Final Presentation of SDS Lab Internship(2024 Winter).

포항공과대학교 산업경영공학과

Statistics and Data Science Lab.

2025/02/06 Jaewon Seo

Outline

1. Introduction
2. Main theme- Diffusion model
3. Experiment- Improving diffusion model by adjusting features of guidance network and its sampling methods
4. Discussion

What I did through internship period

Weekly progress

- ✓ Week 5, December: read paper of various topics- Diffusion model, Wasserstein distance, VAE, etc
- ✓ Week 1~3, January: Selected main topic to study deeper- Diffusion model, read related papers
- ✓ Week 4~5, January: Try toy example in paper and revise it for better output

Papers I read

Papers

- ✓ Ho, J., Jain, A., Abbeel, P. (2020). *Denoising Diffusion Probabilistic Models*. arXiv preprint arXiv:2006.11239.
- ✓ Luo, C. (2022). *Understanding Diffusion Models: A Unified Perspective*. arXiv preprint arXiv:2208.11970.
- ✓ Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B. (2020). *Score-Based Generative Modeling through Stochastic Differential Equations*. arXiv preprint arXiv:2011.13456.
- ✓ Wang, X., Dufour, N., Andreou, N., Cani, M., Abrevaya, V.F., Picard, D., Kalogeiton, V. (2024). *Analysis of Classifier-Free Guidance Weight Schedulers*. arXiv preprint arXiv:2404.13040.
- ✓ Karras, T., Aittala, M., Kynkäänniemi, T., Lehtinen, J., Aila, T., Laine, S. (2024). *Guiding a Diffusion Model with a Bad Version of Itself*. arXiv preprint arXiv:2406.02507.
- ✓ Ho, J., Salimans, T. (2022). *Classifier-Free Diffusion Guidance*. arXiv preprint arXiv:2207.12598.

Guidance and CFG

Guidance

- ✓ Guidance: Intended effect is to decrease the diversity of the samples while increasing the quality of each individual sample
- ✓ Uses the modified score function on sampling

$$\nabla_x \log p_w(x \mid c; \sigma) = \nabla_x \log p_1(x \mid c; \sigma) + (w-1)\nabla_x \log \frac{p_1(x \mid c; \sigma)}{p_0(x \mid c; \sigma)}$$

Classifier free guidance(CFG)

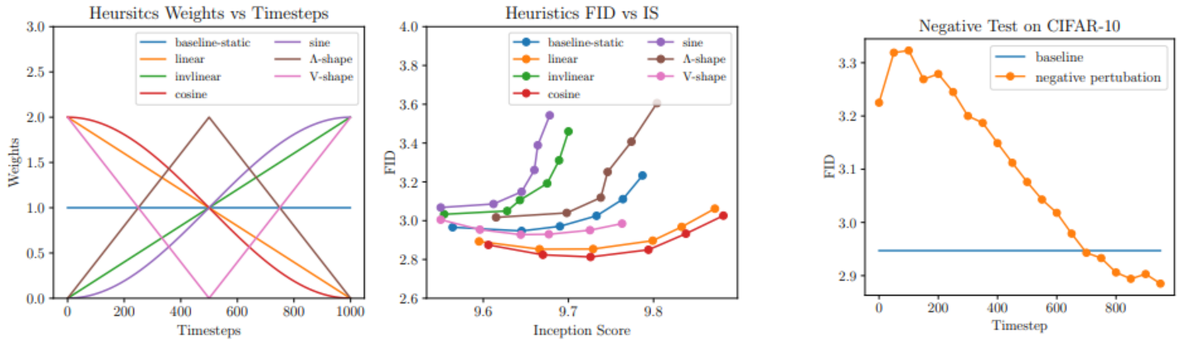
- ✓ Classifier free guidance: Uses conditional model for main network and unconditional model for guidance network
- ✓ Push samples to region that has higher probability density
- ✓ Expected to improve the quality of sample but has various drawbacks

(e.g. over-saturation of colors, limited canonical images, overly simplified images…)

Improve CFG by adjusting weight schedulers

Weight schedulers

- ✓ Increasing guidance monotonically(linear, cosine) improved fidelity and diversity of image
  - ✓ Can be obtained by adjusting one line in code
- ✓ Too much guidance at the beginning of the denoising process is harmful



(a) Experiment on Different Heuristics (b) Negative Perturbation

Figure 6: **Preliminary Analysis on CIFAR-10 (a) Various heuristic curves** with their FID vs. IS performances. **(b) Negative perturbation** by setting the guidance scale to 0 across distinct intervals while preserving static guidance to the rest. By eliminating the weight at the **initial stage** (e.g., $T = 800$), the lowered FID shows an enhancement, whereas removing guidance at higher timesteps leads to worse FID.

