean Gaussian distribution with an identity covariance matrix. Anderson (1982) demonstrates that the forward process has an equivalent reverse-time diffusion process (from T to 0). Thus the generating process is equivalent to solving the diffusion SDE (Song et al., 2020b):

$$\mathrm{d}\mathbf{x}_t = \left[f(t)\mathbf{x}_t - g^2(t)\nabla_{\mathbf{x}}\log q_t(\mathbf{x}_t)\right]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}_t, \quad (1)$$

where $\bar{\mathbf{w}}_t$ represents the Wiener process in reverse time, and $\nabla_{\mathbf{x}}\log q_t(\mathbf{x})$ is the score function. Moreover, Song et al. (2020b) also show that there exists a corresponding deterministic process that shares the same marginal probability densities $q_t(\mathbf{x})$ as (1):

$$\mathrm{d}\mathbf{x}_t = \left[f(t)\mathbf{x}_t - \frac{1}{2}g^2(t)\nabla_{\mathbf{x}}\log q_t(\mathbf{x}_t)\right]\mathrm{d}t.$$

We usually train a score network $\boldsymbol{s}_{\boldsymbol{\theta}}(\mathbf{x}, t)$ parameterized by $\boldsymbol{\theta}$ to approximate the score function $\nabla_{\mathbf{x}}\log q_t(\mathbf{x})$ in (1) by optimizing the denoising score matching loss (Vincent, 2011; Song et al., 2020b):

$$\mathcal{L} = \mathbb{E}_t\left\{\omega_t\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_t}\left[\|\boldsymbol{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - \nabla_{\mathbf{x}}\log q_{0t}(\mathbf{x}_t|\mathbf{x}_0)\|_2^2\right]\right\},$$

where ω_t is a weighting function. While introducing stochasticity in diffusion sampling has been shown to achieve better quality and diversity (Karras et al., 2022; Xue et al., 2023), ODE-based sampling methods (Song et al., 2020a; Zhang & Chen, 2022; Lu et al., 2022; Zhao et al., 2023) are superior when the sampling steps are fewer.

UNet. UNet is an architecture based on convolutional neural networks originally proposed for image segmentation (Ronneberger et al., 2015) but recently proved successful in score estimation (Song & Ermon, 2019; Ho et al., 2020). The U-Net is composed of a group of down-sampling blocks, a group of up-sampling blocks, and long skip connections between the two. See Figure 1 for illustration. Inside the UNet architecture of (Dhariwal & Nichol, 2021), it contains 16 layers of connections from the bottom to the top, where the skip vectors d from the down-sampling component are concatenated with the corresponding up-sampling vectors u. Among these 16 layers, 10 of them have skip vectors that share the same channels as the vectors in the corresponding up-sampling component. In this work, we uncover the significant improvement brought by manipulating the magnitude of skip vectors in the sampling process and provide detailed explanations of these enhancements.

3. Skip-Tuning for Diffusion Sampling

Consider the extreme case where single-step mapping directly generates images from random noises. Although this case has been widely explored in the diffusion distillation setting (Salimans & Ho, 2022; Song et al., 2023; Luo et al.,

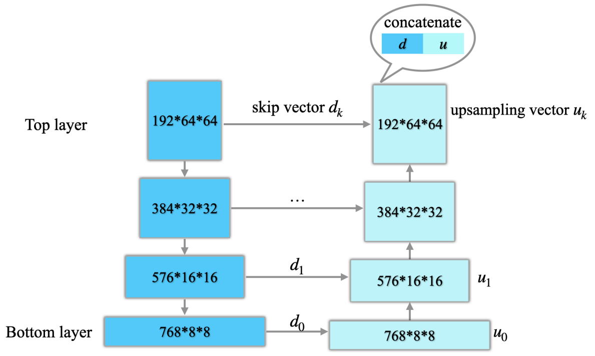


Figure 1. The UNet demonstration figure.

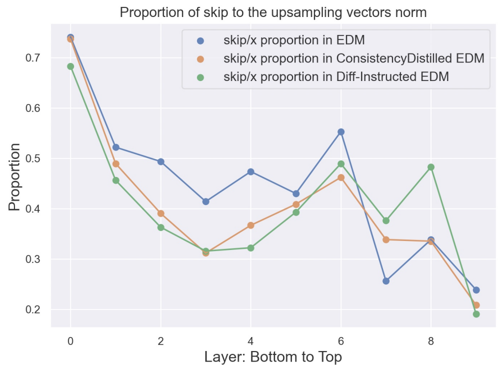


Figure 2. The layerwise down-sampling skip to up-sampling vectors l_2 norm proportion.

2023), the performance is far from optimal by pure sampling methods without extra training. This limitation may be traced back to the capacity of the UNet architecture. In the one-step sampling setting, the UNet acts like a GAN generator (Goodfellow et al., 2014) doing push-forward generation. With data-agnostic choices of the input distribution, the required transformation complexity can be huge, especially when the target distribution is multi-modal or supported on a low-dimensional manifold (Hu et al., 2023).

The skip connection of UNet, which connects the down-sampling and up-sampling components, can be detrimental to the push-forward transformation. To demonstrate, we examine the relative strength that calculates the ratio of l_2 norms between the down-sampling skip vector d versus the up-sampling vectors u in each of the layers, i.e.,

$$\text{prop}_i = \|d_i\|_2/\|u_i\|_2.$$

Figure 2 demonstrates the layerwise prop_i of EDM, CD-distilled EDM (Song et al., 2023) and DI-distilled EDM (Luo et al., 2023). We found that the residual components from the encoder are less pronounced for the distilled UNets. To be more specific, the average layerwise l_2 norm ratio, i.e., $\frac{1}{k}\sum_i^k(\|d_i\|_2/\|u_i\|_2)$ for the base EDM model is 0.446, while those for the distilled models are 0.433 for DI and 0.404 for CD, confirming our hypothesis.

Further, we verify the overall model complexity increase in the distilled EDM network (CD and DI) versus the original EDM on ImageNet 64 in Table 1. Specifically, we choose the l_2 norm of the model gradient (Hu et al., 2023; Negrea et al., 2019; Li et al., 2019) to reflect the complexity of the EDM network U, i.e.,

$$\text{gradient norm}(U) = \mathbb{E}_x\|\text{autograd}_x(U(x))\|_2.$$

Motivated by this observation, we consider manually decreasing the skip connections to improve few-shot diffusion sampling in a training-free fashion.

Definition 3.1 (Skip-Tuning)**.** We introduce skip coefficient ρ_i's to control the relative strength of the skipped down-sampling outputs d_i. Specifically, we add ρ_i in the concatenation of the d_i and u_i, i.e., $\text{concatenate}(d_i \cdot \rho_i, u_i)$. In this work, we only consider $\rho < 1$.

Table 1. Gradient norms of EDM and distilled EDM. The σ values (noise standard deviation) are different because the two distilled models have different initial sigma settings.

	Gradient norm
EDM ($\sigma = 80$)	0.1219
CD EDM ($\sigma = 80$)	0.3525
EDM ($\sigma = 5$)	0.9425
DI EDM ($\sigma = 5$)	8.4765

Through carefully choosing ρ for pre-trained UNet, we can mimic the approximately decreasing l_2 norm ratio observed in Figure 2. Specifically, we adopt the linear interpolation of bottom and top layer ρ_{bottom} and ρ_{top} to match with the pattern(For instance, set the ρ_{bottom} as 0.5 and increase it linearly towards 1.0 for ρ_{top}), i.e.,

$$\Delta\rho = \frac{(\rho_{\text{top}} - \rho_{\text{bottom}})}{k}, \quad \rho_i = \rho_{\text{bottom}} + \Delta\rho \cdot i.$$

To demonstrate its effectiveness, we conduct experiments with pre-trained EDM (Karras et al., 2022) on ImageNet 64. We use the standard class-conditional generation following the settings in (Karras et al., 2022), without extra guidance methods (Dhariwal & Nichol, 2021; Ho & Salimans, 2022; Ma et al., 2023b; Liu et al., 2024). The few-step sampling results with the Heun and UniPC (Zhao et al., 2023) are reported in Table 2. With less than 10 NFEs, our Skip-Tuning can improve the FID by around 100%.

Table 2. EDM Skip-Tuning with few-step sampling. ρ stands for the linear interpolation from the bottom to the top layer.

	Sampler	step	NFE	FID
EDM	Heun	5	9	35.12
EDM (ρ:0.55 to 1.0)	Heun	5	9	18.71
EDM	UniPC	9	9	5.88
EDM (ρ:0.68 to 1.0)	UniPC	9	9	2.92

Remark 3.2 (Beyond existing architecture). The modified skip coefficient cannot be absorbed into existing model parameters, due to the placement of the input within the group normalization[1], SiLU activation function, and convolution function in the forward function. The nonlinearity of the SiLU activation prevents the study of the skip coefficient value within the convolution function.

Skip-Tuning offers extra flexibility to pretrained diffusion models in a training-free fashion. Besides the surprising effectiveness in few-shot diffusion sampling, we also test out its performance for distilled UNet in one-step generation. In Table 3, we can observe a significant improvement over the baseline. It is worth mentioning that the ideal ρ for distilled UNets are close to 1.0 (CD: 0.91; DI: 0.98) due to the implicit reduction of skip connections through the distillation process, as confirmed by the lower skip norm proportion of distilled models in Figure 2.

Table 3. Skip-Tuning in distilled EDM (CD: Consistency Distillation, DI: Diff-Instruct). *: results reported in original papers. †: In our reproduction, we replaced flash attention with standard attention for better GPU compatibility.

	NFE	FID
CD EDM*	1	6.20
CD EDM†	1	6.85
CD EDM†(ρ: 0.91 to 0.96)	1	5.56
DI EDM*	1	4.24
DI EDM†	1	4.16
DI EDM†(ρ_{TOP}: 0.98)	1	3.98

In Figure 3, we demonstrate the monotone increase in the complexity of the EDM network U by diminishing the downsampling vector d in the skip concatenation ($\rho < 1$), where the model complexity is estimated by the gradient norm.

