ETH zürich

Organisationseinheit verbal
optional auf 2 Zeilen

Guiding a Diffusion Model with a Bad Version of Itself

Jakob Hütteneder

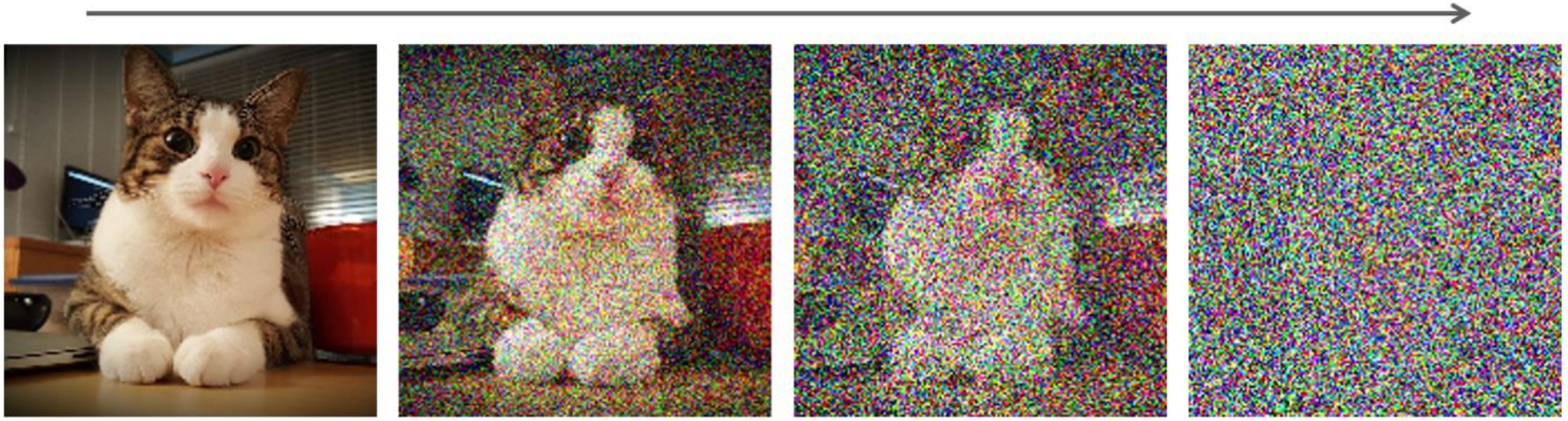
“A photograph of a cat wearing a superman costume”

Image generated using Stable Diffusion.

“A photograph of an astronaut riding a horse”

Image generated using Stable Diffusion.

Diffusion – short introduction



“Mushroom”

“Palace”

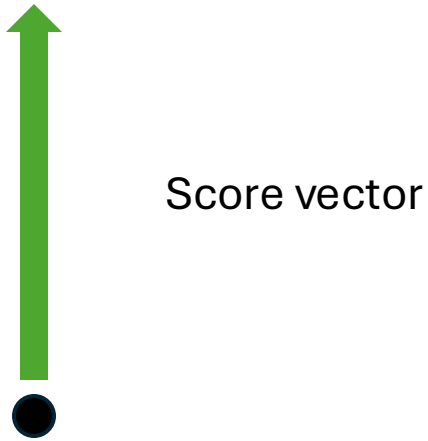
[1]

Score matching

Score vector

Classifier Free Guidance - CFG

- How can prompt alignment be boosted?

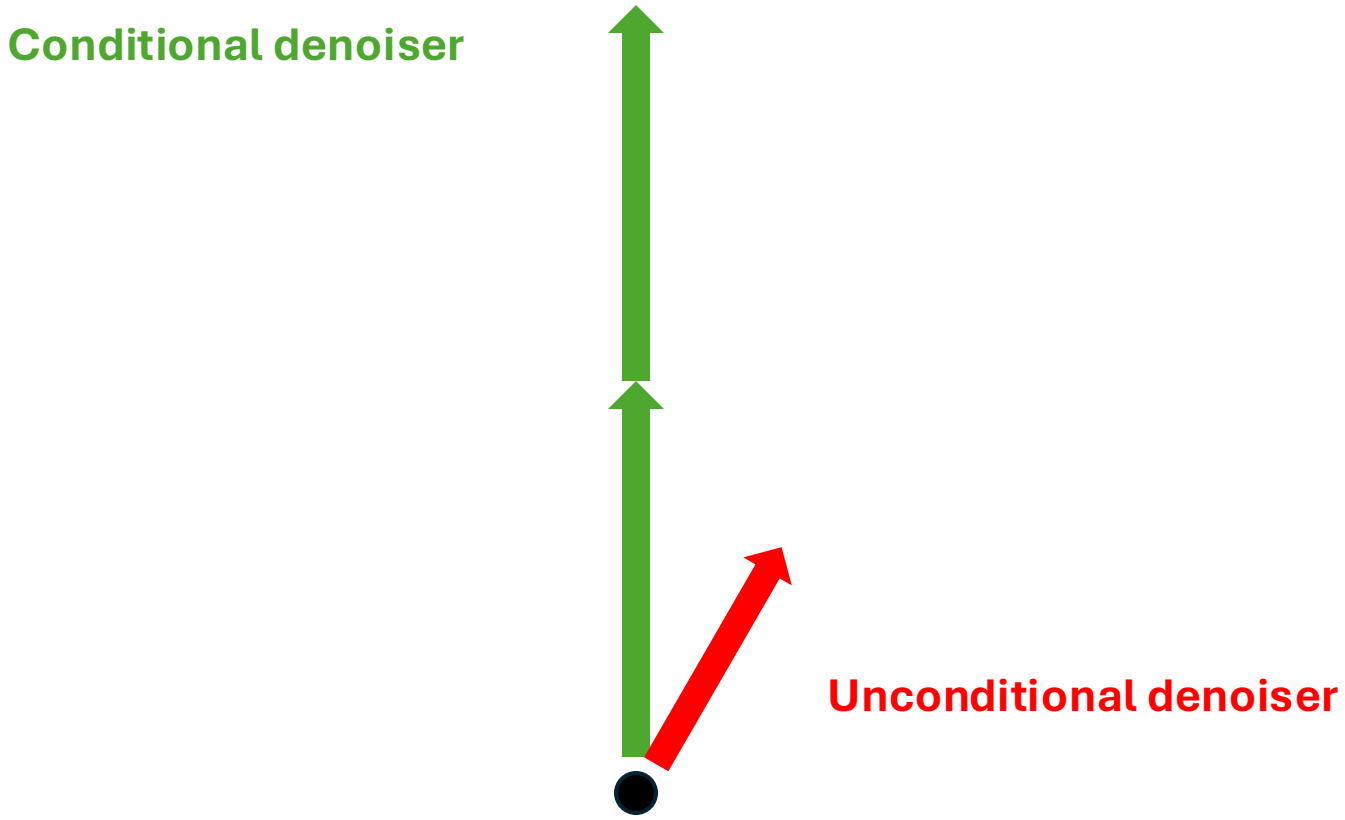


[2]