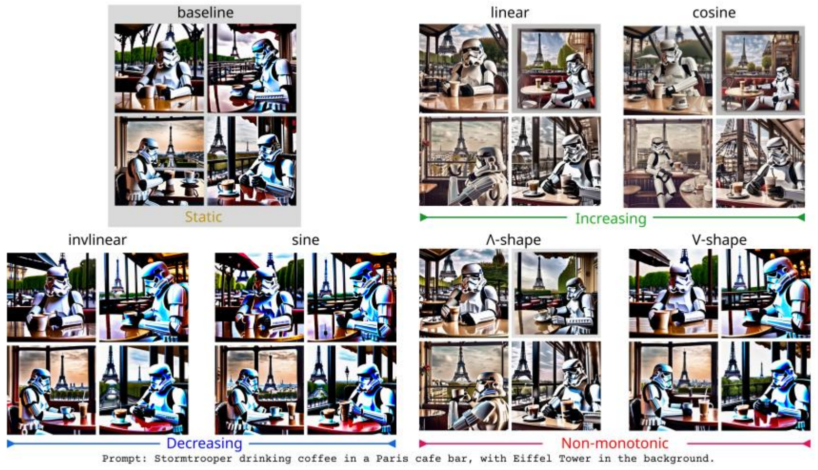
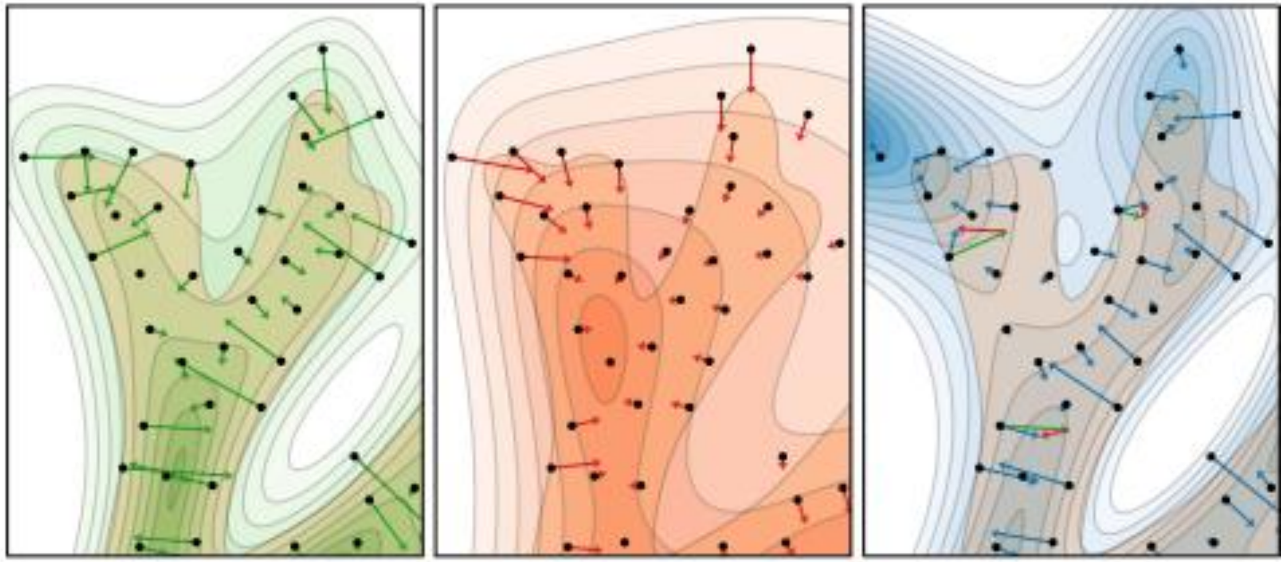


Figure 2: **Examples of all heuristics** on SDXL. Increasing ones (*linear* and *cosine*) enhance fidelity, textual adherence and diversity.

Autoguidance

Autoguidance

- ✓ Use guidance network as bad version(less trained) of main network(less iteration, less layer, etc)
- ✓ It does not require additional train of unconditional model
- ✓ Less trained model shows less trained pdf(not matching with ground truth distribution)

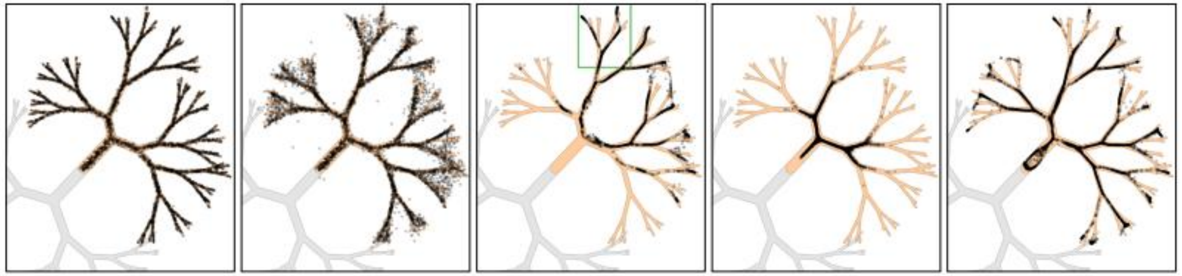


(a) $p_1(\mathbf{x}|\mathbf{c};\sigma_{\text{mid}})$ (b) $p_0(\mathbf{x};\sigma_{\text{mid}})$ (c) Ratio p_1/p_0