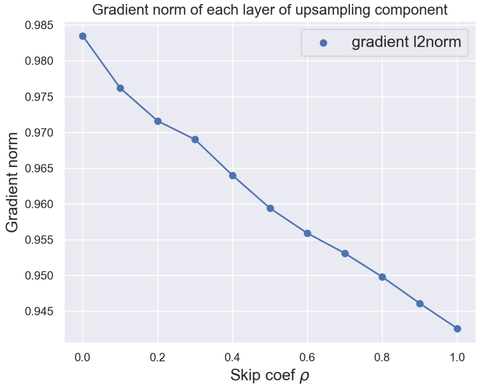


Figure 3. The gradient l_2 norm changes with skip coefficient ρ.

4. Breaking the ODE-Sampling Limit

Our proposed Skip-Tuning has demonstrated surprising effectiveness in improving few-shot diffusion sampling. A natural question that follows is whether the improvement can still be significant if we increase the number of sampling steps. Current sampling methods are mostly based on ODE solvers which discretize the diffusion ODE according to specific schemes. As the sampling steps increase, the discretization error approaches zero, and FID scores will also saturate to a limit.

Remark 4.1. Most current distillation methods (e.g. Progressive Distillation Salimans & Ho (2022), Consistency Model Song et al. (2023)) learn the map of ODE trajectory from noise to data, which is simulated through the ODE-sampling limit of the teacher model.

In this section, we further test the limit of Skip-Tuning to see how it fares with the state-of-the-art DPMs, e.g., EDM (Karras et al., 2022), EDM-2 (Karras et al., 2023), LDM (Rombach et al., 2022), UViT(Bao et al., 2023).

We begin with EDM on ImageNet, where existing literature indicates that any ODE sampler, with arbitrary sampling steps, cannot get FID below 2.2 (Karras et al., 2022). Surprisingly, as showcased in Table 4, our Skip-Tuning EDM surpasses the previous ODE-sampling limit with just 19 NFEs (FID: 1.75).

Furthermore, by increasing the sampling steps to 39 NFEs in Table 5, our Skip-Tuning on the original EDM (Karras et al., 2022) (FID: 1.57) can even beat the heavily optimized EDM-2 (Karras et al., 2023) (FID: 1.58). Similar conclusions can be drawn from the sampling results on AFHQv2 (Choi et al., 2020; Karras et al., 2021) 64×64 in Table 6.

[1] Oftentimes, the concatenation will first go through a normalization layer, e.g., GroupNorm in EDM. Our proposed Skip-Tuning mainly affects the residual connection within each UNet block (details can be found in Appendix C)

Table 4. Skip-Tuning in EDM with ODE sampling. ρ in the bracket stands for the linear interpolation from the bottom to the top layer.

	Sampler	steps	NFE	FID
EDM	Heun	10	19	3.64
EDM(ρ: 0.78 to 1.0)	Heun	10	19	1.88
EDM	UniPC	19	19	2.60
EDM(ρ: 0.82 to 1.0)	UniPC	19	19	1.75
EDM	UniPC	39	39	2.21
EDM(ρ: 0.83 to 1.0)	UniPC	39	39	1.57

Table 5. ODE sampling limit. The EDM checkpoint for baseline and the Skip-Tuning is from (Karras et al., 2022). The EDM-2-S results are from (Karras et al., 2023).

	NFE	Mparams	FID
EDM	79	296	2.22
EDM(ρ: 0.83 to 1.0)	39	296	1.57
EDM-2-S	63	280	1.58

Table 6. Skip-Tuning in EDM with ODE sampling on AFHQv2 64×64.

	Sampler	steps	NFE	FID
EDM	UniPC	9	9	4.47
EDM(ρ: 0.75 to 1.0)	UniPC	9	9	3.85
EDM	UniPC	19	19	2.13
EDM(ρ: 0.87 to 1.0)	UniPC	19	19	2.03
EDM	UniPC	39	39	2.05
EDM(ρ: 0.90 to 1.0)	UniPC	39	39	1.96

To demonstrate the stability of Skip-Tuning in enhancing the sampling performance, we conduct experiments on varying skip coefficients ρ under different steps of UniPC sampling shown in Figure 4. The FID curves all exhibit U-shaped patterns under different NFEs. For NFE = 9, the "sweet point" of the skip coefficient for the U-shaped FID curve is between 0.65 and 0.70. This can be attributed to the increased network complexity requirement in few-step settings. For NFE = 39 (which converges well, as the FID of 2.21 for $\rho = 1$ matches the result of 511 NFEs Heun sampling (Karras et al., 2022)), the ρ sweet point lies around 0.85. We summarize the findings as follows:

- With a fixed skip coefficient, the FID score improves monotonically as the number of sampling steps increases.
- For a given sampling step, there exists an optimal skip coefficient range.
- With increasing sampling steps, the optimal skip coefficient monotonically increases towards a limit below 1.

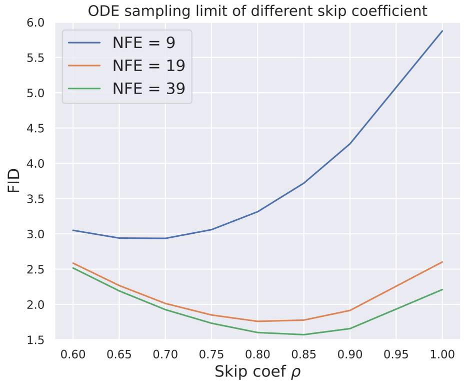


Figure 4. ODE UniPC sampling results of different skip coefficients and steps.

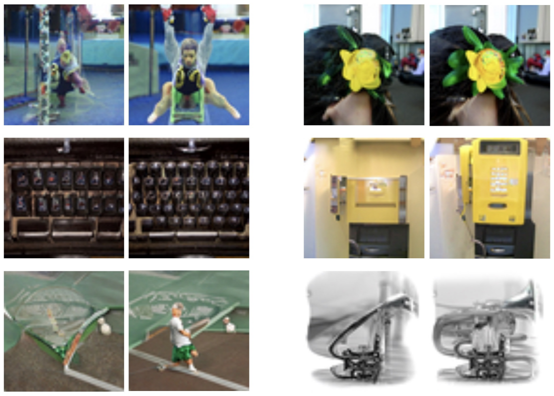


Figure 5. The left-hand side 64x64 figures are sampled from ODE 10 steps (FID: 3.64); the right-hand side figures are sampled from ODE 10 steps with Skip-Tuning $\rho = 0.78$ (FID: 1.88).

Besides EDM, our Skip-Tuning can also improve other DPMs consisting of skip connection designs, including LDM (Rombach et al., 2022) and UViT(Bao et al., 2023), as presented in Table 7.

Table 7. Skip-Tuning in LDM and UViT in 256x256 ImageNet.

	Steps	FID
LDM	5	12.97
LDM(ρ: 0.83 to 1.0)	5	11.29
LDM	10	4.91
LDM(ρ: 0.95 to 1.0)	10	4.67
LDM	20	4.25
LDM(ρ: 0.994 to 1.0)	20	4.13
UViT	50	2.32
UViT(ρ: 0.82 to 1.0)	50	2.21

In addition to the remarkable improvement in quantitative metrics, Figures 5 and 6 visually demonstrate that Skip-Tuning contributes to object and semantic enrichment. For instance, the flower picture (right-hand side of first row) in

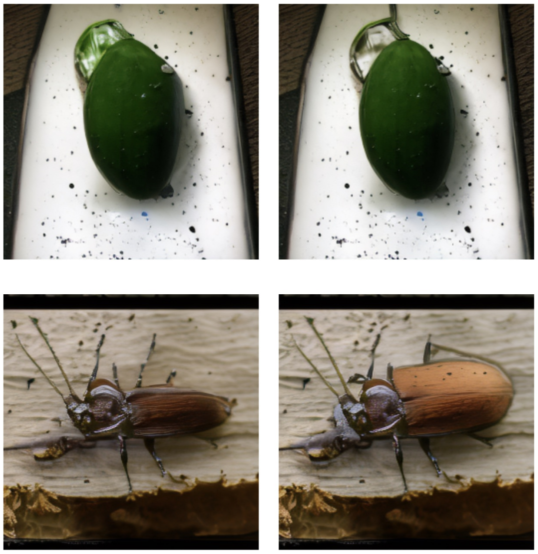


Figure 6. The left-hand side 256x256 figures are sampled from LDM in 10 steps (FID: 4.91); the right-hand side figures are sampled from LDM 10 steps with Skip-Tuning $\rho = 0.78$ (FID: 4.67).

Figure 5 is decorated with leafy details and a more vibrant yellow color after Skip-Tuning. We further test the effectiveness of Skip-Tuning in text-to-image settings. We applied Skip-Tuning in Stable Diffusion 2 [2] to generate 768x768 images for evaluation, as shown in Figure 7. The skip coefficient $\rho = 0.8$ and 20 sampling steps are used throughout the experiments. Based on the comparisons in Figure 7, Skip-Tuning primarily serves to repair objects and enhance their semantic content.

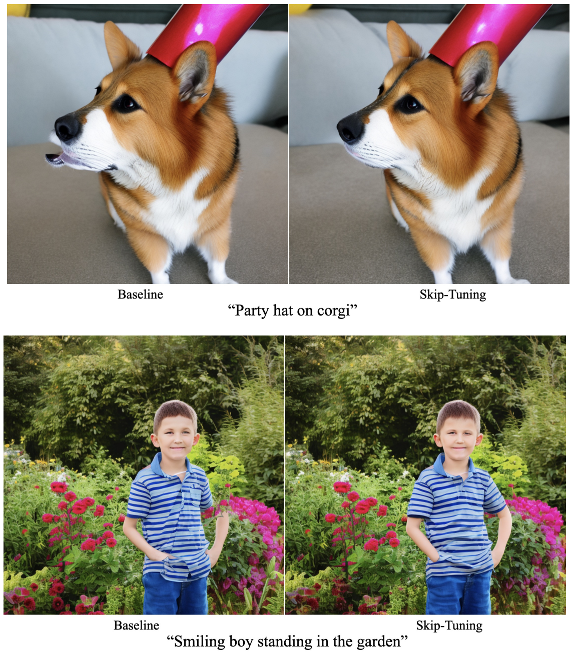
5. Demystifying Skip-Tuning

In this section, we thoroughly examine how Skip-Tuning contributes to diffusion model sampling. As emphasized before, the training and sampling of DPMs are decoupled. Now that Skip-Tuning offers significant post-training sampling improvement, the first question to investigate is its effect on diffusion training loss.