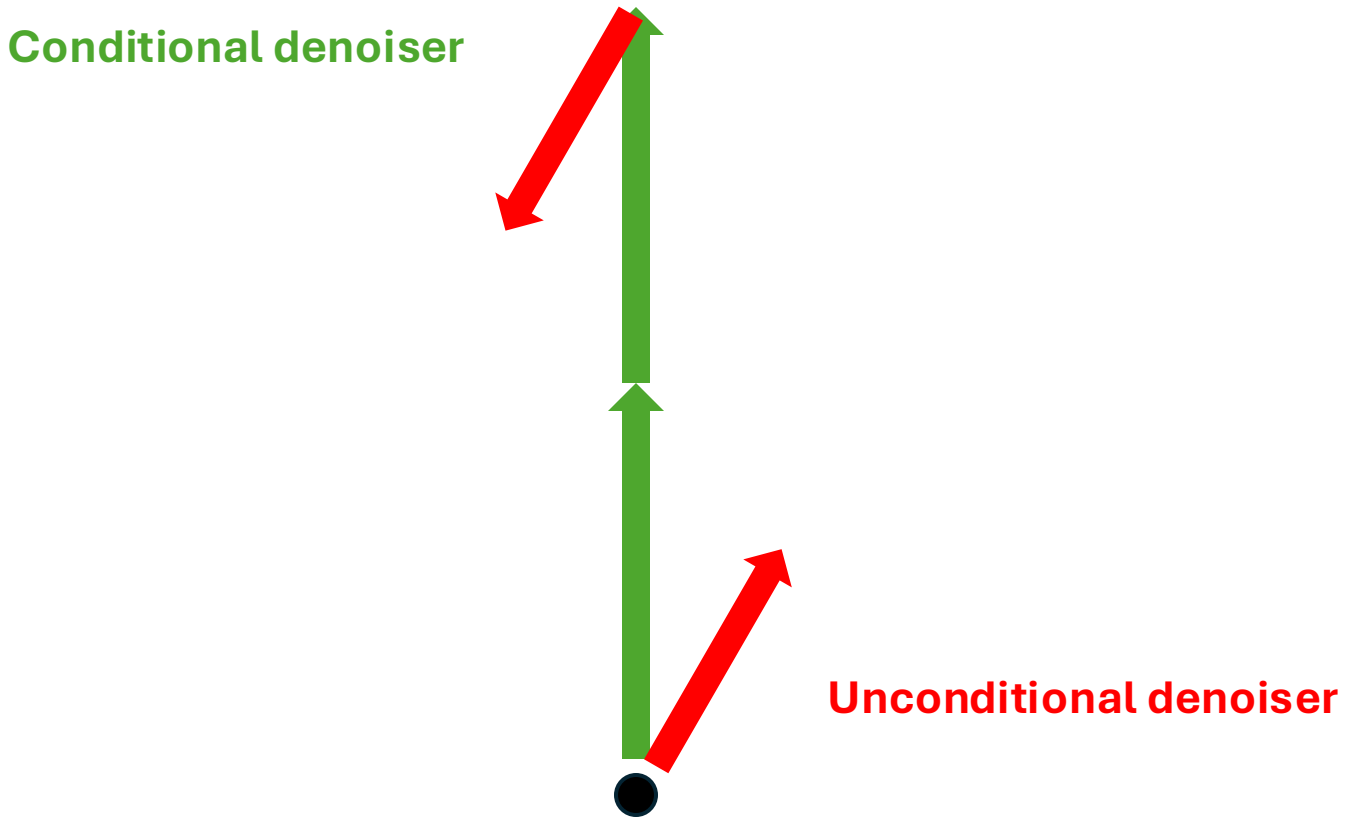
Classifier Free Guidance - CFG

“More prompt”