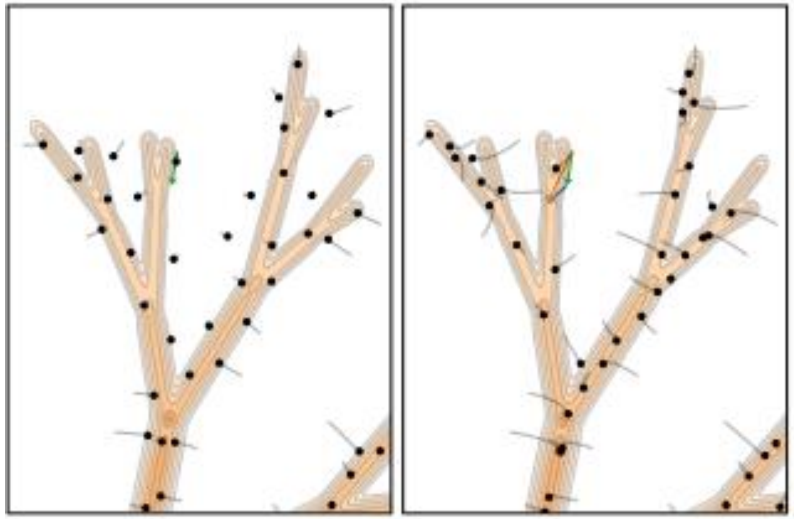
Autoguidance

Autoguidance

- ✓ Using guidance network shows more samples located on manifold comparing to model without guidance
- ✓ Comparing to CFG, Autoguidance shows widely distributed sample while showing less outliers(which is not located on branch)
- ✓ It shows that model with guidance pushes samples to the branch manifold(high density region)



(a) Ground truth (b) No guidance (c) Classifier-free guidance (d) Naive truncation (e) Autoguidance (ours)



(d) No guidance (e) CFG with $w = 4$

Autoguidance

Selecting guidance network

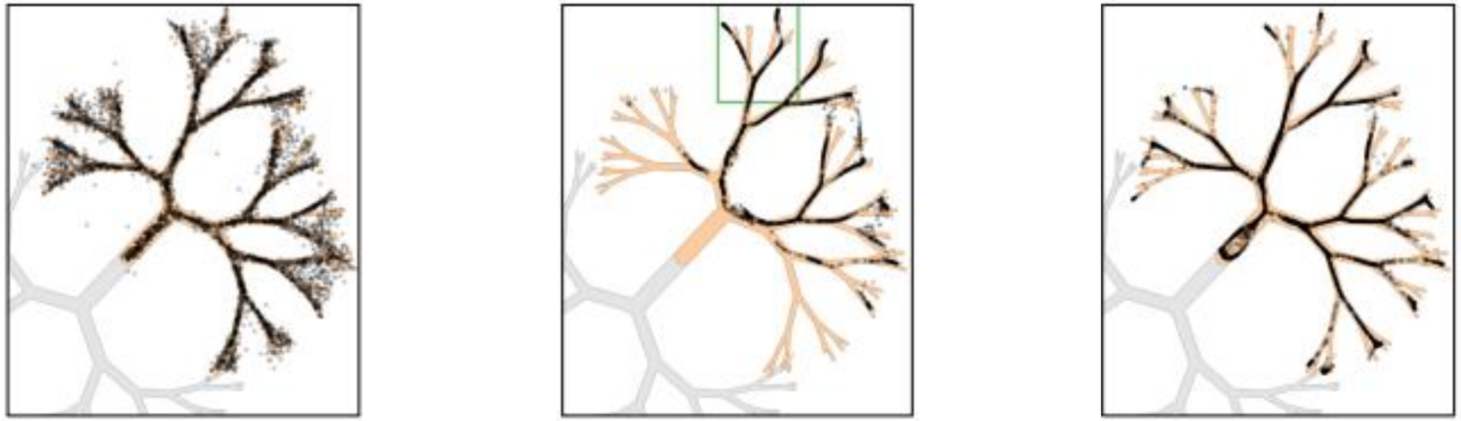
- ✓ Using guidance network pushes samples near the branch manifold comparing to model without guidance
- ✓ Comparing to CFG, Autoguidance shows samples are distributed evenly(also located on minor branch region)
- ✓ Both CFG and Autoguidance shows more samples are located on branch manifold
- ✓ It shows that model with guidance pushes samples to the branch manifold(high density region)

$$p_w(x \mid c;\sigma) \propto p_1(x \mid c;\sigma)\left(\frac{p_1(x \mid c;\sigma)}{p_0(x \mid c;\sigma)}\right)^w$$

Experiment

Current problem

- ✓ No guidance: Samples are located on non-manifold area
- ✓ CFG: Samples are concentrated on specific branches → Lack of diversity
- ✓ Autoguidance: Samples seems to be uniformly distributed among branches but still has concentration on specific branches



(b) No guidance　　(c) Classifier-free guidance　　(e) Autoguidance (ours)