5.1. Denoising Score Matching

Consider the denoising score-matching loss below

$$\mathcal{L}_{\text{pixel}} = \mathbb{E}_t \left\{ \omega_t \mathbb{E}_{\mathbf{x}} \left[\| \mathbf{x}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - \mathbf{x} \|_2^2 \right] \right\}.$$



“Party hat on corgi”

“Smiling boy standing in the garden”

Figure 7. The left-hand side 768x768 figures are sampled from stable diffusion 2 with 20 sampling steps; the right-hand side figures are sampled with Skip-Tuning $\rho = 0.8$.

Table 8 compares the score-matching losses of the original EDM and its checkpoints with Skip-Tuning ($\rho = 0.8$), where we can see that Skip-Tuning makes the pixel loss worse. This is anticipated since the baseline EDM checkpoint is optimized under this pixel loss $\mathcal{L}_{\text{pixel}}$. Then, why can the quality be significantly improved (FID improved from 3.64 to 1.88) while the validation loss is higher? As it turns out, instead of the original pixel space, Skip-Tuning can result in a decreased denoising score-matching loss in the *feature space* of various discriminative models f, as described below:

$$\mathcal{L}_{\text{feature}} = \mathbb{E}_t \left\{ \omega_t \mathbb{E}_{\mathbf{x}} \left[\| f(\mathbf{x}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)) - f(\mathbf{x}) \|_2^2 \right] \right\}.$$

Table 8 lists losses measured in the feature space of Inception-V3 (Szegedy et al., 2016), ResNet-101 (He et al., 2016) (trained on ImageNet with the output dimension of 2048), and CLIP-ViT (Radford et al., 2021) image encoder (trained on web-crawled image-caption pairs and public datasets; the output dimension is 1024). In the Skip-Tuning setting, the score-matching losses in the feature space of classifiers and the CLIP encoder all dropped, indicating improved score-matching estimates in the discriminative model feature space.

In Table 9, we extend the comparison of score-matching loss in the ResNet101 feature space ($\mathcal{L}_{\text{ResNet-101}}$) across different sampling σ levels. The results demonstrate that the improvement in feature-space score-matching achieved by Skip-Tuning is not uniform over time (σ) and is particularly noticeable for intermediate noise values (sampling stages). This observation serves as motivation for exploring time-dependent Skip-Tuning in the next section.

[2] checkpoints: `https://huggingface.co/stabilityai/stable-diffusion-2/blob/main/768-v-ema.ckpt`

Table 8. EDM score-matching losses in pixel, discriminative feature, and CLIP image encoder space.

	BASELINE (FID:3.64)	SKIP-TUNINGρ : 0.8 (FID:1.88)
$\mathcal{L}_{\text{PIXEL}}$	0.5238	0.5253
$\mathcal{L}_{\text{INCEPTION-V3}}$	4.3466	4.3219
$\mathcal{L}_{\text{RESNET-101}}$	30.8421	30.7297
$\mathcal{L}_{\text{CLIP-VIT}}$	12.6432	12.4550

Table 9. Comparison of score-matching loss in the ResNet101 feature space ($\mathcal{L}_{\text{ResNet-101}}$) between the baseline EDM and Skip-Tuning EDM. The σ values are selected from 5 steps of ODE sampling.

σ	BASELINE	SKIP-TUNING
0.002	99.9295	99.5523
0.1698	27.1737	27.4448
2.5152	13.7342	13.6390
17.5278	14.2074	12.9545
80.0	12.6893	12.6827

5.2. Noise Level Dependence

In our exploration of the time-dependent properties of Skip-Tuning, we aimed to identify the time interval that provides the greatest FID improvement during diffusion sampling. To achieve this, we conducted an exhaustive window search. By dividing the sigma interval $[0.002, 80]$ into 13 non-overlapping sub-intervals, each consisting of only 4 steps of the sampling process, we performed Skip-Tuning separately within each sub-interval. The original model was used outside of these intervals. The exhaustive search results in Figure 8 reveal that Skip-Tuning during the middle stage of the σ range contributes the most to sampling performance. This observation is consistent with the lower score-matching loss in the ResNet101 feature space ($\mathcal{L}_{\text{ResNet-101}}$) achieved by Skip-Tuning at the middle σ stage, as shown in Table 9.

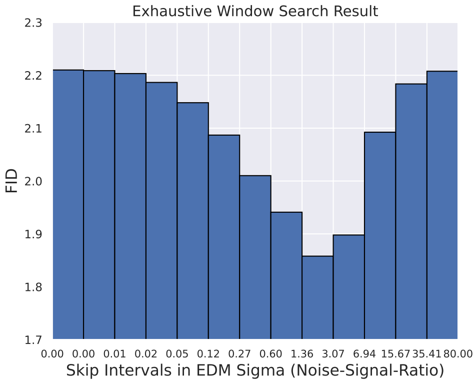


Figure 8. Exaustive window search

Besides, we further verify that different diffusion models favor different time schedules of Skip-Tuning based on their training objectives. Figure 11 in the Appendix displays the two opposite linear interpolations of ρ across the sampling time: "increasing ρ" represents ρ linearly increased from value ρ_0 at time 0 to 1.0 at time T while "decreasing ρ" represents the ρ at time 0 linearly decreased from value 1.0 to ρ_0 at time T. The rationale is that at different time steps, the required complexity from the score network is different. With noise prediction models such as LDM, the task becomes easier as noise level σ increases while it is the opposite for data prediction models such as EDM. As we have established that decreasing ρ increases the network complexity, the ideal schedule for ρ should be correspondingly inverse.

Table 10 compares the impact of different time-dependent ρ orders on sampling performance. The EDM model favors the decreasing ρ order, resulting in a smaller skip coefficient at time T (allowing less noise to pass through) and a larger skip coefficient at time 0 (yielding increasingly clean images). Conversely, the LDM and UViT models prefer the increasing ρ order, indicating a reversed preference for time-dependent skip coefficients.

Table 10. Comparison of ρ time-dependent order among EDM, LDM, and UViT. The increasing order indicates a linear increase of ρ from ρ_0 to 1.0 over time 0 to T, while the decreasing order signifies a linear decrease of ρ from 1.0 to ρ_0 over time 0 to T.

	STEPS	INCREASING ρ	DECREASING ρ
EDM(ρ_0:0.78)	10	1.98	1.88
LDM(ρ_0:0.95)	10	4.67	5.15
UVIT(ρ_0:0.82)	50	2.21	2.47

5.3. Skip-Tuning vs Fine-Tuning

After revealing that Skip-Tuning contributes to score-matching in the discriminative feature space, a natural question occurs: can we achieve the same improvement by fine-tuning the diffusion model based on *score-matching loss in feature space*? To address this question, we conduct two types of experiments, only fine-tuning the skip coefficient ρ and full fine-tuning with all the UNet parameters. Surprisingly, both results indicate that direct fine-tuning can lead to sampling performance deterioration and is not comparable to Skip-Tuning.

Fine-tuning ρ. Table 11 lists the sampling results obtained after fine-tuning ρ using the score-matching loss in ResNet101 feature space. Directly fine-tuning ρ will drive some skip coefficients greater than 1, leading to a significant decline in performance. The generated images are almost noises, as indicated by the exploded FID. To eliminate the possibility of $\rho > 1$, we then apply a Sigmoid function to constrain $\rho \in (0, 1)$. The results are significantly improved but not as good as direct Skip-Tuning.

Table 11. EDM skip coefficient ρ fine-tuned with score-matching loss in ResNet101 feature space on ImageNet 64x64.

	STEPS	NFE	FID
EDM	5	9	35.12
EDM ρ FINE-TUNED	5	9	215.09
EDM SIGMOID(ρ) FINE-TUNED	5	9	18.92
EDM	10	19	3.64
EDM ρ FINE-TUNED	10	19	112.15
EDM SIGMOID(ρ) FINE-TUNED	10	19	2.77

Full fine-tuning. Table 12 presents the fine-tuning of the full network parameters of EDM checkpoint using a hybrid loss combining vanilla score matching and score-matching in the feature space. Initially, there was a slight performance improvement, but as training progressed, it deteriorated. Similarly, fine-tuning struggles to match the quality and stability achieved by Skip-Tuning.

$$\mathcal{L}_{\text{hybrid}} = \mathcal{L}_{\text{pixel}} + \mathcal{L}_{\text{feature}}.$$

The experiment results show that naively incorporating the Inception-V3 as a feature extractor in the fine-tuning loss does not produce significant and consistent improvement compared with Skip-Tuning. Our comparisons in this section indicate that improving the score-matching loss in the feature space is only one aspect of Skip-Tuning and its effectiveness cannot be encapsulated by naive fine-tuning. In the next part, we take a look at how Skip-Tuning affects the inverse process of diffusion sampling.

Table 12. EDM fine-tuned with Inception-V3 modeling score-matching loss.

	MIMG	NFE	FID
EDM (INITIAL)	0	19	2.60141
EDM	4	19	2.58764
EDM	10	19	2.51128
EDM	30	19	3.81702
EDM	60	19	6.02844

5.4. Inverse Process

Simulating the diffusion ODE from time 0 to time T, we inverse the data to (approximately) a Gaussian noise. This raises the question of whether skip tuning can improve the results of the inversion process. We evaluate the distance between the inverted (pseudo) Gaussian noise and the ground truth Gaussian distribution using Mean Maximum discrepancy (MMD) as a metric. A brief introduction to MMD can be found in Appendix B. Specifically, we inverse 10k images to get 10k noises and calculate the MMD distance between 10k generated noises and 10k ground truth noises. The experiments are conducted several times and the average results are reported in Tabel 13. For each kernel, we normalize the baseline result to 1.

$$\mathrm{d}\mathbf{x}_t = \left[f(t)\mathbf{x}_t - \frac{1}{2} g^2(t) \nabla_{\mathbf{x}} \log q_t(\mathbf{x}_t) \right] \mathrm{d}t. \quad (2)$$

Table 13. Comparison of MMD distance.

