RESCALING INTERMEDIATE FEATURES MAKES TRAINED CONSISTENCY MODELS PERFORM BETTER

Junyi Zhu[†], **Zinan Lin**[‡], **Enshu Liu**[§], **Xuefei Ning**[§], **Matthew B. Blaschko**[†*]

ABSTRACT

In the domain of deep generative models, diffusion models are renowned for their high-quality image generation but are constrained by intensive computational demands. To mitigate this, consistency models have been proposed as a computationally efficient alternative. Our research reveals that post-training rescaling of internal features can enhance the one-step sample quality of these models without incurring detectable computational overhead. This optimization is evidenced by an obvious improvement in Frèchet Inception Distance (FID). For example, with our rescaled consistency distillation (CD) model, FID on the ImageNet dataset reduces from 6.2 to 5.2, on the LSUN-cat dataset from 10.9 to 9.5. Closer inspection of the generated images reveals that this enhancement may originate from improved visual details and clarity.

1 INTRODUCTION AND RELATED WORKS

Diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021; Rombach et al., 2021) have achieved unprecedented success across different generative tasks of multiple domains. Diffusion models progressively remove noise from random initialized samples, hence inducing high computational cost (time) during inference. Many works attempt to accelerate the inference of diffusion models by reducing the number of denoising iterations (Salimans & Ho, 2022; Lu et al., 2022; Liu et al., 2023b; Song et al., 2023). Specifically, Song et al. (2023) propose consistency models, which manage to map any point at any time of a probability flow (PF) trajectory to the trajectory's starting point in a single step and obtain competitive sample quality.

Prior work shows that the training objective of diffusion models is not directly related to their image quality (Liu et al., 2023a). Therefore, we anticipate the opportunity to improve the image quality by tweaking the parameters of *pre-trained* consistency models. While there exist many ways to do that, we observe a perhaps surprising phenomenon: *simply tuning two numbers that control the scale of intermediate features of pre-trained consistency models could improve the image quality*. In particular, the two scalars multiply the embedding feature and the features passing from encoder to decoder in the U-Net (Ronneberger et al., 2015) respectively. Intuitively, the introduced scalars regulate the intensity of conditioning and the reuse of input's high-resolution features, thereby affecting the generated images. The optimal rescaling scalars can be found through grid search with respect to image quality metrics such as FID using a small sampling batch. *This method does not involve any costly training process*.

Empirically, we find that our approach can improve the FID by a large margin. While FID may not always be consistent with the visual evaluation (Kirstain et al., 2023), our experimental results in Sections 3 and C show that after rescaling, we could indeed obtain more satisfying examples, i.e. examples showing more realistic features or fewer artifacts. Moreover, our work reveals a novel way to elicit new output of the off-the-shelf consistency models (e.g. the same object with different poses, see Fig. 7), beyond adjusting the initial noise.

2 METHOD

Our method rescales the intermediate features of the consistency models. To this end, we introduce two new scalars s_0, s_1 to multiply the output of the embedding layer and the encoder-decoder connections. Fig. 1 illustrates our modification. A code snippet is provided in §A. Our idea is simple but turns out to be effective. We find the best values of s_0, s_1 through grid search within a range around 1. For each combination of s_0 and s_1, we generate a few hunderds examples (500 for ImageNet, 200

*†: KU Leuven, ‡: Microsoft Research, §: Tsinghua University. Correspondence to jzhu@esat.kuleuven.be

for LSUN). Since only one-step inference is conducted, this grid search is cheap. For an 11×11 grid, the search can be done within a couple of hours using an NVIDIA TITAN XP.

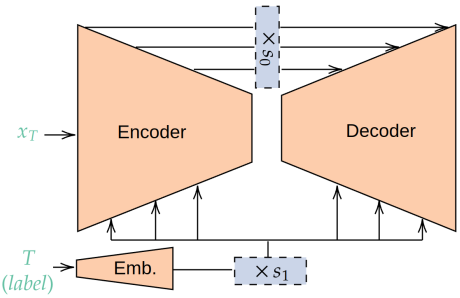


Figure 1: A schematic illustration of the rescaling operations.