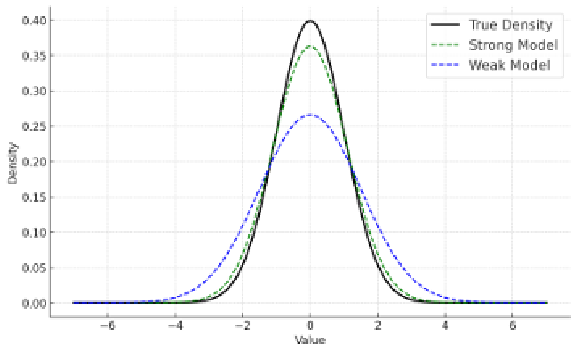
Autoguidance

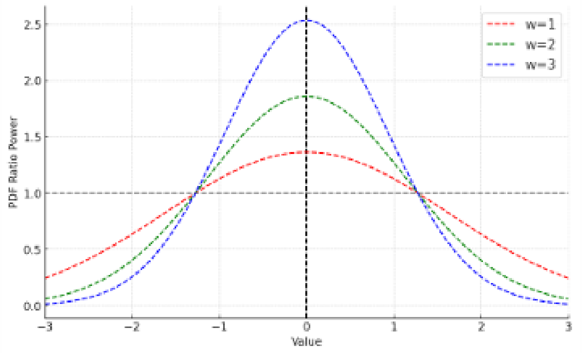
Motivation

- Autoguidance suggest a p_0 as degraded p_1 which close enough to p_1 but informative.

$$p_w(\mathbf{x} \mid \mathbf{c}; \sigma) \propto p_1(\mathbf{x} \mid \mathbf{c}; \sigma) \left(\frac{p_1(\mathbf{x} \mid \mathbf{c}; \sigma)}{p_0(\mathbf{x} \mid \mathbf{c}; \sigma)} \right)^w$$

- (Opinion) Autoguidance boosted performance by emphasizing the distribution's mode.
  - They choose p_1 and p_0 targeted same distribution p_{data}. i.e. share same modes.
  - Due to differences in quality, the ratio was set to highlight areas around the modes.





- (Opinion) Error correction is just the result of emphasizing mode.
- All we have to do is find p_0 more spread than p_1 while targeting the same distribution.

Autoguidance

HeavyTailGuidance

- I suggest p_0 with perturbed with heavy tail noise, to spread out the distribution than p_1.

$$\frac{p_1(\mathbf{x} \mid \mathbf{c}; \sigma)}{p_0(\mathbf{x} \mid \mathbf{c}; \sigma)} = \frac{p_{\text{data}}(\mathbf{x}|\mathbf{c}) * \mathcal{N}(\mathbf{x}; \mathbf{0}, \sigma^2\mathbf{I})}{p_{\text{data}}(\mathbf{x}|\mathbf{c}) * \mathcal{T}_\nu(\mathbf{x}; \mathbf{0}, \sigma'^2\mathbf{I})} \text{ or } \frac{p_{\text{data}}(\mathbf{x}|\mathbf{c}) * \mathcal{N}(\mathbf{x}; \mathbf{0}, \sigma^2\mathbf{I})}{p_{\text{data}}(\mathbf{x}|\mathbf{c}) * \mathcal{N}(\mathbf{x}; \mathbf{0}, \sigma^2_{\text{larger}}\mathbf{I})}$$

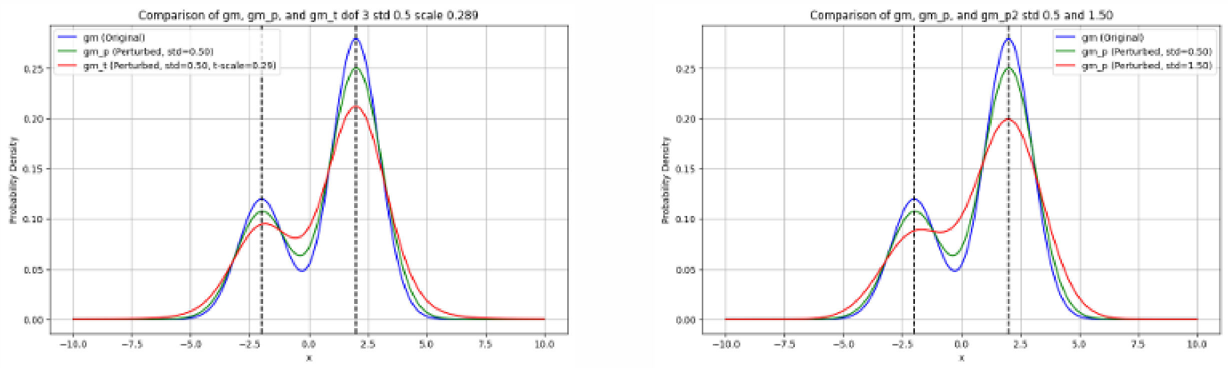


Figure 5: blue : $p_{\text{data}}(\mathbf{x}|\mathbf{c})$, green : numerator and red : denominator

- It requires computation to train the model, but the following advantages are expected.
  1. Control the spread out level, directly : tail's order and noise level.
  2. Heavy-tailed model can capture the entire distribution, but degraded model can't.