MMD KERNEL	STEPS	$\rho = 1$	$\rho = 0.7$
LINEAR KERNEL	9	1	0.9793
RBF KERNEL	9	1	1.0000
LAPLACIAN KERNEL	9	1	1.0000
SIGMOID KERNEL	9	1	0.9592
IMQ KERNEL	9	1	1.0143
POLYNOMIAL KERNEL	9	1	0.9912
COSINE KERNEL	9	1	0.9879

The results demonstrate that Skip-Tuning decreases the discrepancy between the inverted noise and the standard Gaussian noise under most kernels, aligning with the generating process.

5.5. Relationship with Stochastic Sampling

Stochastic sampling can be viewed as an interpolation of diffusion ODE and Langevin diffusion as follows:

$$\begin{aligned}\mathrm{d}\mathbf{x}_t = & \left[f(t)\mathbf{x}_t - \frac{1}{2} g^2(t) \nabla_{\mathbf{x}} \log q_t(\mathbf{x}_t) \right] \mathrm{d}t \\ & - \frac{\tau^2(t)}{2} g^2(t) \nabla_{\mathbf{x}} \log q_t(\mathbf{x}_t) \mathrm{d}t + \tau(t) g(t) \mathrm{d}\bar{\mathbf{w}}_t.\end{aligned} \quad (3)$$

Stochastic sampling can surpass the ODE sampling limit by injecting additional noise during sampling (Song et al., 2020b; Karras et al., 2022; Xue et al., 2023). Karras et al. (2022) asserts that the implicit Langevin diffusion in stochastic sampling drives the sample towards the desired marginal distribution at a given time that corrects the error in earlier sampling steps. Xue et al. (2023) give an inequality on KL divergence to show the superiority of stochastic sampling.

However, the stochastic strength $\tau(t)$ during stochastic sampling affects the sampling. Karras et al. (2022) also provides empirical results on the ImageNet-64 dataset: stochastic sampling can improve the FID score of the baseline model from 2.66 to 1.55, and from 2.22 to 1.36 for the EDM model. They also observed that the optimal amount of stochastic strength for the EDM model is much lower than the baseline model. We conduct extra experiments to explore the effect of the skip coefficient combined with stochastic sampling. The experiment results are shown in Fig. 9, the sweet point of the stochastic strength decreases as the skip coefficient decreases. We find that a slight Skip-Tuning can improve the stochastic sampling for all stochastic strength (ρ = 0.95 over ρ = 1).

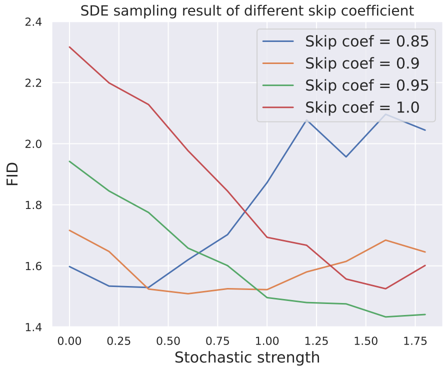


Figure 9. Combination of skip tuning and stochastic sampling

6. Related Work

FreeU Most related to our work is FreeU (Si et al., 2023), where the authors analyzed the contribution of skip connection in the views of image frequency decomposition. However, this does not capture the whole picture. In Figure 12 of the Appendix, we conduct wavelet transformation of the original figures and compare the score-matching loss of the pre-trained EDM checkpoint and its checkpoint with skip connection diminished to 80% ($\rho = 0.8$) under pixel and wavelet transformed space. The results in Table 15 reveal that, despite the FID improvement from 3.64 to 1.88, the score-matching losses in all wavelet frequency spaces increase. This suggests that the enhancement in generation quality is not directly linked to a better score-matching loss in the frequency space. On the other hand, our method does not contain Fourier transform and inverse Fourier transform, which requires additional computational cost. We add a detailed analysis of the difference in the operation level with FreeU in Appendix C. In terms of visual evaluation, we compare Skip-Tuning with FreeU in Stable Diffusion 2 in Figure 13 of the Appendix. We observe that FreeU mainly changes the image aesthetics, enhancing image contrast and highlighting the object but may lose fidelity; In contrast, our Skip-Tuning contributes to object enrichment and quality improvement without losing authenticity.

Diffusion architectures Efforts have been devoted to analyzing diffusion model architectures and proposing improved designs for improved training. Karras et al. (2023) conducted extensive experiments and improved the well-accepted ADM network in terms of weight normalization, block design, and exponential moving averaging training schedule. Huang et al. (2023) uncovers the impact of skip connection in stabilizing and speeding up diffusion training. Bao et al. (2023) points out that the design of skip concatenation plays a crucial role in achieving high-quality training. SCedit (Jiang et al., 2023) incorporates a fine-tuned non-linear projection component within the skip connection for controllable image generation. In contrast, our Skip-Tuning does not require extra model components to the existing UNet, saving both the training and inference costs. In terms of FID evaluation, SCedit does not exhibit a substantial improvement compared to Skip-Tuning. Ma et al. (2023a) analyzes the skip connection in improving self-supervised learning as well. In clear contrast, Skip-Tuning is a post-training design that significantly enhances the sampling performance without additional training.

Evaluation metrics Evaluating the quality of generated images is a challenging task. The FID metric has been widely used for such a purpose. However, there is still a perceivable gap between FID and human evaluation. Chong & Forsyth (2020) highlighted the bias of FID in finite sample evaluation. Jung & Keuper (2021) assesses the less sensitivity of FID to various augmentations and attributes the Inception-V3 as the cause. Parmar et al. (2022) analyzes the impact of low-level preprocessing on FID metrics, while Jayasumana et al. (2023) challenges the key assumption of FID regarding normal distribution. To provide a comprehensive evaluation of Skip-Tuning, we include other metrics such as Inception Score (IS), Precision, Recall, and Mean Maximum Discrepancy in Inception-V3 feature space (IMMD) in Table 14.

Table 14. Other evaluation metrics

	EDM	EDM SKIP TUNING
FID↓	2.21	**1.57**
IS↑	47.55	**57.64**
PRECISION↑	0.719	**0.752**
RECALL↑	**0.639**	0.625
IMMD↓	0.521	**0.335**

7. Discussion

Our proposed Skip-Tuning breaks the limit of ODE sampling, improving both the existing UNet diffusion model (teacher model) generation quality and enhancing the distilled diffusion model (student model) in one-step sampling. Through extensive investigation, we attribute the success of Skip-Tuning to improved score-matching in the discriminative feature space and a smaller discrepancy between inversed noise and ground truth Gaussian noise. These findings not only deepen our understanding of the UNet architecture but also demonstrate the remarkably useful nature of Skip-Tuning as a post-training method for enhancing diffusion generation quality. In future work, we will explore UNet inside models of different modalities to further investigate its potential.

Impact Statement

This paper presents work whose goal is to advance the field of diffusion sampling. By enhancing the efficiency and quality of image generation, this method can democratize access to high-quality visual content, benefiting industries such as entertainment, education, and marketing. However, the potential for misuse, such as creating realistic fake images and deepfakes, poses ethical challenges. Therefore, establishing robust ethical guidelines and detection mechanisms is essential to balance innovation with responsibility, ensuring these technologies serve the public good.

References

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., and Zhu, J. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22669–22679, 2023.

Choi, Y., Uh, Y., Yoo, J., and Ha, J.-W. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Chong, M. J. and Forsyth, D. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6070–6079, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Hu, T., Chen, F., Wang, H., Li, J., Wang, W., Sun, J., and Li, Z. Complexity matters: Rethinking the latent space for generative modeling. *arXiv preprint arXiv:2307.08283*, 2023.

Huang, Z., Zhou, P., Yan, S., and Lin, L. Scalelong: Towards more stable training of diffusion model via scaling network long skip connection. *arXiv preprint arXiv:2310.13545*, 2023.

Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., and Kumar, S. Rethinking fid: Towards a better evaluation metric for image generation. *arXiv preprint arXiv:2401.09603*, 2023.

Jiang, Z., Mao, C., Pan, Y., Han, Z., and Zhang, J. Scedit: Efficient and controllable image diffusion generation via skip connection editing. *arXiv preprint arXiv:2312.11392*, 2023.

Jung, S. and Keuper, M. Internalized biases in fréchet inception distance. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., and Aila, T. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35: 26565–26577, 2022.

Karras, T., Aittala, M., Lehtinen, J., Hellsten, J., Aila, T., and Laine, S. Analyzing and improving the training dynamics of diffusion models. *arXiv preprint arXiv:2312.02696*, 2023.

Kingma, D. P., Salimans, T., Poole, B., and Ho, J. Variational diffusion models. *arXiv preprint arXiv:2107.00630*, 2021.

Li, J., Luo, X., and Qiao, M. On generalization error bounds of noisy gradient methods for non-convex learning. *arXiv preprint arXiv:1902.00621*, 2019.

Liu, X., Hu, T., Wang, W., Kawaguchi, K., and Yao, Y. Referee can play: An alternative approach to conditional generation via model inversion. *arXiv preprint arXiv:2402.16305*, 2024.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022.

Luo, W., Hu, T., Zhang, S., Sun, J., Li, Z., and Zhang, Z. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *arXiv preprint arXiv:2305.18455*, 2023.

Ma, J., Hu, T., and Wang, W. Deciphering the projection head: Representation evaluation self-supervised learning. *arXiv preprint arXiv:2301.12189*, 2023a.

Ma, J., Hu, T., Wang, W., and Sun, J. Elucidating the design space of classifier-guided diffusion generation. *arXiv preprint arXiv:2310.11311*, 2023b.

Negrea, J., Haghifam, M., Dziugaite, G. K., Khisti, A., and Roy, D. M. Information-theoretic generalization bounds for sgld via data-dependent estimates. *Advances in Neural Information Processing Systems*, 32, 2019.

Parmar, G., Zhang, R., and Zhu, J.-Y. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11410–11420, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

Salimans, T. and Ho, J. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Si, C., Huang, Z., Jiang, Y., and Liu, Z. Freeu: Free lunch in diffusion u-net. *arXiv preprint arXiv:2309.11497*, 2023.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Wu, Y. and He, K. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Xiao, C., Zhong, P., and Zheng, C. Bourgan: Generative networks with metric embeddings. *Advances in neural information processing systems*, 31, 2018.

Xue, S., Yi, M., Luo, W., Zhang, S., Sun, J., Li, Z., and Ma, Z.-M. Sa-solver: Stochastic adams solver for fast sampling of diffusion models. *arXiv preprint arXiv:2309.05019*, 2023.