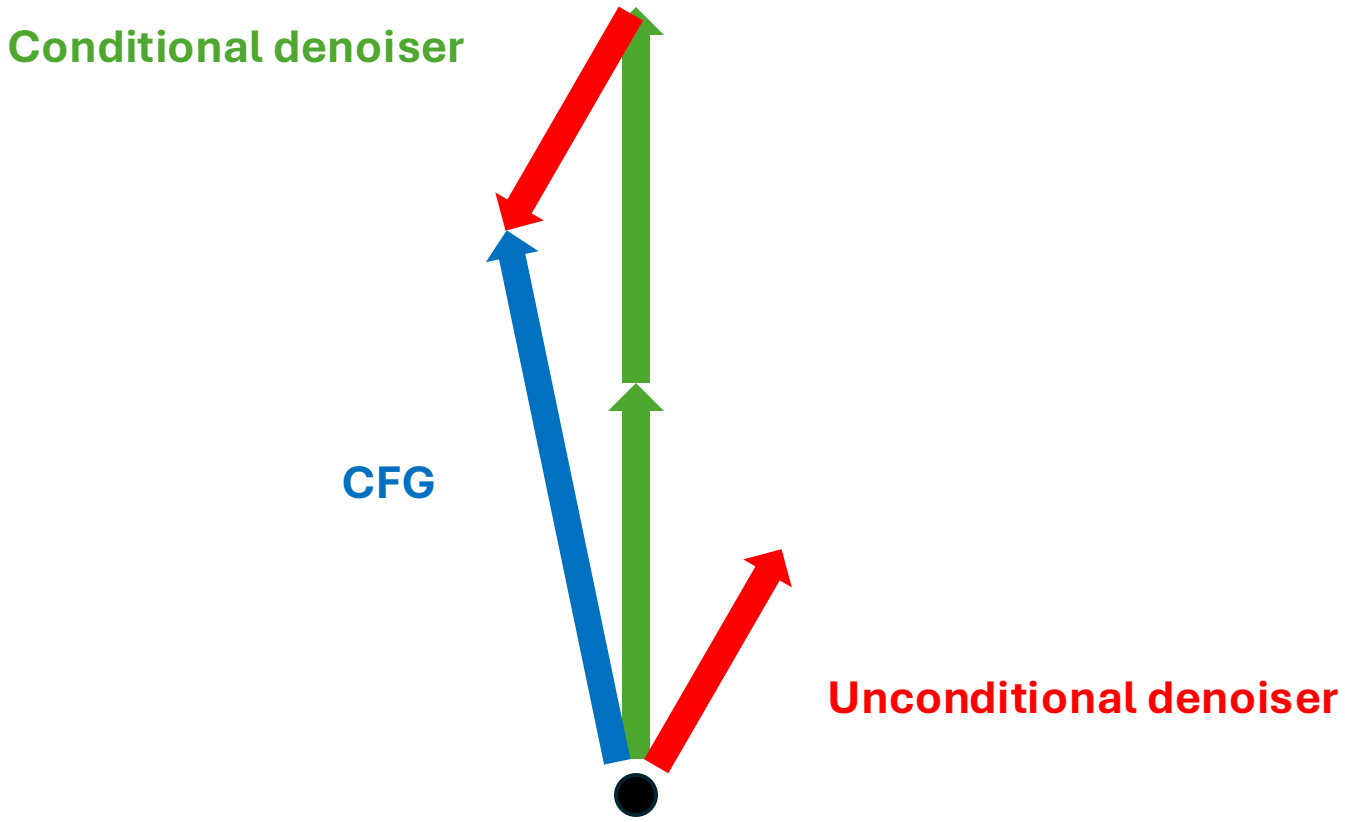
Classifier Free Guidance - CFG



Classifier Free Guidance - CFG



Classifier Free Guidance - CFG



Classifier Free Guidance - CFG

$$D_w(\mathbf{x};\sigma,\mathbf{c}) = wD_1(\mathbf{x};\sigma,\mathbf{c}) + (1-w)D_0(\mathbf{x};\sigma,\mathbf{c})$$

	D_1	D_0
CFG	Conditional Denoiser	Unconditional Denoiser

$w = 1$ $w = 2$ $w = 3$

“Mushroom”

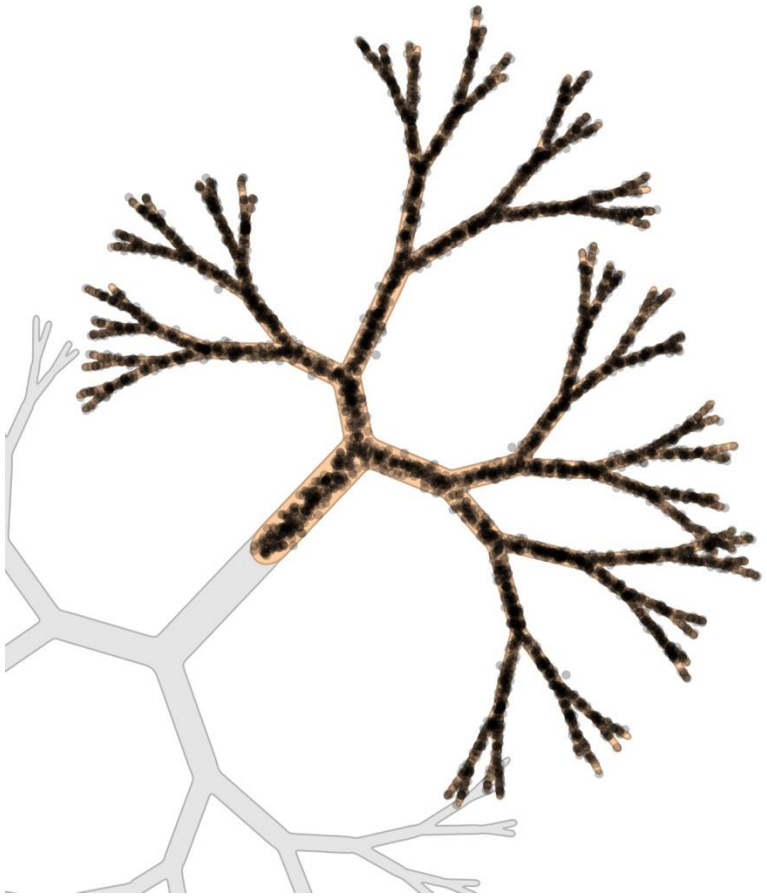
“Palace”

[1]

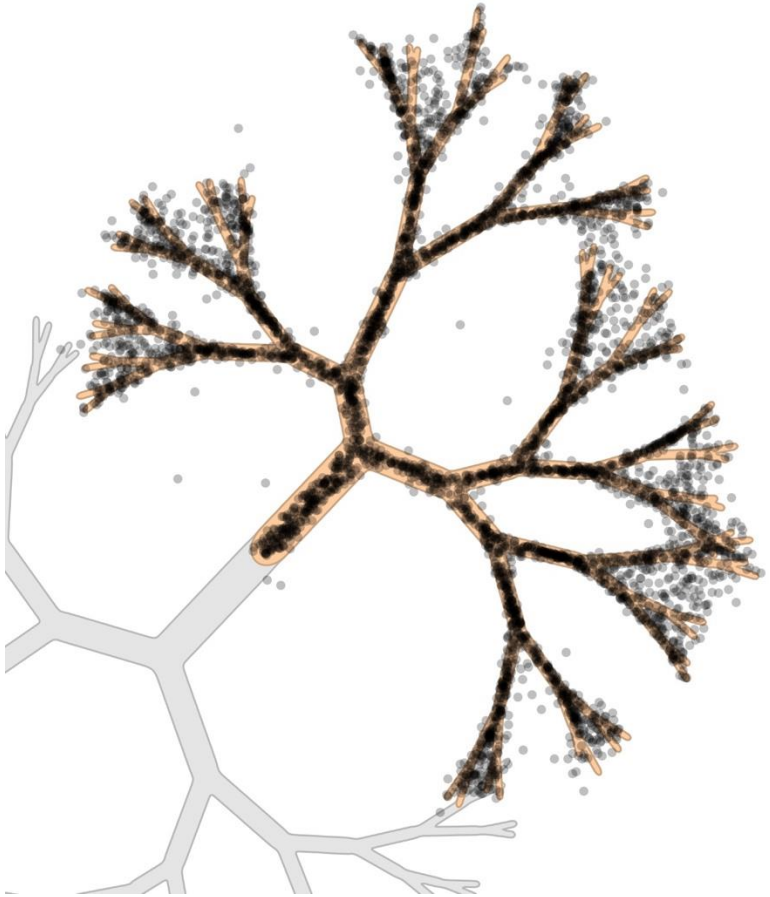
Score Matching leads to image artifacts

- Behaves similarly to maximum likelihood estimation
- Extreme penalties for underestimating likelihood of any sample
- Can restrain model‘s ability to focus on common patterns