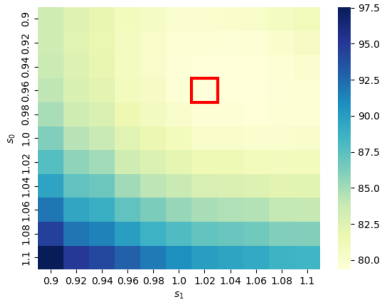


Figure 2: FIDs on ImageNet against varying combination of rescaling scalars. The red square highlights the lowest FID value.

3 Results

We set up the experiments using official codes and model checkpoints of consistency models (Song et al., 2023), more details of the experimental setup are given in §A.

Numerical Metrics against Rescaling. We first observe that the original configuration, i.e. $s_0 = 1, s_1 = 1$, usually does not achieve the lowest FID. Fig. 2 shows a heatmap of FIDs computed on ImageNet using a CD model. Tab. 1 provides a more comprehensive numerical comparison between the original models and the rescaled models, which are found through grid search on different datasets using FIDs. Selected scalars for rescaled models and more heatmaps are presented in §B.

	ImageNet			LSUN-cat			LSUN-bedroom		
	FID ↓	Prec. ↑	Rec. ↑	FID ↓	Prec. ↑	Rec. ↑	FID ↓	Prec. ↑	Rec. ↑
CD	6.19	0.68	**0.63**	10.88	0.65	0.36	8.20	**0.68**	0.31
rescaled CD	**5.21**	**0.71**	0.60	**9.52**	**0.66**	**0.38**	**7.57**	0.67	**0.35**
CT	12.83	0.71	**0.47**	20.64	0.56	0.24	16.03	0.59	0.16
rescaled CT	**11.07**	**0.78**	0.41	**19.99**	**0.63**	**0.25**	**13.62**	**0.72**	**0.17**

Table 1: Metrics computed using 50K examples. Best results are in bold.

Visual Quality against Rescaling. We further investigate the impact of rescaling on generation quality. Fig. 3 shows an example that rescaling assists in generating reasonable object. Additionally, we find that rescaling can elicit new output of the off-the-shelf models, see §C for more results.

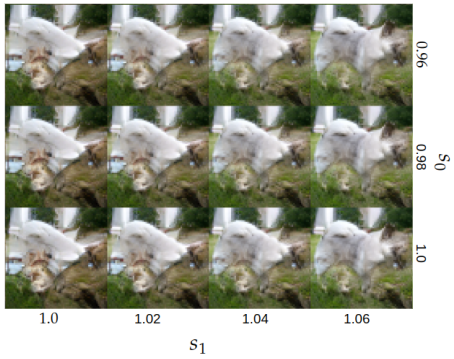


Figure 3: Images generated using varing rescaling scalars s_0, s_1. Zoom in to discern subtleties. These results are produced by the CD model, which was trained on the ImageNet dataset. At the original configuration ($s_0 = 1, s_1 = 1$), image generation fails to produce recognizable objects. Incrementally increasing s_1 reveals a distinct pug figure. We also observe that reducing s_0 aids in artifacts mitigation. An extended series of these images is displayed in Fig. 6.

4 Discussion and Conclusion

Our research demonstrates that rescaling the intermediate features within consistency models not only has the potential to enhance generation quality but also fosters a more diverse output. We propose two directions for future work: i) Applying different scalars to different layers and time steps (in case of few-step generation); ii) Extending our method's evaluation to encompass other frameworks, such as Rectified Flow (Liu et al., 2023b) and LCM (Luo et al., 2023), etc.

ACKNOWLEDGEMENT

This research received funding from the Flemish Government (AI Research Program) and the Research Foundation - Flanders (FWO) through project number G0G2921N.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. 2023.

Enshu Liu, Xuefei Ning, Zinan Lin, Huazhong Yang, and Yu Wang. OMS-DPM: Optimizing the model schedule for diffusion probabilistic models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 21915–21936. PMLR, 2023a.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023b.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan LI, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 5775–5787. Curran Associates, Inc., 2022.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 23–29 Jul 2023.

Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, abs/1506.03365, 2015.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

A EXPERIMENTS

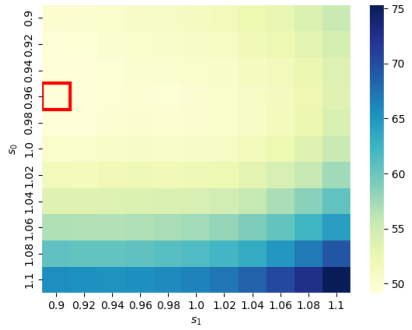
Our experiments are based on the official code and model checkpoints of consistency models (Song et al., 2023), which are available on GitHub (https://github.com/openai/consistency_models). We only evaluate checkpoints using the LPIPS metric (Zhang et al., 2018), which is claimed by Song et al. (2023) to be the optimal metric for both consistency distillation (CD) and consistency training (CT). Additionally, we evaluate the models trained on ImageNet (Deng et al., 2009) and LSUN datasets (Yu et al., 2015). To rescale the intermediate features of the consistency model, we need to modify the `forward` function of the `UNetModel` class in `unet.py`. As shown in the following code snippet A, rescaling operations are conducted in Line 19 and Line 23.

When comparing numerical metrics and visual quality against different combinations of rescaling scalars, *we always use the same initial noises (and classes in the conditioned setting) to generate images.*

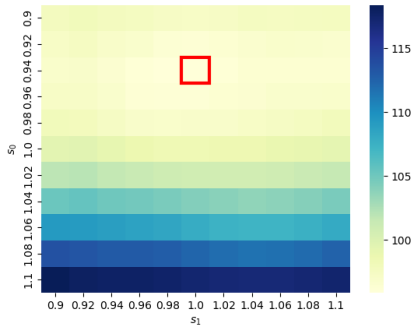
```
1  def forward(self, x, timesteps, y=None, scalar0=1., scalar1=1.):
2      """
3      Apply the model to an input batch.
4
5      :param x: an [N x C x ...] Tensor of inputs.
6      :param timesteps: a 1-D batch of timesteps.
7      :param y: an [N] Tensor of labels, if class-conditional.
8      :return: an [N x C x ...] Tensor of outputs.
9      """
10     assert (y is not None) == (
11         self.num_classes is not None
12     ), "must specify y if and only if the model is class-conditional"
13
14     hs = []
15     emb = self.time_embed(timestep_embedding(timesteps, self.
       model_channels))
16     if self.num_classes is not None:
17         assert y.shape == (x.shape[0],)
18         emb = emb + self.label_emb(y)
19     emb.mul_(scalar1)
20     h = x.type(self.dtype)
21     for module in self.input_blocks:
22         h = module(h, emb)
23         hs.append(h * scalar0)
24     h = self.middle_block(h, emb)
25     for module in self.output_blocks:
26         h = th.cat([h, hs.pop()], dim=1)
27         h = module(h, emb)
28     h = h.type(x.dtype)
29     return self.out(h)
```

	ImageNet		LSUN-cat		LSUN-bedroom	
	s_0	s_1	s_0	s_1	s_0	s_1
rescaled CD	0.96	1.02	0.94	1.0	0.96	0.9
rescaled CT	1.0	1.36	0.84	1.1	0.8	1.0

Table 2: Scalars selected for the experiments in Tab. 1.



(a) FIDs on LSUN-bedroom.

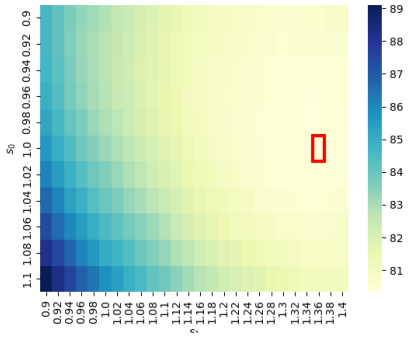


(b) FIDs on LSUN-cat.