Xue, S., Liu, Z., Chen, F., Zhang, S., Hu, T., Xie, E., and Li, Z. Accelerating diffusion sampling with optimized time steps. *arXiv preprint arXiv:2402.17376*, 2024.

Zhang, Q. and Chen, Y. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022.

Zhao, W., Bai, L., Rao, Y., Zhou, J., and Lu, J. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *arXiv preprint arXiv:2302.04867*, 2023.

Appendix

A. Other Details

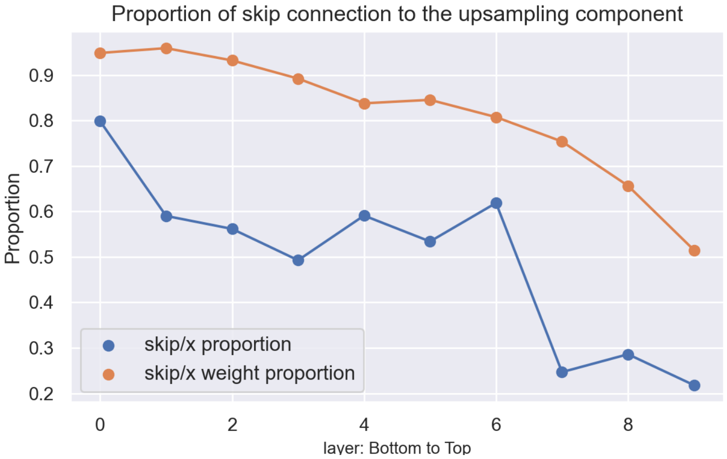


Figure 10. The skip vector and up-sampling component norm proportion. The skip vector and up-sampling weights norm proportion.

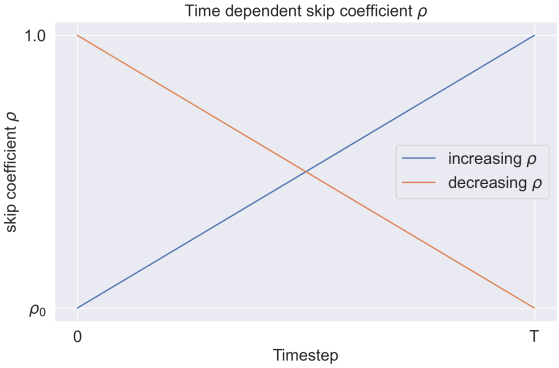


Figure 11. The time-dependent linear interpolation of skip-coefficient ρ.

Table 15. Score-matching loss in pixel and frequency space. 'LL', 'LH,' 'HL', and 'HH' represent frequency spectrum 'Approximation', ' Horizontal detail', 'Vertical detail', 'Diagonal detail' respectively.

	BASELINE (FID:3.64)	$\rho = 0.8$ (FID:1.88)
$\mathcal{L}_{\text{PIXEL}}$	0.5238	0.5253
$\mathcal{L}_{\text{LL}}$	1.6160	1.6221
$\mathcal{L}_{\text{LH}}$	0.2264	0.2267
$\mathcal{L}_{\text{HL}}$	0.2258	0.2260
$\mathcal{L}_{\text{HH}}$	0.1408	0.1409

B. Details on Mean Maximum Discrepancy (MMD)

Maximum Mean Discrepancy (MMD) (Gretton et al., 2006; 2012) is a kernel-based statistical test used as a two-sample test to determine whether two samples come from the same distribution. The MMD statistic can be viewed as a discrepancy

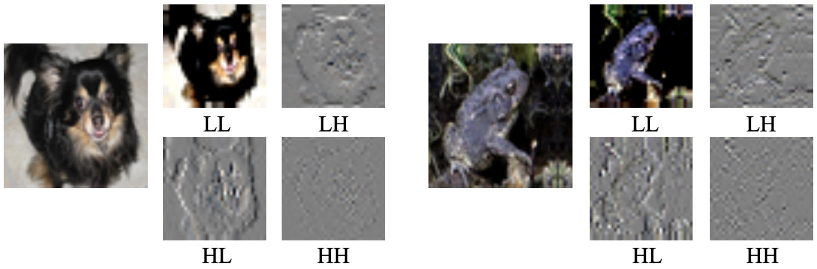


Figure 12. The wavelet transformation of figures. ’LL’, ’LH,’ ’HL’, and ’HH’ represent frequency spectrum ’Approximation’, ’ Horizontal detail’, ’Vertical detail’, ’Diagonal detail’ respectively.

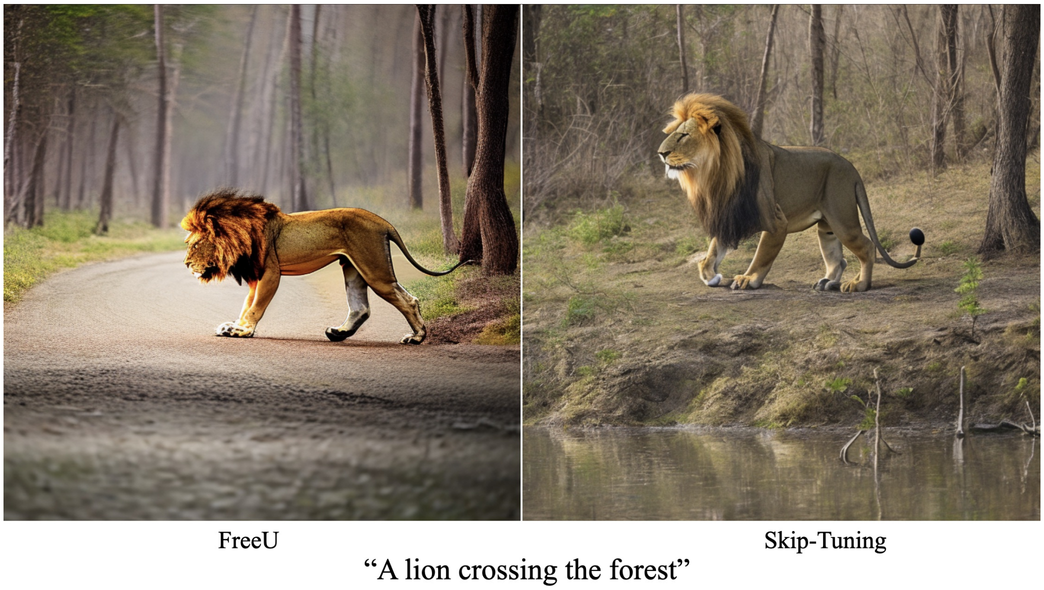


Figure 13. The comparison of Skip-Tuning and FreeU in Stable Diffusion 2 in generating text-to-images in 768x768 resolution with 20 sampling steps. FreeU is based on hyper-parameter settings reported by the author (b1:1.4, b2: 1.6, s1: 0.9, s2: 0.2). We fixed the skip coefficient $\rho = 0.8$ of Skip-Tuning through all experiments to avoid manual fine-tuning.

between two distributions. Given distribution P and Q, a feature map ϕ maps P and Q to feature space F. Denote the kernel function $k(x, y) = \langle \phi(x), \phi(y) \rangle_F$, the MMD distance with respect to the positive definite kernel k is defined by:

$$\text{MMD}^2(P, Q) = \|\mu_P - \mu_Q\|_F^2 = \mathbb{E}_P[k(X, X)] - 2\mathbb{E}_{P,Q}[k(X, Y)] + \mathbb{E}_Q[k(Y, Y)] \tag{4}$$

In practice, we only have two empirical distributions $\hat{P} = \sum_{i=1}^{m} \delta(x_i)$ and $\hat{Q} = \sum_{i=1}^{n} \delta(y_i)$ independently sampled from P and Q, we have the following unbiased empirical estimator of the MMD distance:

$$\widehat{\text{MMD}}^2(P, Q) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(y_i, y_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j) \tag{5}$$

C. Details on Group Normalization in UNetBlock

```
def forward(self, x, emb):
    orig = x
```

```
x = self.conv0(silu(self.norm0(x)))

params = self.affine(emb).unsqueeze(2).unsqueeze(3).to(x.dtype)
if self.adaptive_scale:
   scale, shift = params.chunk(chunks=2, dim=1)
   x = silu(torch.addcmul(shift, self.norm1(x), scale + 1))
else:
   x = silu(self.norm1(x.add_(params)))

x = self.conv1(torch.nn.functional.dropout(x, p=self.dropout, training=self.training))
x = x.add_(self.skip(orig) if self.skip is not None else orig)
x = x * self.skip_scale

if self.num_heads:
   q, k, v = self.qkv(self.norm2(x)).reshape(x.shape[0] * self.num_heads, x.shape[1]
       // self.num_heads, 3, -1).unbind(2)
   w = AttentionOp.apply(q, k)
   a = torch.einsum('nqk,nck->ncq', w, v)
   x = self.proj(a.reshape(*x.shape)).add_(x)
   x = x * self.skip_scale
return x
```

Group Normalization (Wu & He, 2018) is a normalization layer that divides channels into groups and normalizes the features within each group. It is a natural question what is the effect of Skip-Tuning under the impact of the group normalization layer? The UNetBlock takes the input of concatenation of linearly scaled features of skipped down-sampling parts and upsampling parts. The linear scaling will vanish after the first group normalization layer in UNetBlock with at most one exception group. However, the inner skip connection `x = x.add_(self.skip(orig)if self.skip is not None else orig)` maintains the information of Skip-Tuning.

We conduct an experiment to verify that the proposed Skip-Tuning is approximately equivalent to only changing the scale in `orig` variable. Specifically, we maintain the input of UNetBlock unchanged and multiply the scaling factor only on the corresponding channels of `orig` variable. We adopt the settings in Tab. 5, which achieves 1.57 FID score with 39 NFEs. In comparison, we do not observe a performance drop: only changing the scale in `orig` variable yields an FID score of 1.58.

We also experiment in another direction which only changes the scale of `self.norm(0)` variable and maintains the `orig` variable invariant. Surprisingly, we also do not observe a performance drop: only changing the scale in `self.norm(0)` variable yields an FID score of 1.57.

Remark C.1. FreeU (Si et al., 2023) adds an inflation coefficient (> 1) on the backbone features. The impact of the group normalization layer on FreeU is similar. Thus we speculate that the inflation coefficient also works on `orig` variable. From this viewpoint, the operations of Skip-Tuning and FreeU are different.