Score Matching leads to outliers



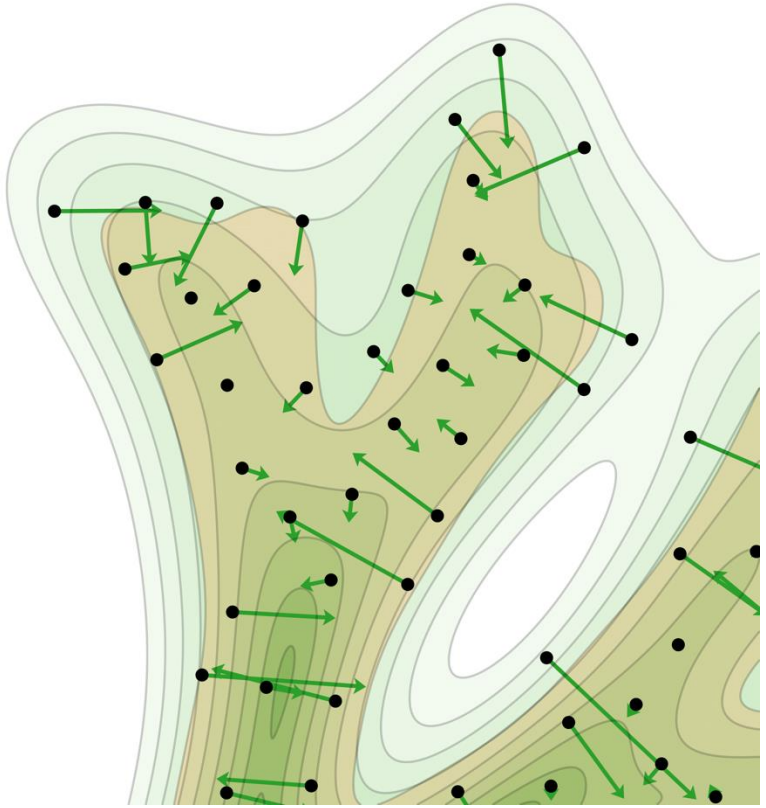
Ground Truth



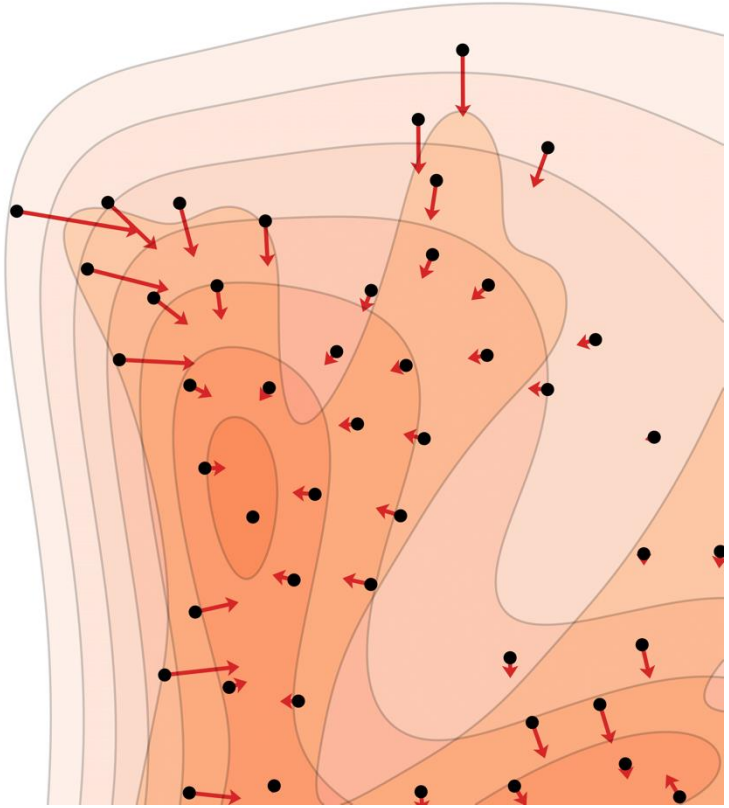
No Guidance

[1]

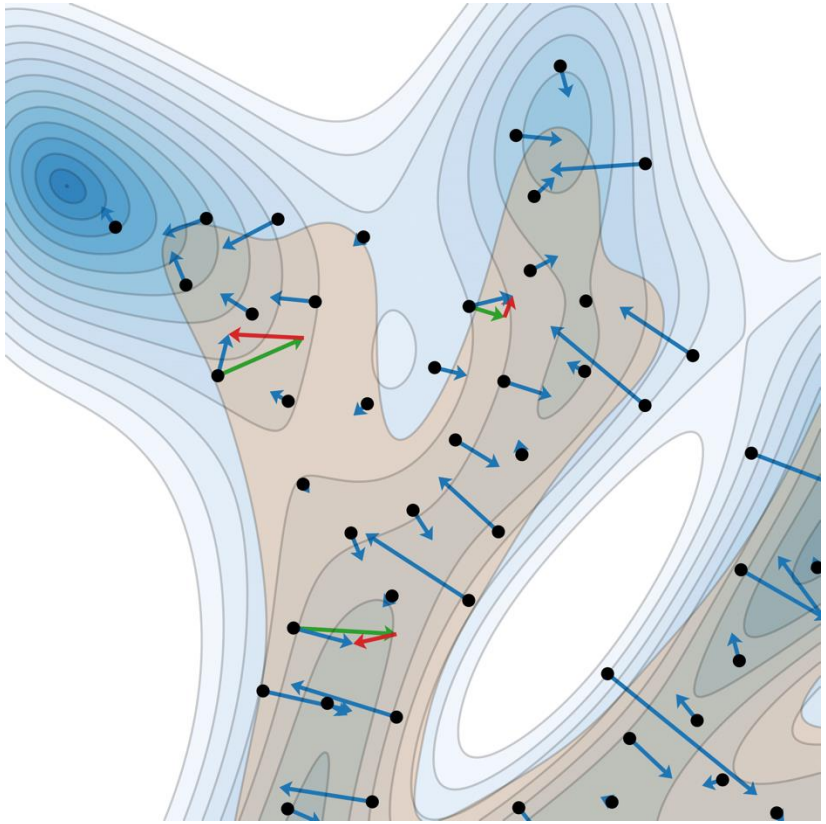
How does CFG work?