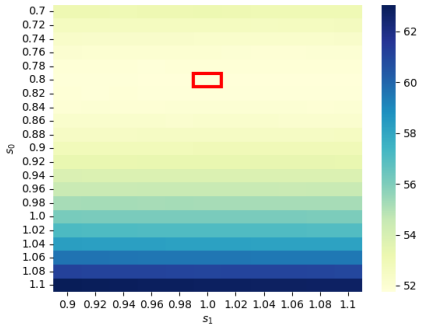
Figure 4: FIDs of CD model against varying combinations of rescaling scalars. The red square highlights the lowest FID value.

B More Results on Numerical Metrics

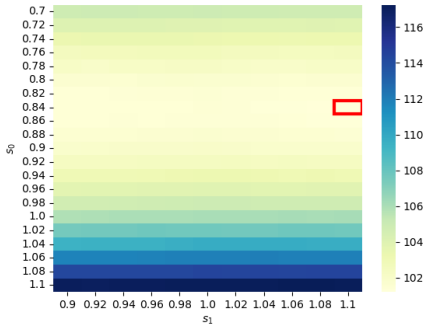
For CD models, we find that it is efficient to search within range of $[0.9, 1.1]$ for both s_0 and s_1. The results of heatmaps are shown in Figs. 2 and 4. For CT models, we find it is beneficial to extend the search range, the heatmaps are shown in Fig. 5. Using the best combination of s_0 and s_1 (see Tab. 2), we numerically evaluate rescaled CD models and rescaled CT models, and results are recorded in Tab. 1. Overall, the rescaled models achieve better metrics than the original models.



(a) FIDs on ImageNet.



(b) FIDs on LSUN-bedroom.



(c) FIDs on LSUN-cat.

Figure 5: FIDs of CT model against varying combinations of rescaling scalars. The red square highlights the lowest FID value.