D. Additional Samples

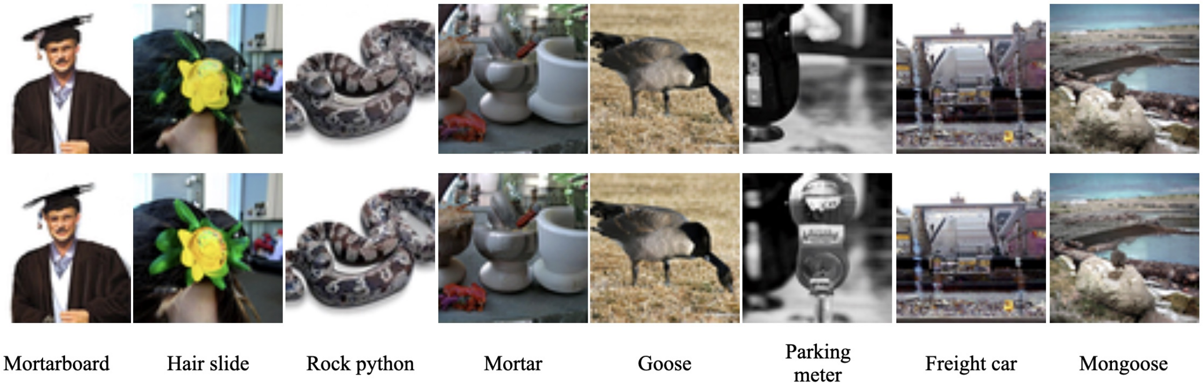


Figure 14. Image sampled from EDM model with ODE Heun sampling for 10 steps(19NFE). The random seed is set continuously from 33 to 40.

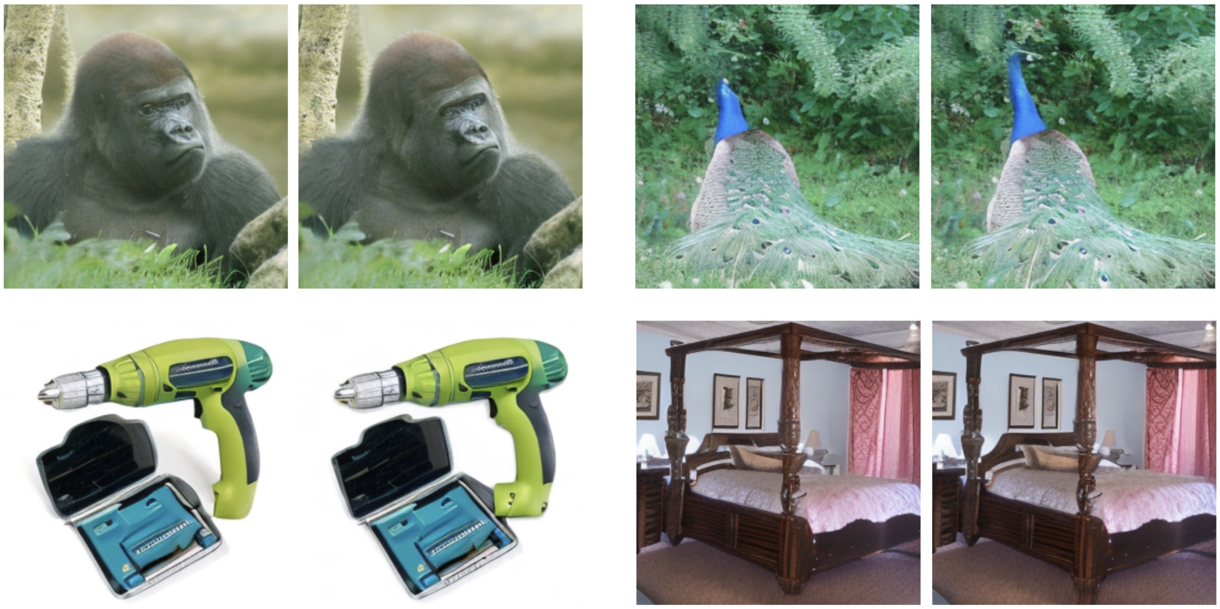


Figure 15. The left-hand side 256x256 figures are sampled from UViT 50steps(FID: 2.31), the right-hand side figures are sampled from UViT 50steps with $\rho = 0.82$ (FID: 2.21).

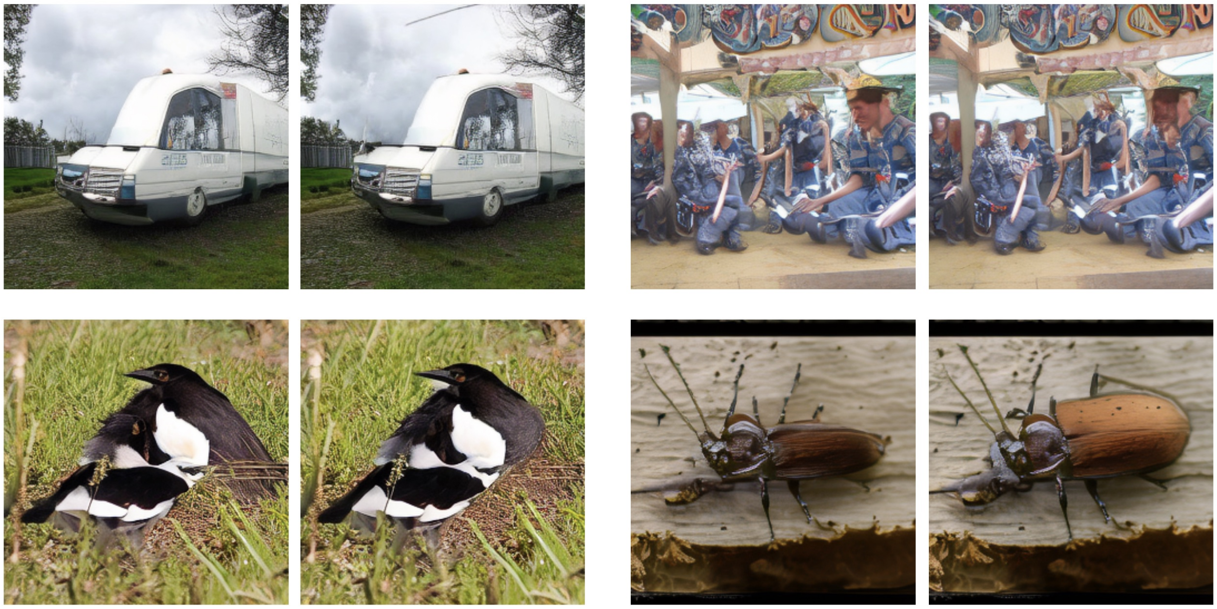


Figure 16. The left-hand side 256x256 figures are sampled from LDM 10steps(FID: 4.91), the right-hand side figures are sampled from LDM 10steps with $\rho = 0.95$ (FID: 4.67).

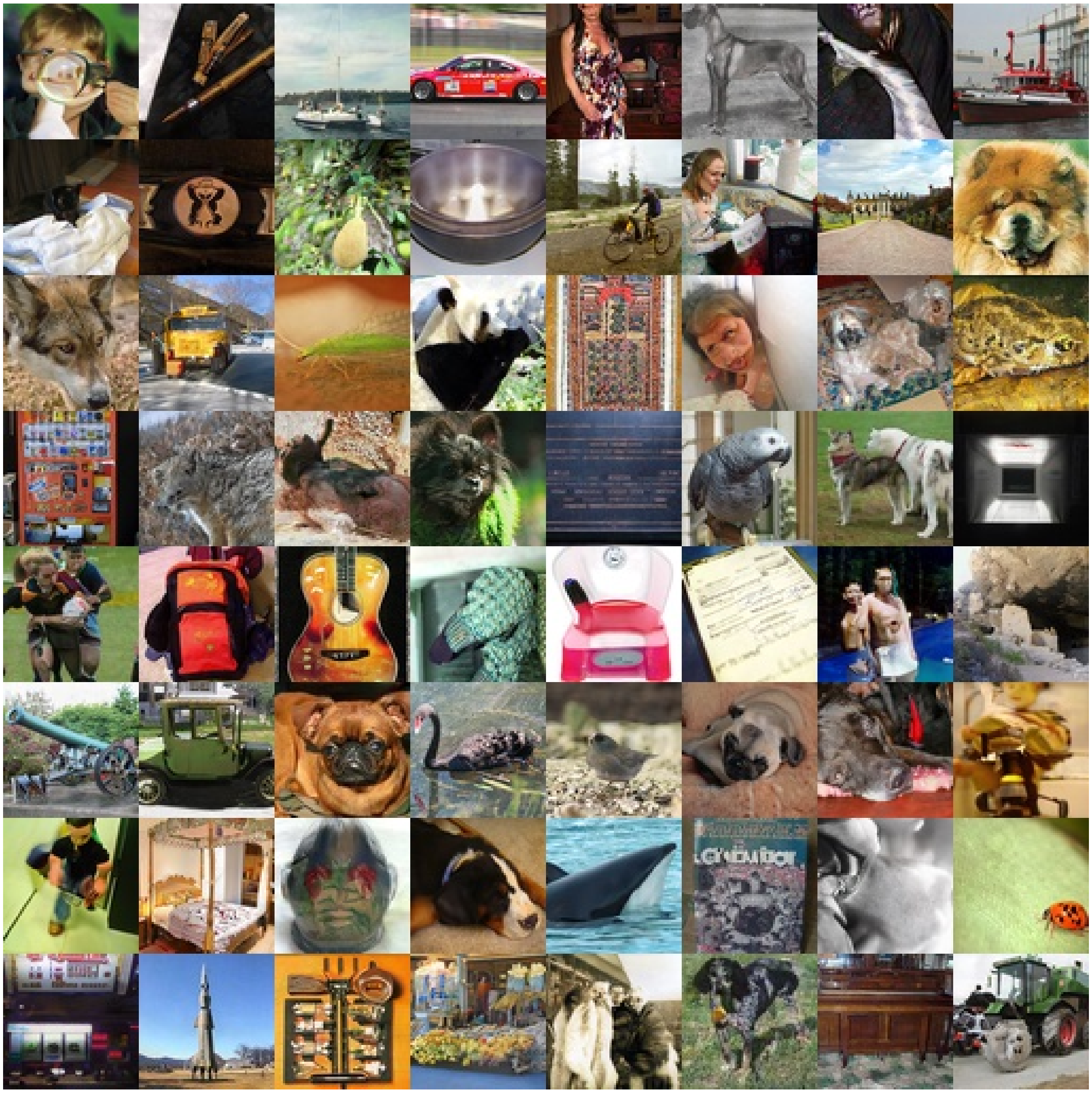


Figure 17. Image sampled from EDM model with NFE = 9 and ρ : 0.68 to 1.0 (FID = 2.92).