Experiment

Motivations and expectations

- ✓ It has been shown that increasing guidance monotonically improves sample quality
- ✓ Heavy tail guidance are expected to improve sample quality
  - ✓ Heavy tail guidance allows do not requires extra model for guidance network
- ✓ Use methods above to improve sample quality

Experiment

Sigma Scaling

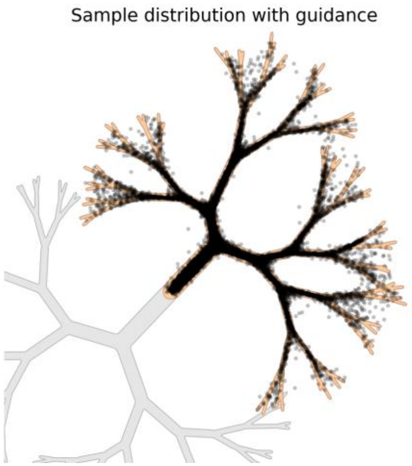
- ✓ For monotonically decreasing sigma of main model, set sigma of guidance model as multiplying constant to sigma of main model
- ✓ Adjust constant k and observe the change of output

$$\sigma_{guidance} = k\sigma$$

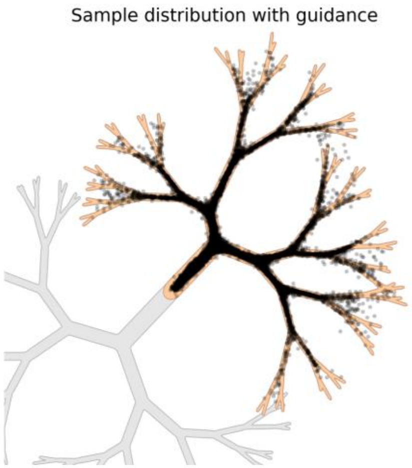
- ✓ As scaling factor k increases, sample tend to be concentrated around the main branch (high density region)
- ✓ Small k reduces outliers and still preserve diversity comparing to model without guidance