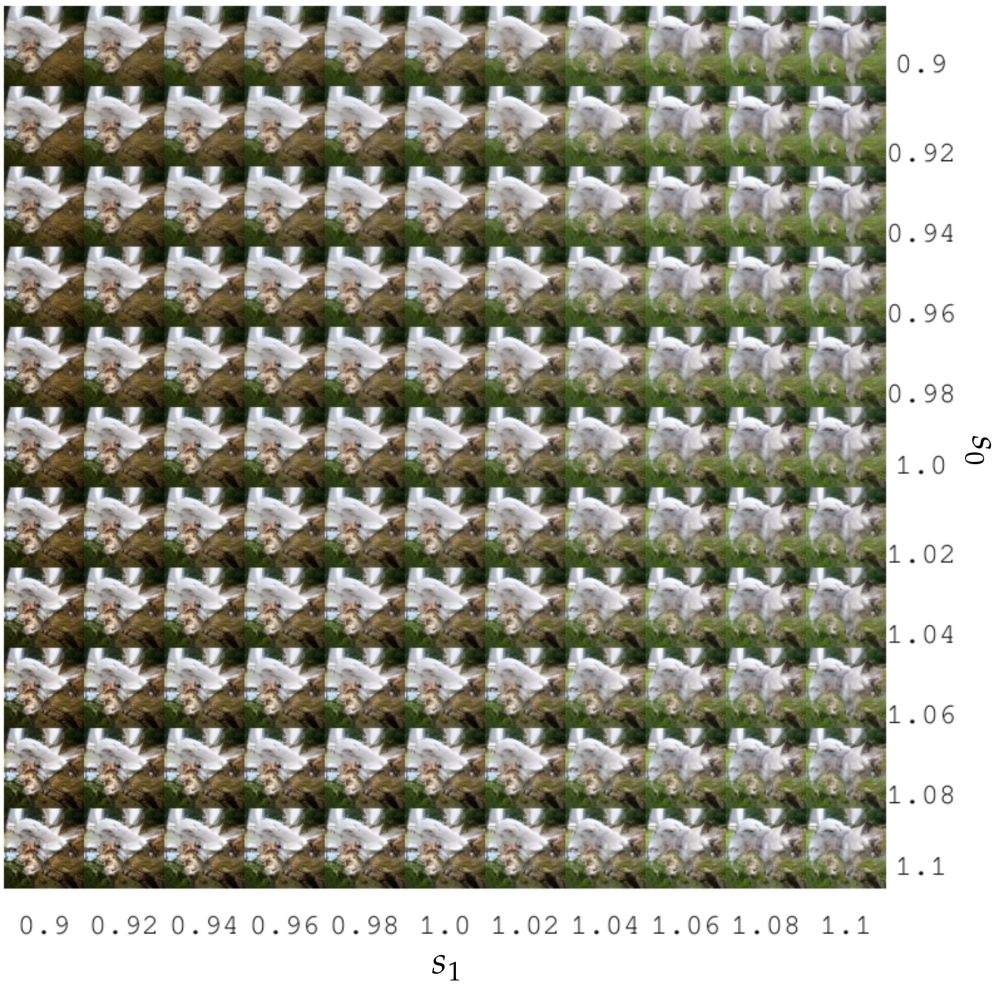


Figure 6: Images generated by a CD model trained on the ImageNet dataset.

C More Results on Visual Quality

In this section, we demonstrate the impact of rescaling on visual quality using image examples.

Figure 6 illustrates that scaling up s_1 aids in generating the desired objective in the conditioned setting. At the original configuration ($s_0 = 1, s_1 = 1$), there is no discernible object. However, as s_1 increases, a pug figure emerges. It is noteworthy that the condition for this image is "Norwegian elkhound" instead of "pub", but scaling up s_1 still pushes the generation result towards the specific condition.

Fig. 7 shows that scaling up s_1 results in a new outcome, as a dog appears to be turning its head from its right side towards the screen. Interestingly, it seems to be the same dog in different poses.

Fig. 8 shows that the generated cat's left eye is corrupted at the original configuration ($s_0 = 1, s_1 = 1$). However, adjusting towards smaller s_0 and s_1 values (e.g., $s_0 = 0.9, s_1 = 0.9$) resolves this issue.

Fig. 9 reveals that transitioning from the configuration ($s_0 = 0.9, s_1 = 0.9$) to ($s_0 = 1.1, s_1 = 1.1$) causes the bedside lamp to dim.

Effects of the rescaling operation: We inspect several hundred generation results and our observations can be summarized as follows: i) We note that decreasing s_0 can mitigate artifacts. Compared to the row where $s_0 = 1.1$, images in the row with $s_0 = 0.9$ exhibit smoother textures in both

the object and the background; ii) For the conditioned setting, e.g. ImageNet, tuning up s_1 appears to emphasize the condition in the generated result, e.g. the condition becomes clearer (see Fig. 6) or the object adopts a facing-forward pose (see Fig. 7); iii) The configurations yielding the lowest FIDs do not always produce the visually best images. However, the regions of high visual quality in the image grid generally correspond to the regions of low FID values in the corresponding heatmaps; iv) The best scaling scalars are different across datasets, which is reasonable since the feature distribution is not consistent across datasets.