Conditional



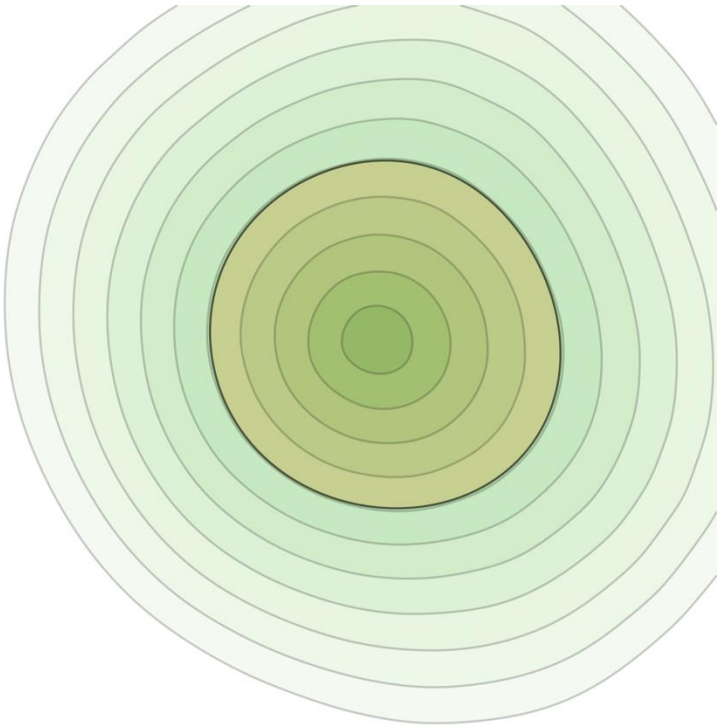
Unconditional

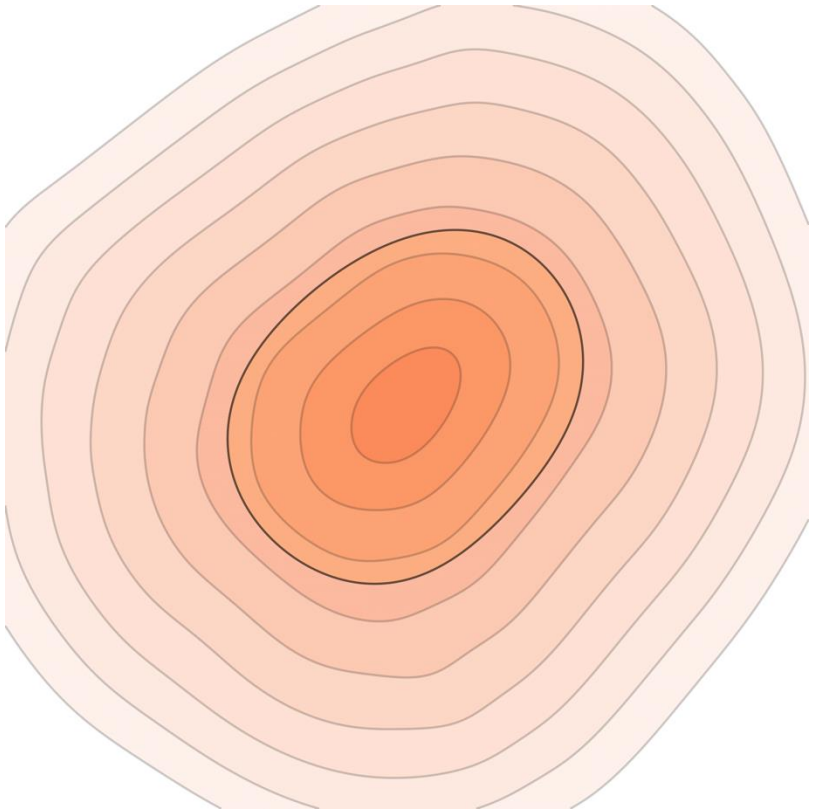


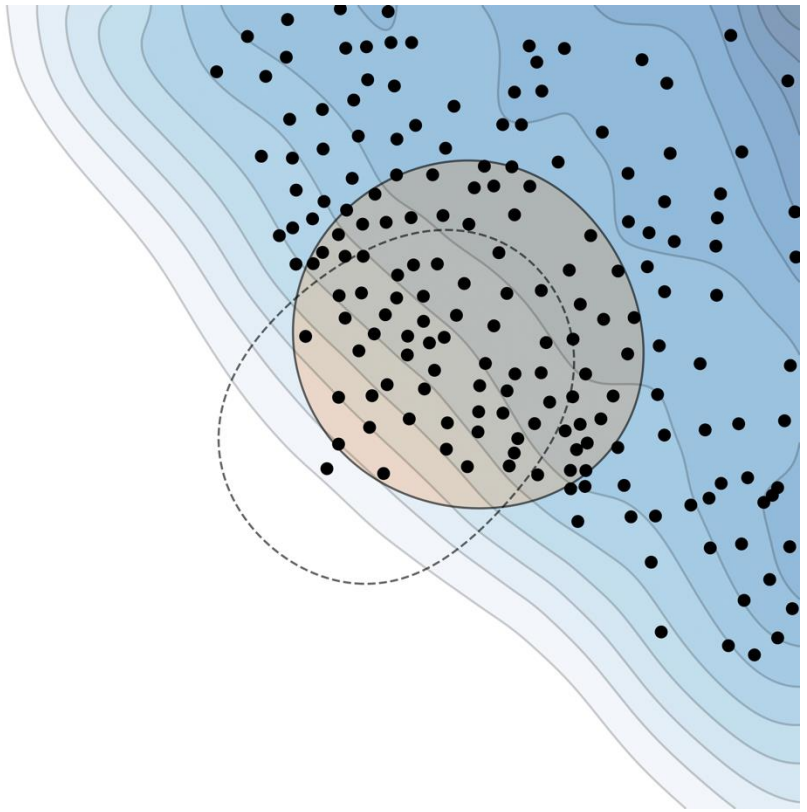
Conditional - Unconditional

[1]