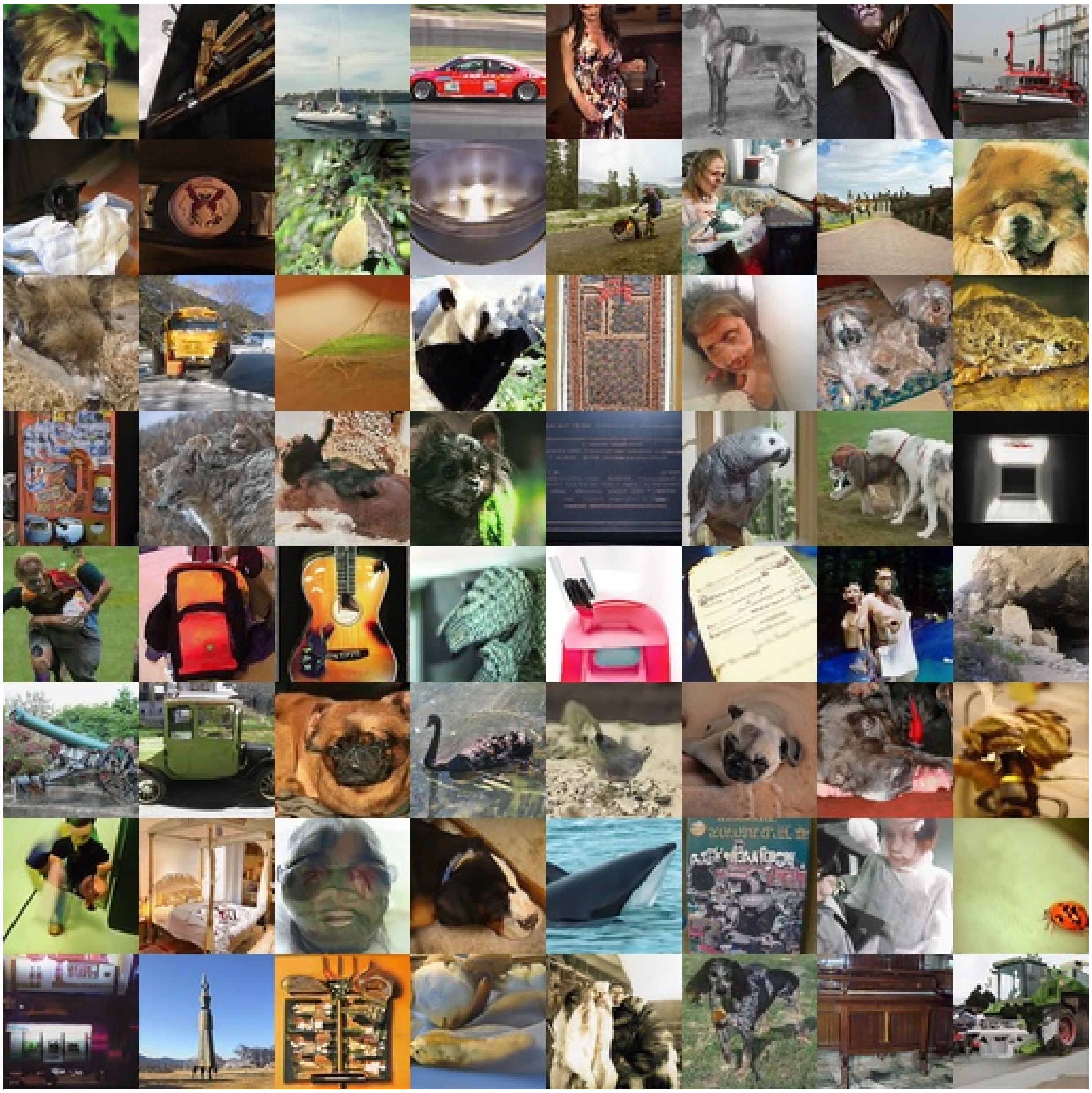


Figure 18. Image sampled from EDM model with NFE = 9 and ρ : 1.0 to 1.0 (FID = 5.88).

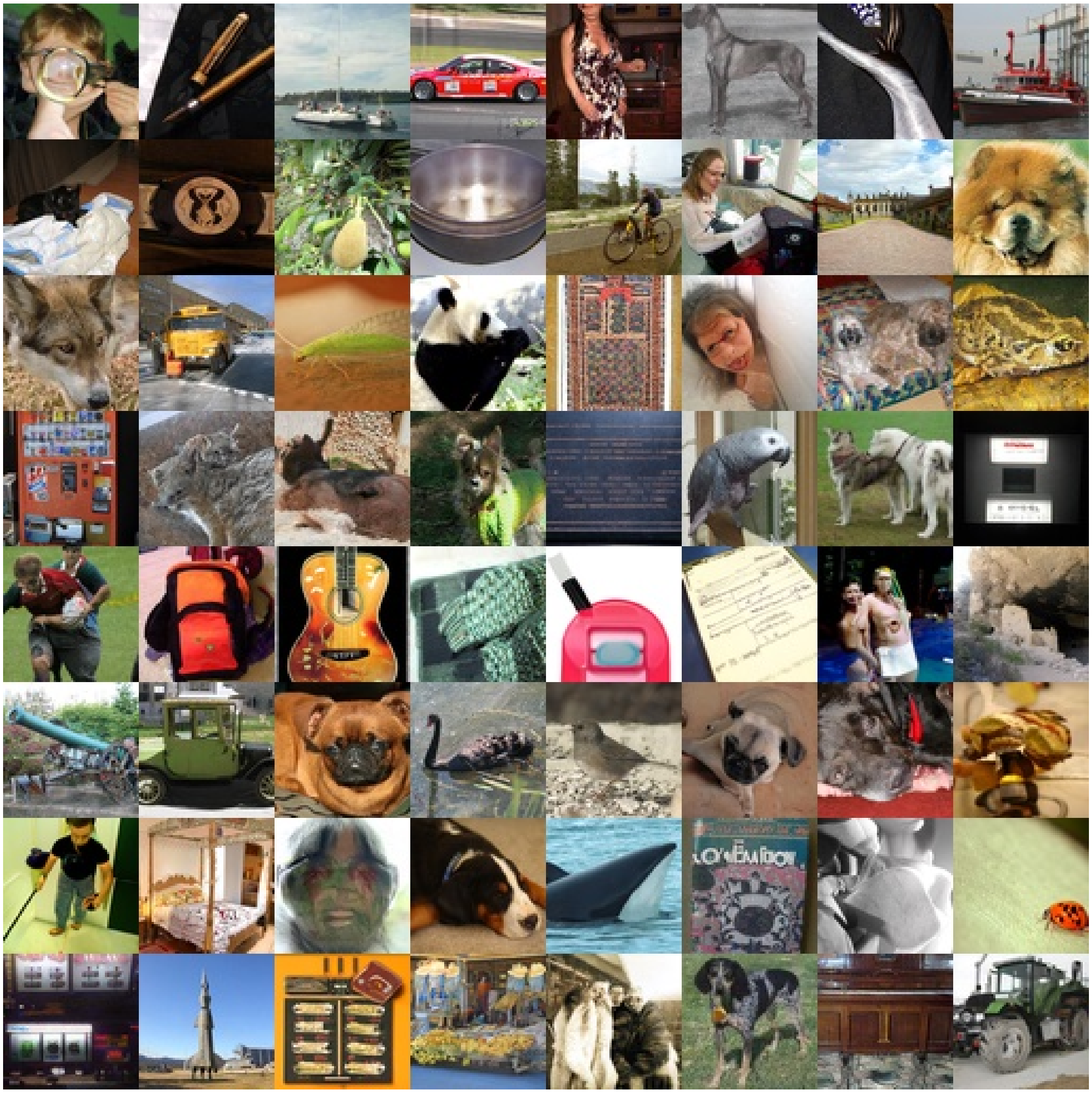


Figure 19. Image sampled from EDM model with NFE = 19 and ρ : 0.82 to 1.0 (FID = 1.75).