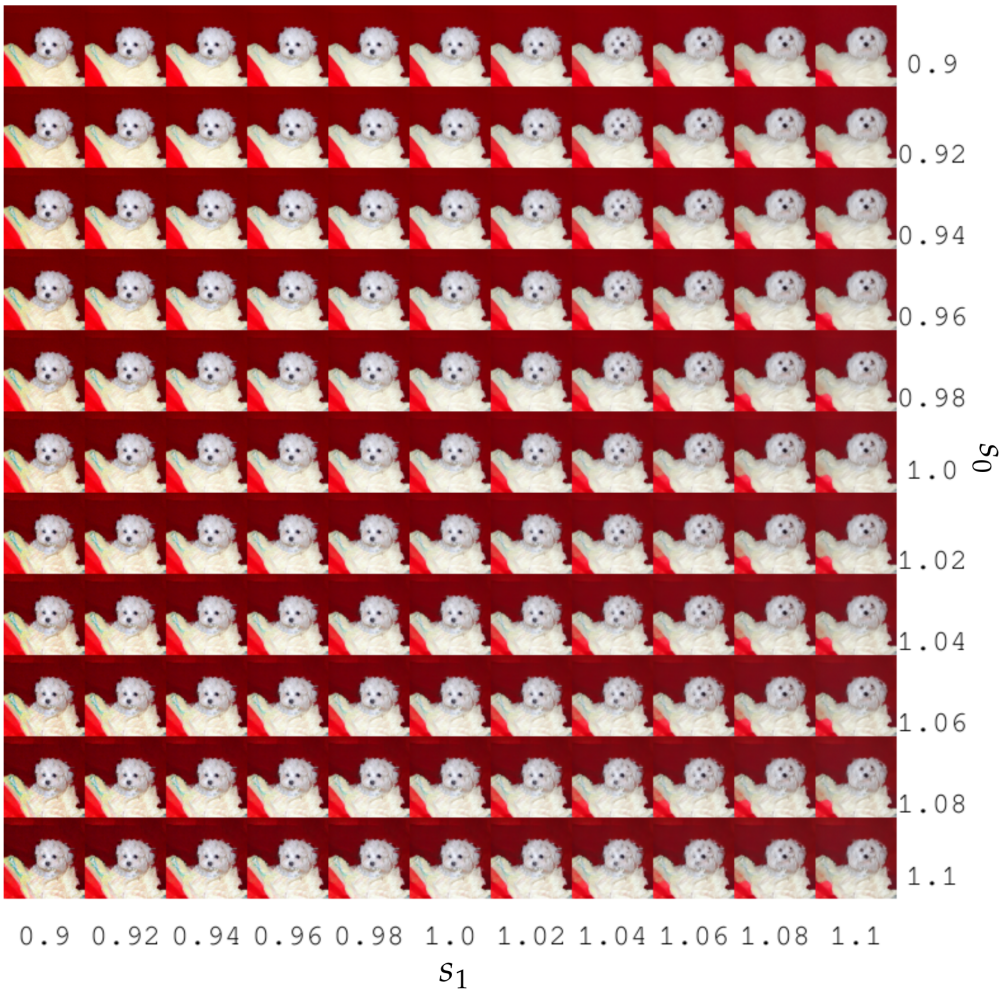


Figure 7: Images generated by a CD model trained on the ImageNet dataset.