Experiment

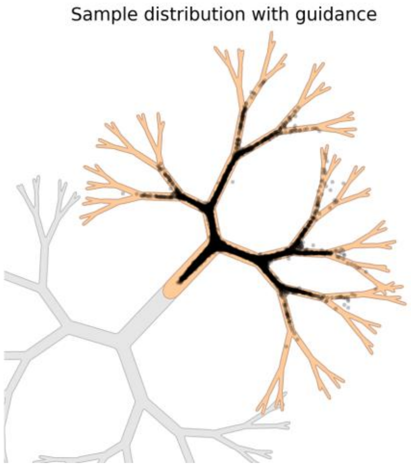
<k=1.00>



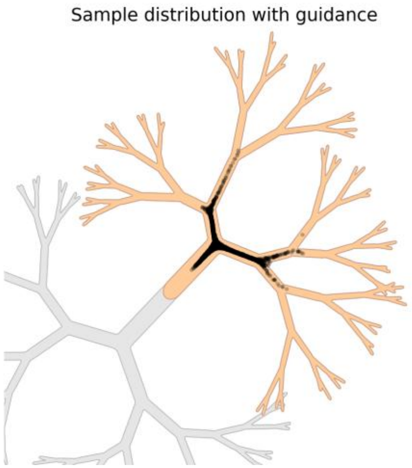
<k=1.02>



<k=1.04>



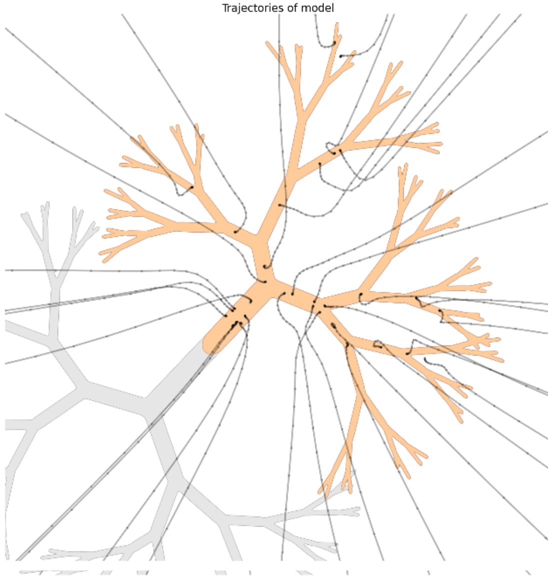
<k=1.10>



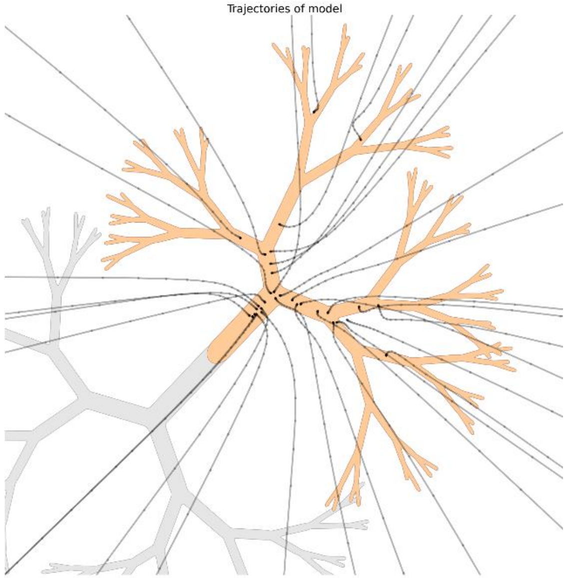
<k=1.20>

Experiment

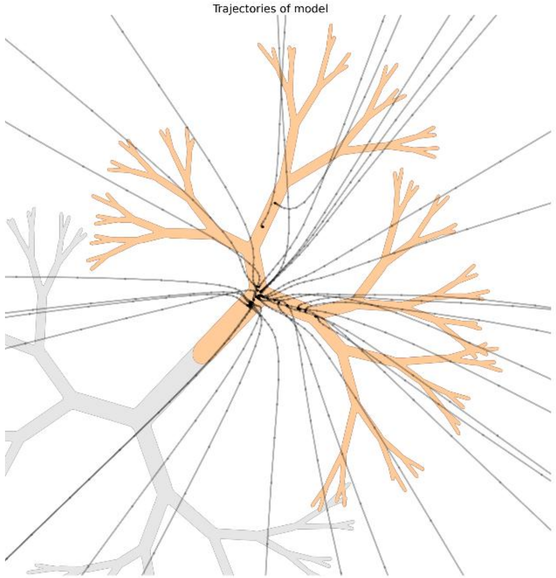
Trajectory of sample when applying sigma scaling



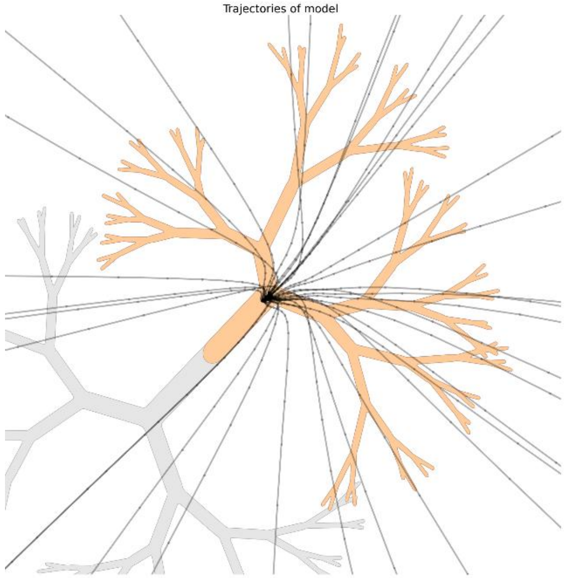
<No guidance>



<k=1.04>



<k=1.10>



<k=1.20>

Experiment

Weight scheduler and sigma scaling

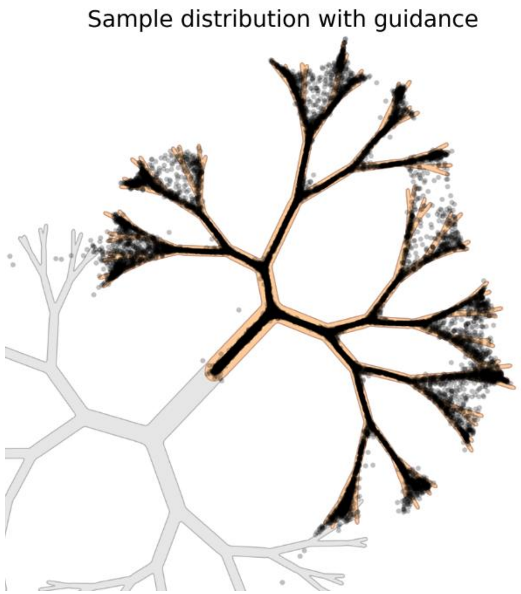
- ✓ For 32 timesteps in toy example, increase guidance weight following using cosine curve

$$w_t = \frac{1}{2} guidance_{max} (1 + cos \frac{t}{T} \pi)$$

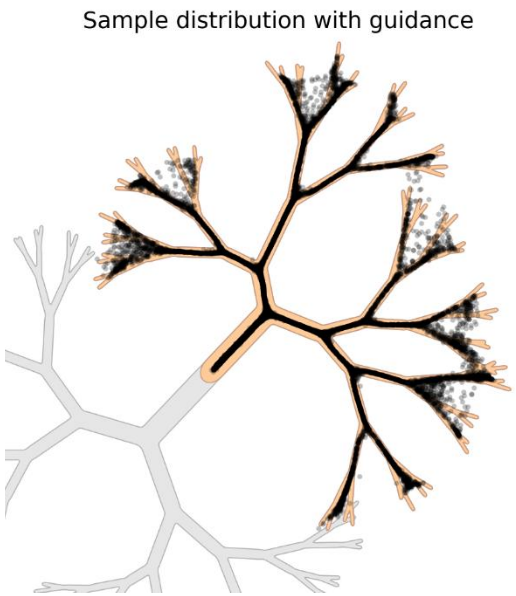
- ✓ Since it is not able to apply weight scheduler solely on model which uses identical main network and guidance network, it is necessary to apply both weight scheduler and sigma scaling method at the same time.