CFG







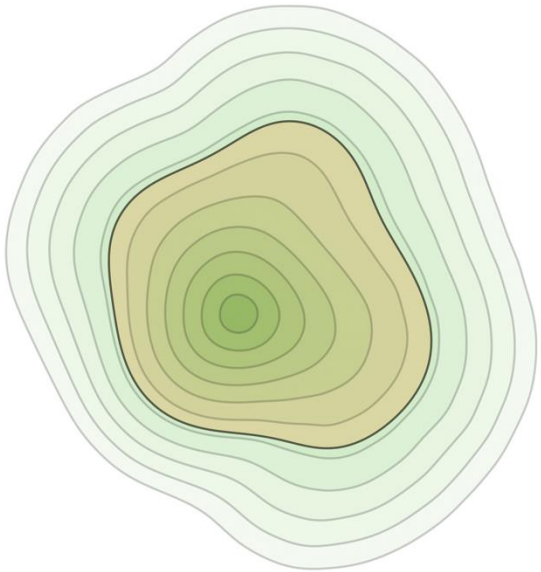
Conditional

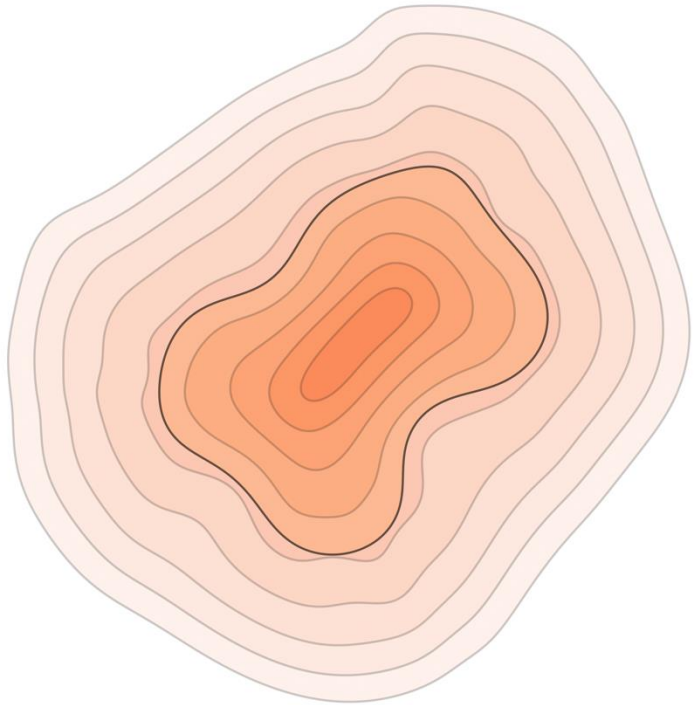
Unconditional

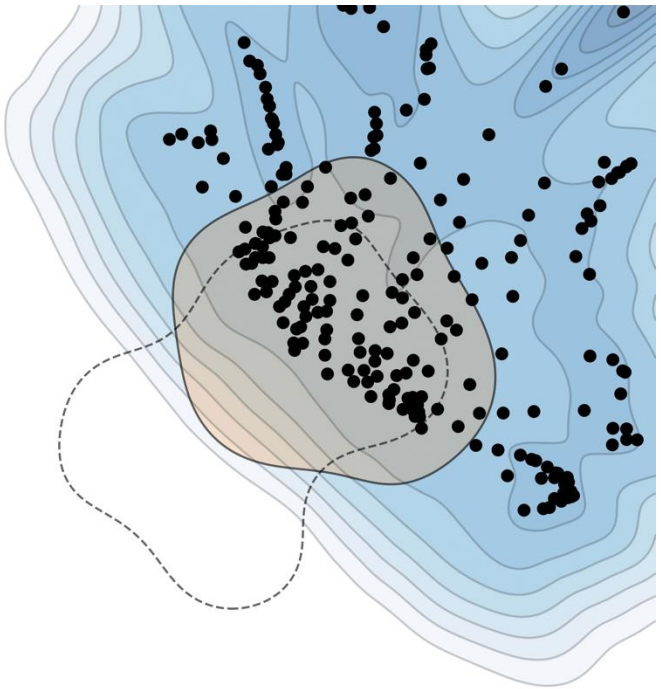
Conditional - Unconditional

[1]