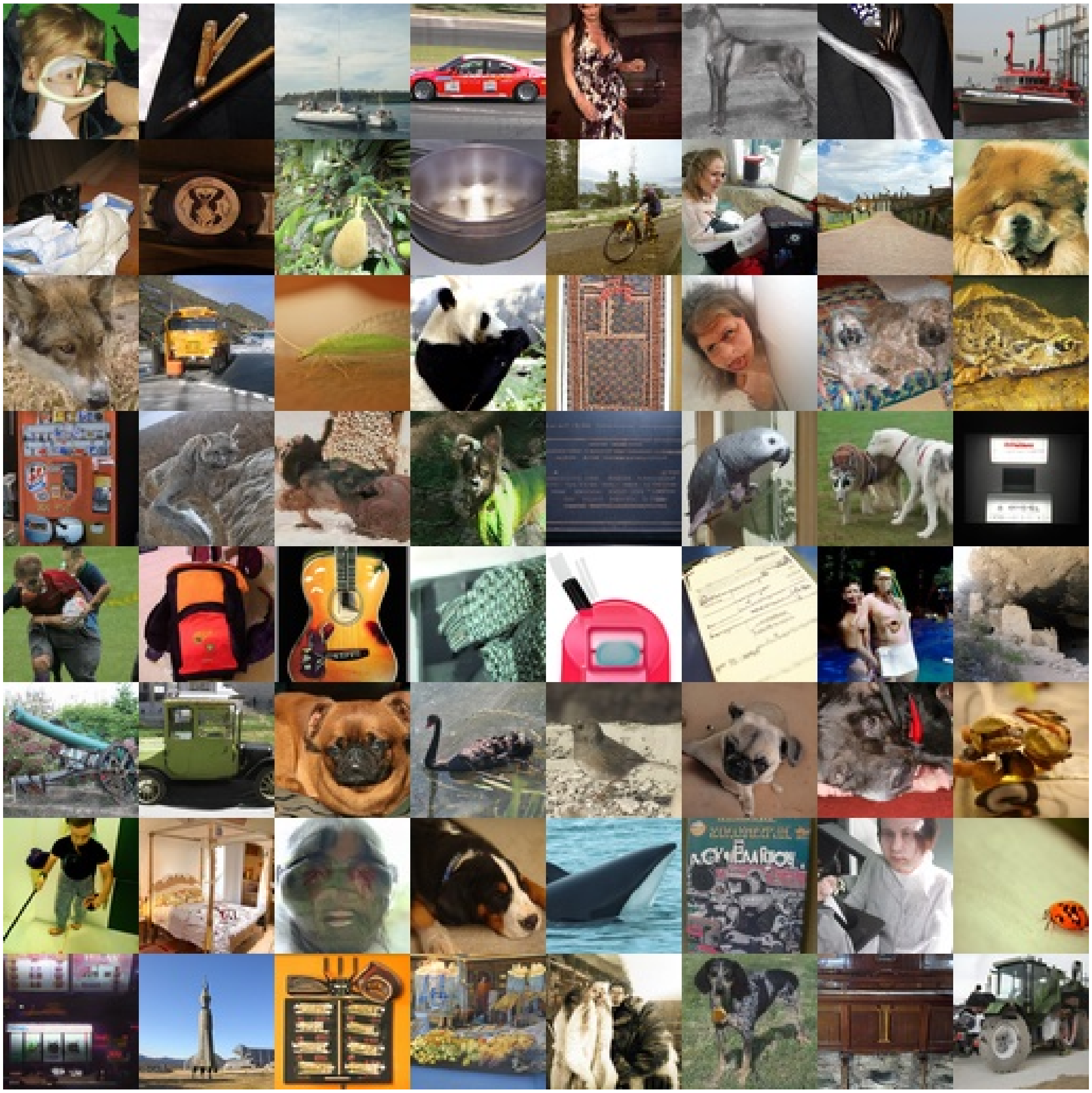


Figure 20. Image sampled from EDM model with NFE = 19 and ρ : 1.0 to 1.0 (FID = 2.60).

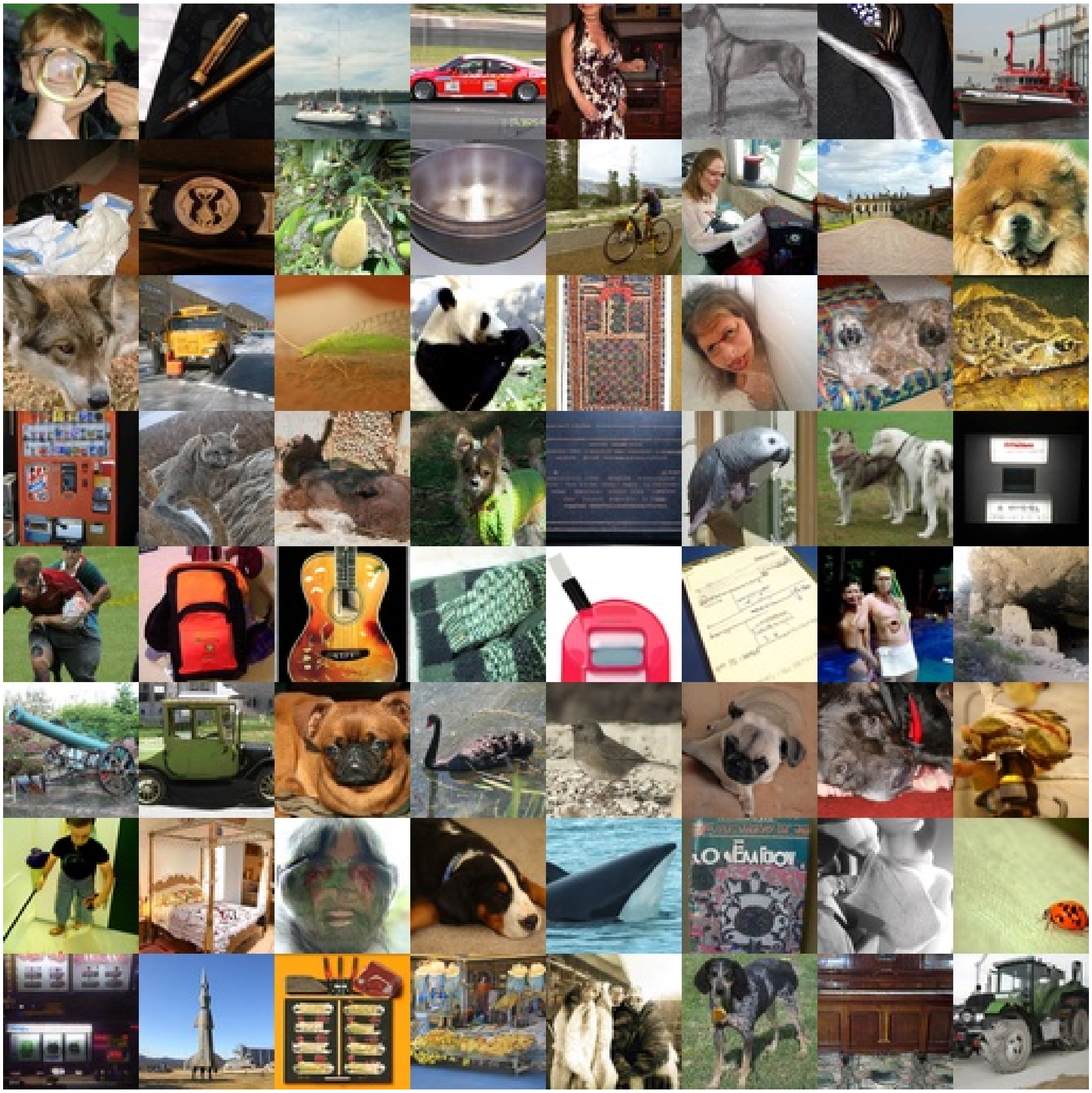


Figure 21. Image sampled from EDM model with NFE = 39 and ρ : 0.83 to 1.0 (FID = 1.57).

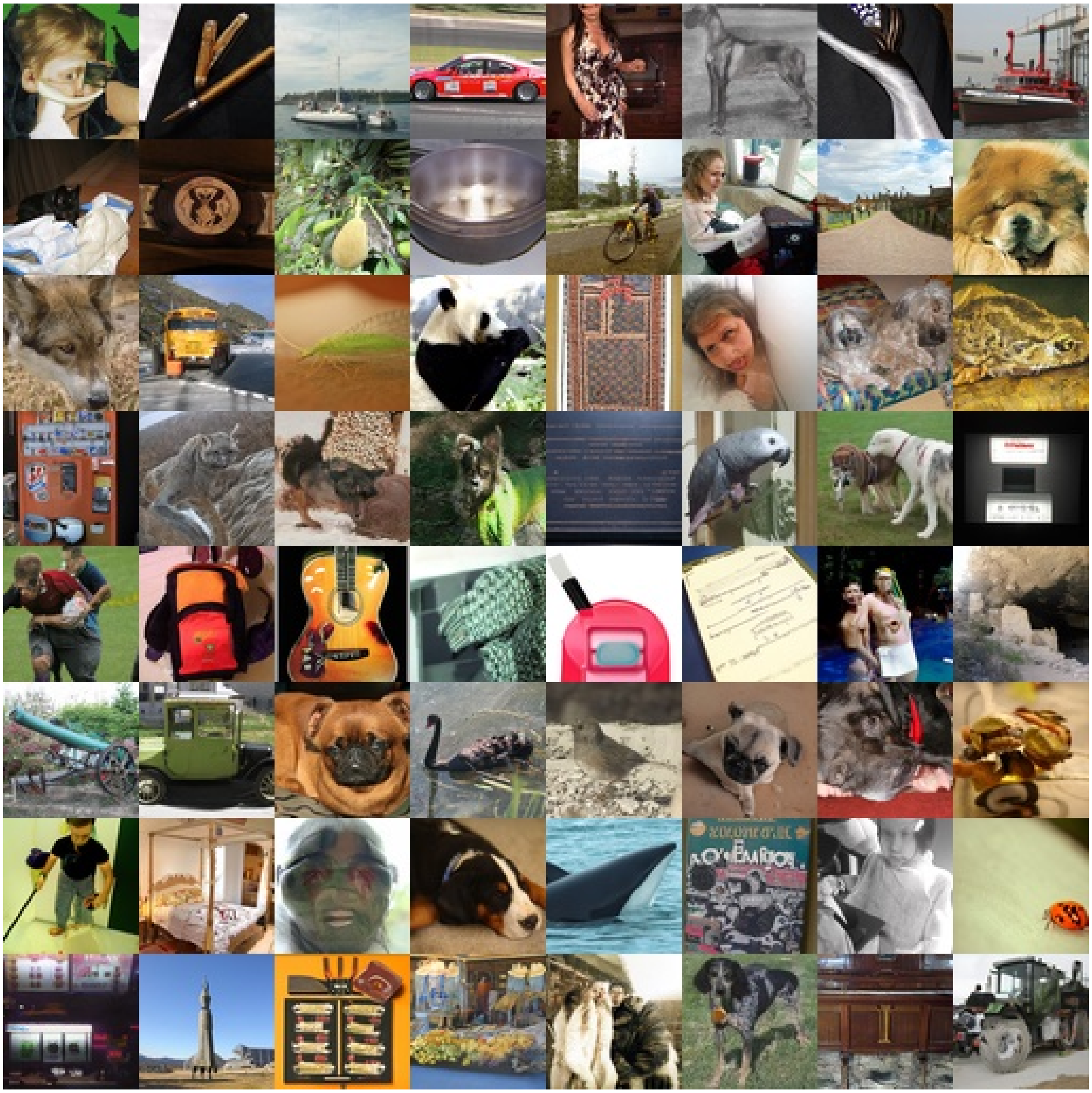


Figure 22. Image sampled from EDM model with NFE = 39 and ρ : 1.0 to 1.0 (FID = 2.21).