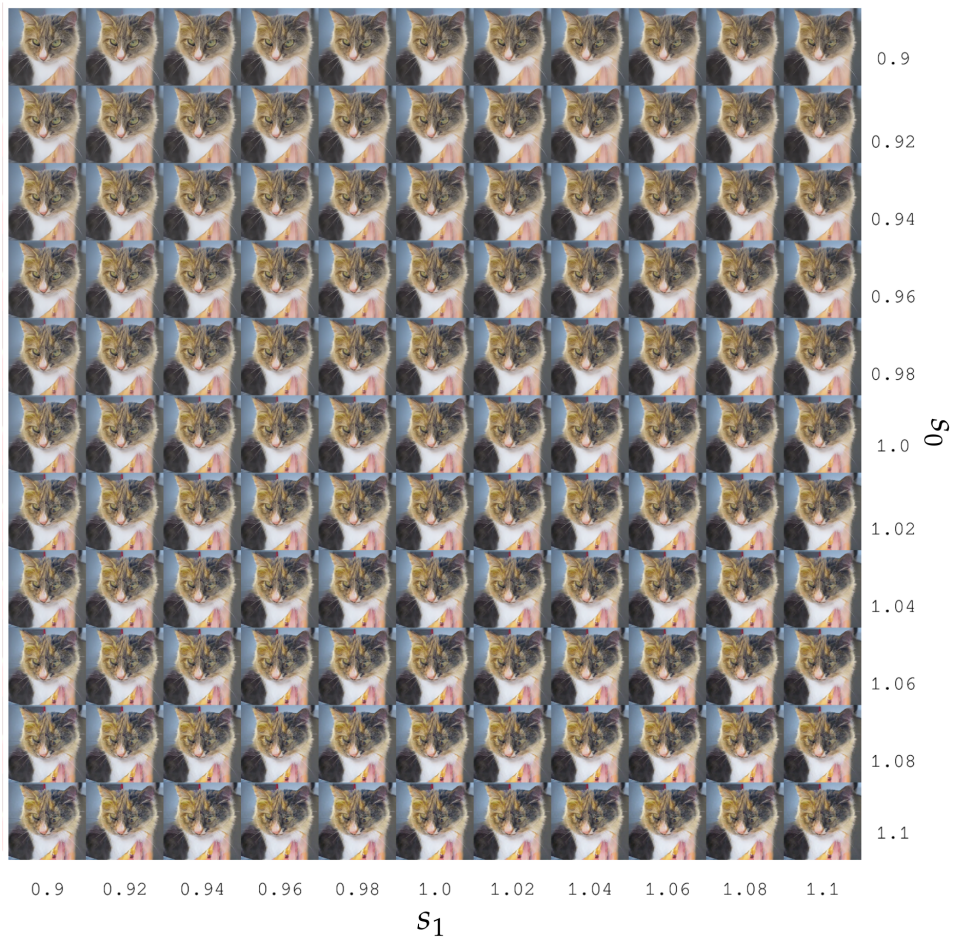


Figure 8: Images generated by a CD model trained on the LSUN-cat dataset.

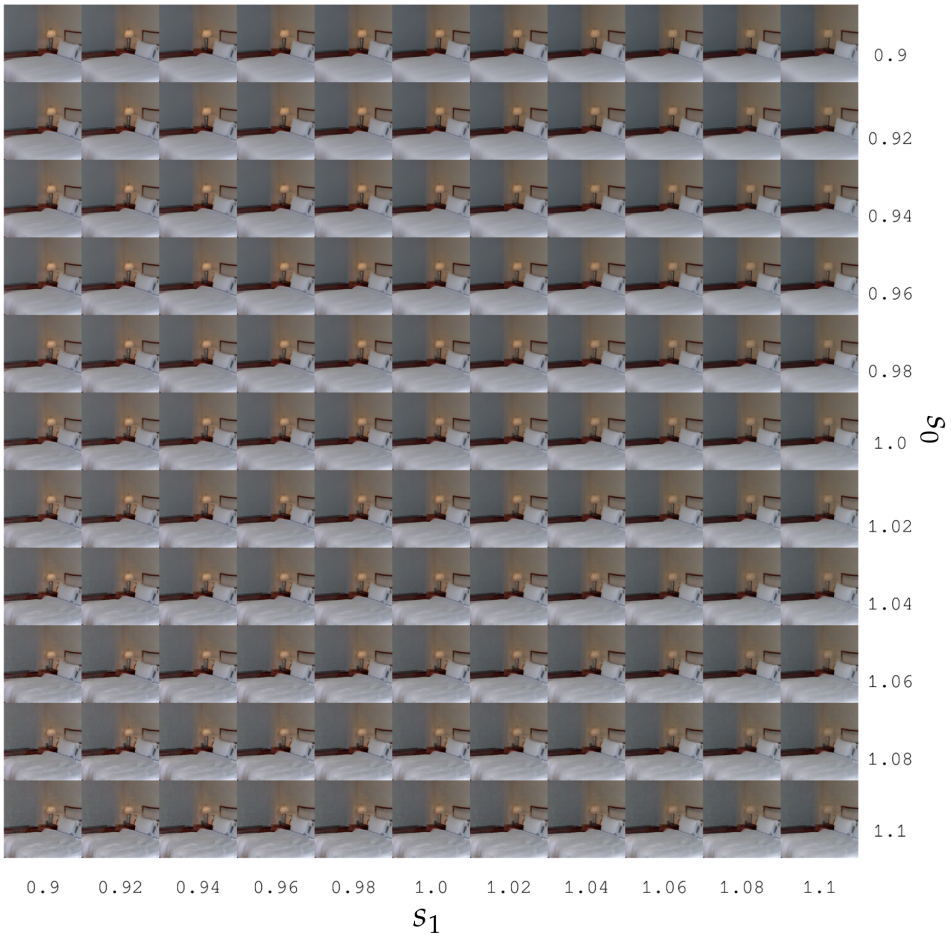


Figure 9: Images generated by a CD model trained on the LSUN-bedroom dataset.