CFG







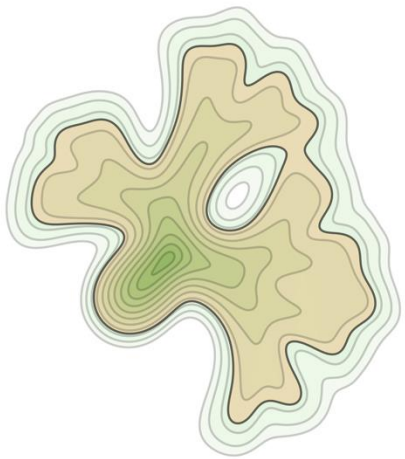
Conditional

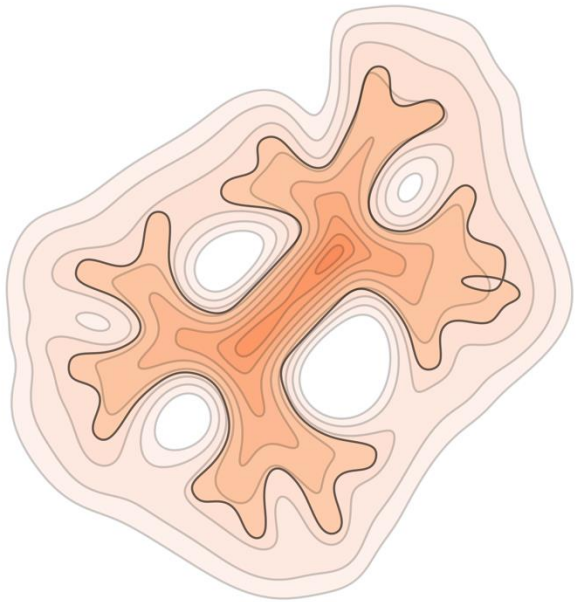
Unconditional

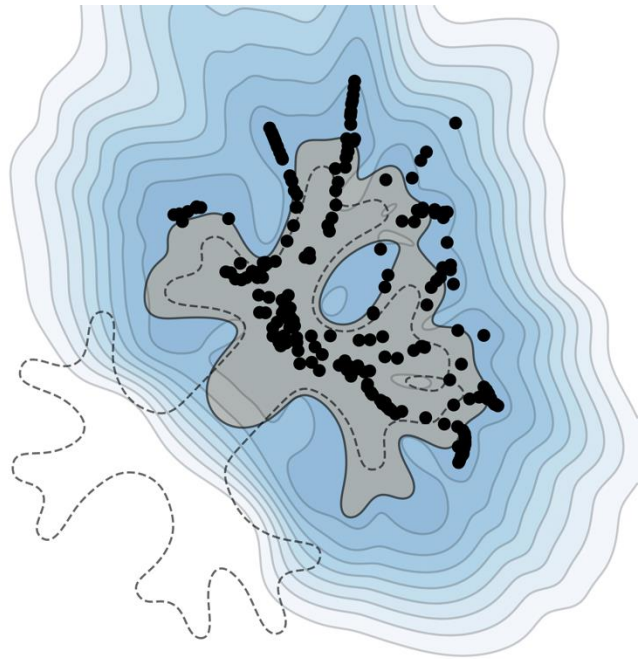
Conditional - Unconditional

[1]

CFG







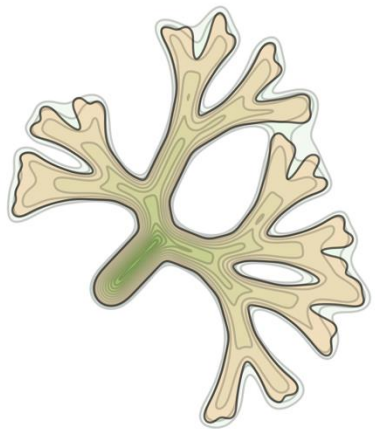
Conditional

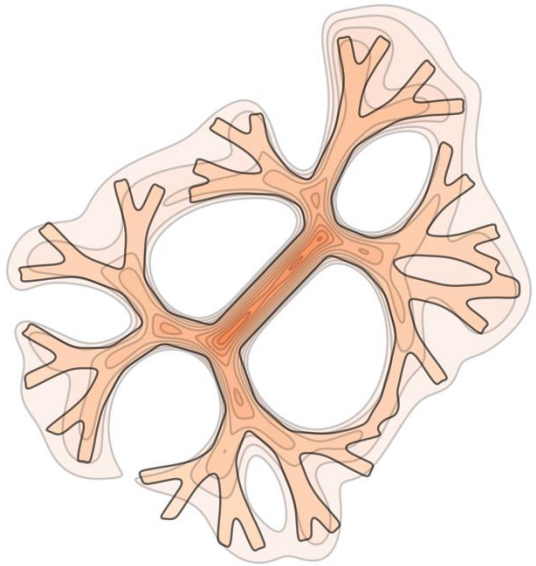
Unconditional

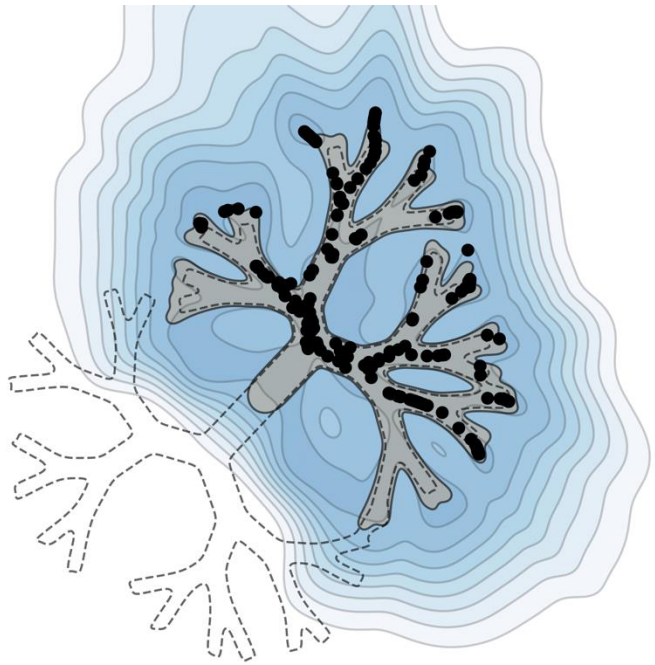
Conditional - Unconditional

[1]