Experiment

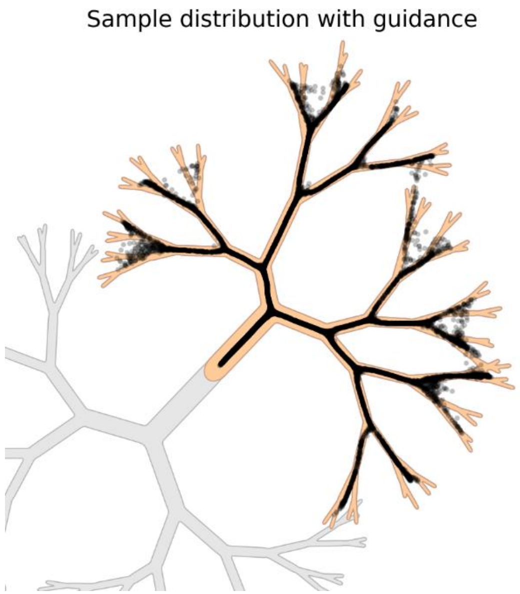
<No guidance>



<k=1.20, w=4>



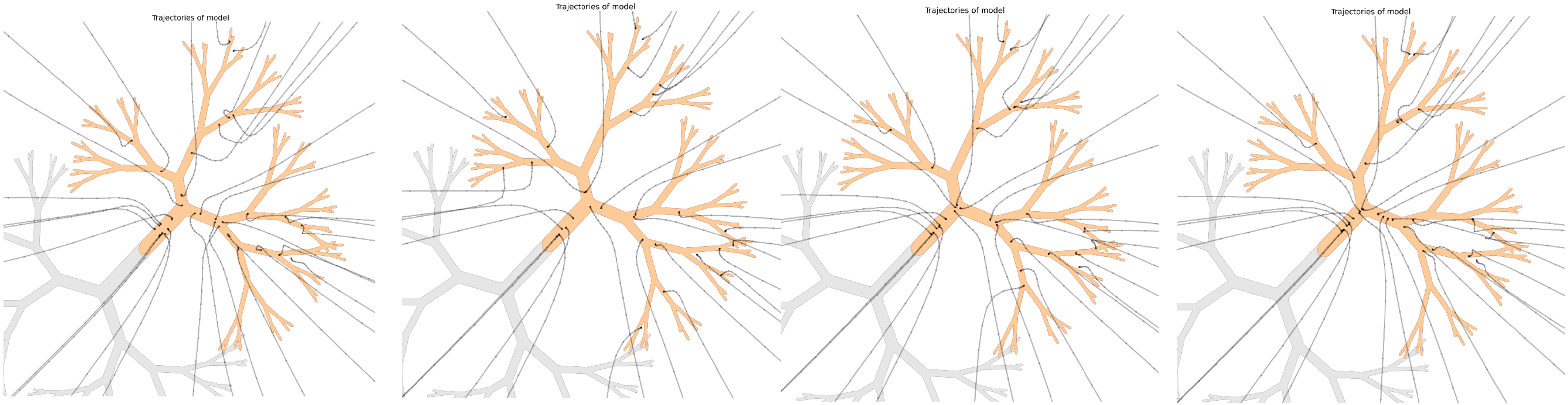
<k=1.20, w=6>



<k=1.20, w=8>

Experiment

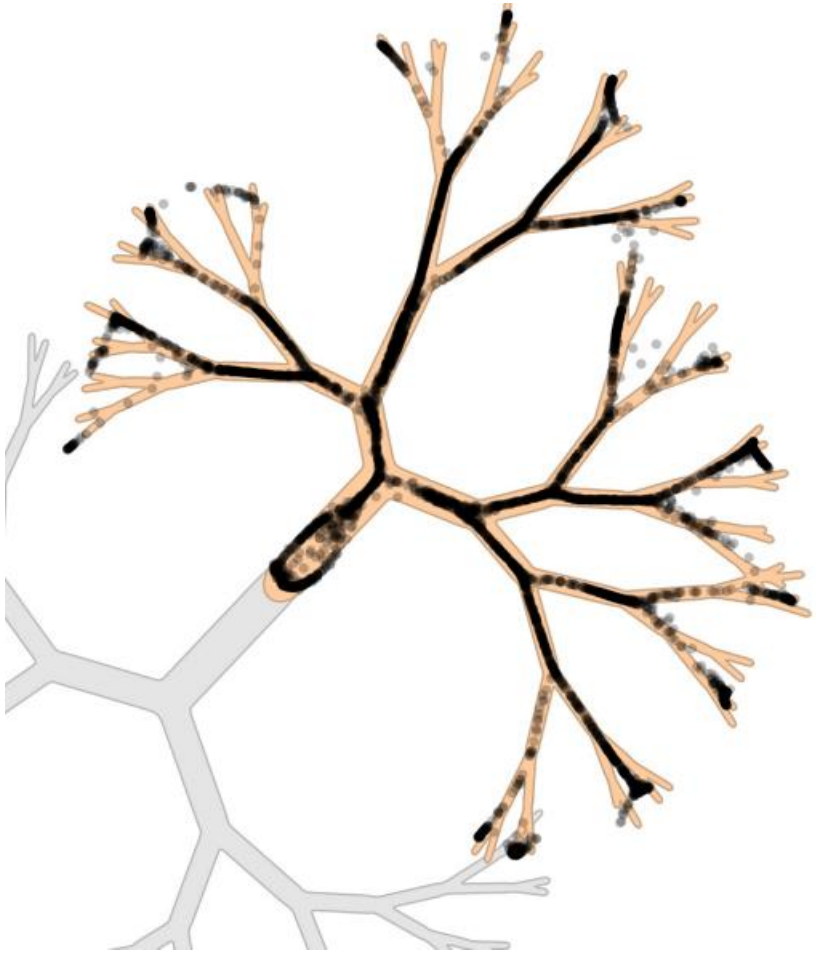
Trajectory of sample when applying sigma scaling and weight scheduler

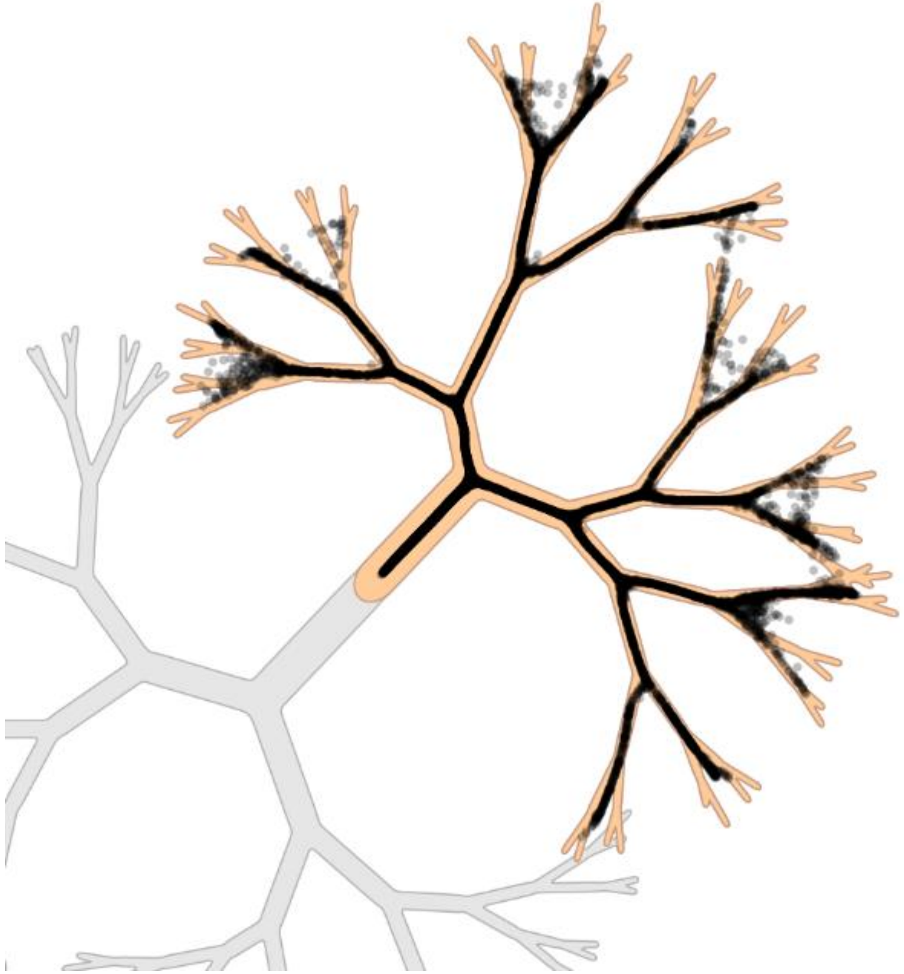


<No guidance> <k=1.20, w=4> <k=1.20, w=6> <k=1.20, w=8>

Experiment

Comparing with autoguidance





Experiment

Interpretation of output

- ✓ Sigma scaling tends to decrease outliers as k increases, but too large k pulls samples to large density region
- ✓ Weight scheduler should be used with sigma scaling with same time and larger weight pulls samples to the large density region
- ✓ Using weight scheduler and sigma scaling can add diversity of sample comparing to using only sigma scaling
- ✓ Comparing to autoguidance, applying weight scheduler and sigma scaling adds diversity but still has problem of outliers
- ✓ It is not sure that those methods will show performance in real image dataset.

Applying on image generation

Prompt: "a photograph of an astronaut riding a horse"

<baseline>

<linear scheduler>

<cosine scheduler>

Applying on image generation

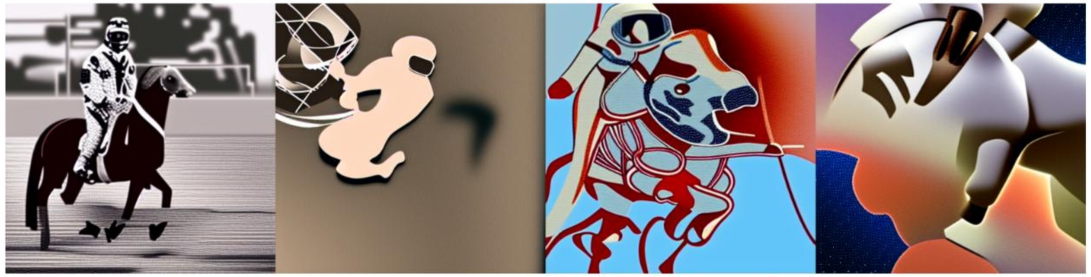
Prompt: "a photograph of an astronaut riding a horse"

Heavy tail guidance with multiply 1.2

Heavy tail guidance with multiply 1.4

Heavy tail guidance with multiply 1.6

Heavy tail guidance with multiply 1.200 + cosine scheduler



Heavy tail guidance with multiply 1.400 + cosine scheduler

Heavy tail guidance with multiply 1.600 + cosine scheduler