CFG







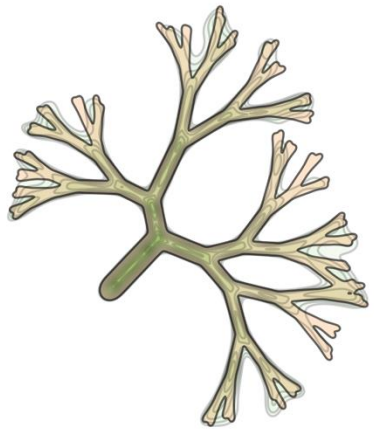
Conditional

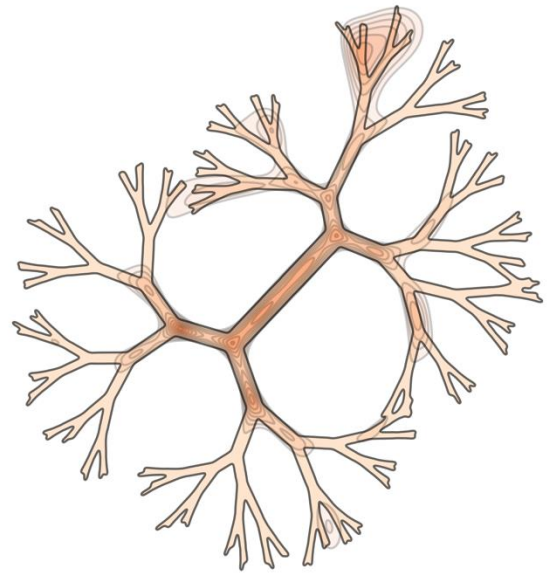
Unconditional

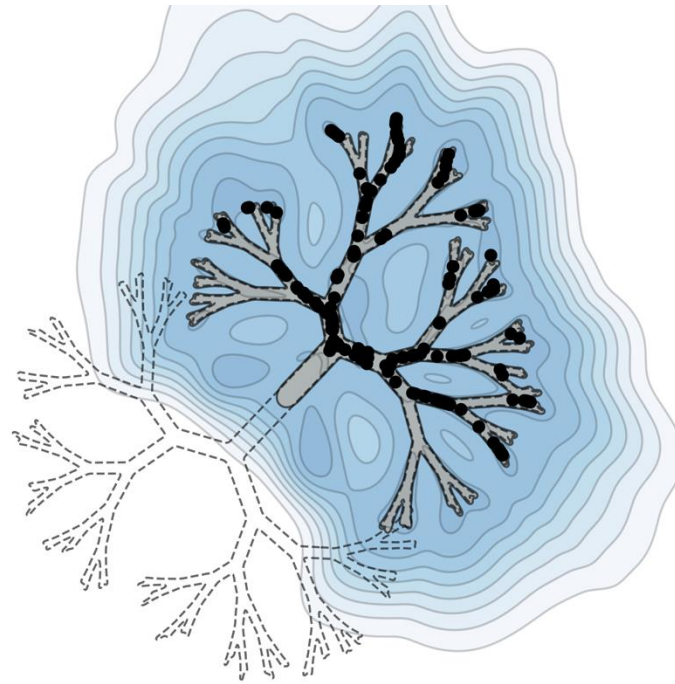
Conditional - Unconditional

[1]

CFG







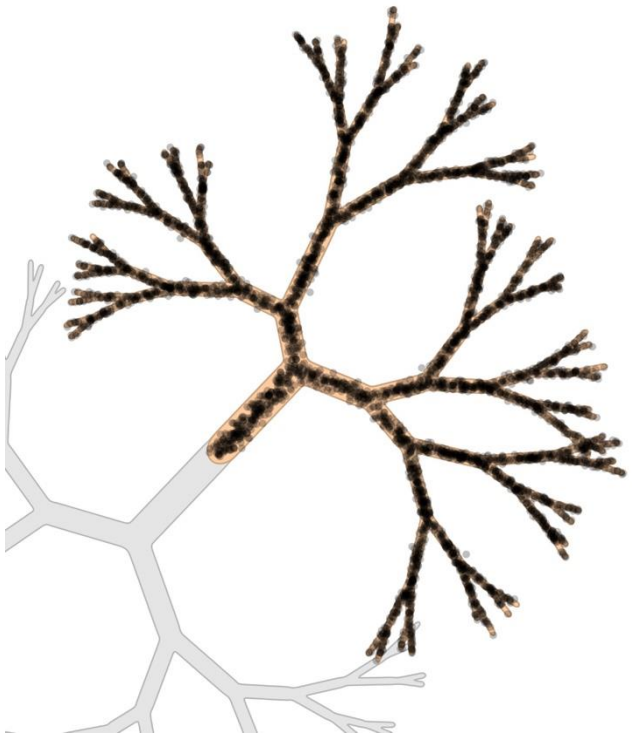
Conditional

Unconditional

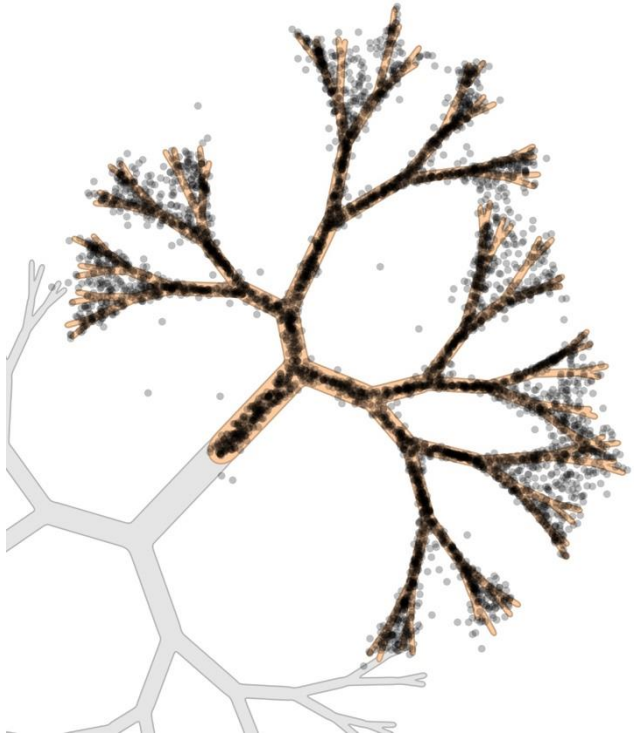
Conditional - Unconditional

[1]

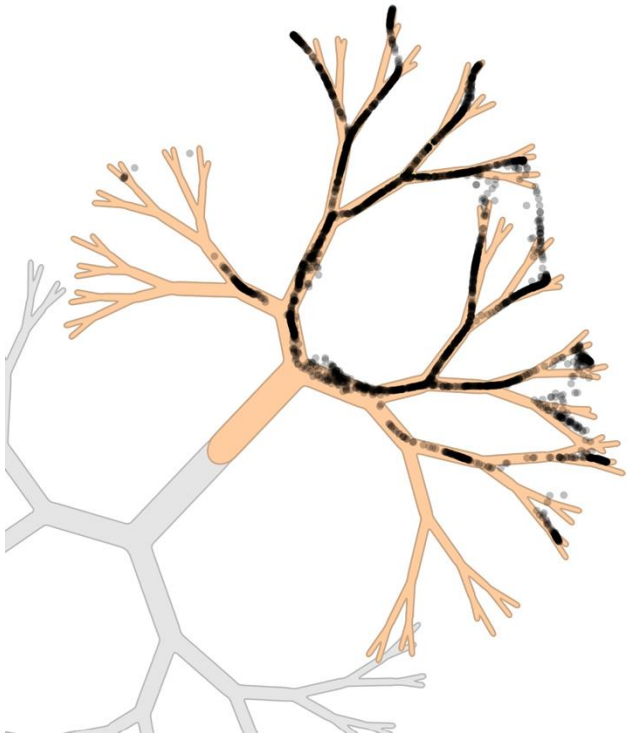
How is it in comparison?



Ground Truth



No Guidance



CFG

[1]

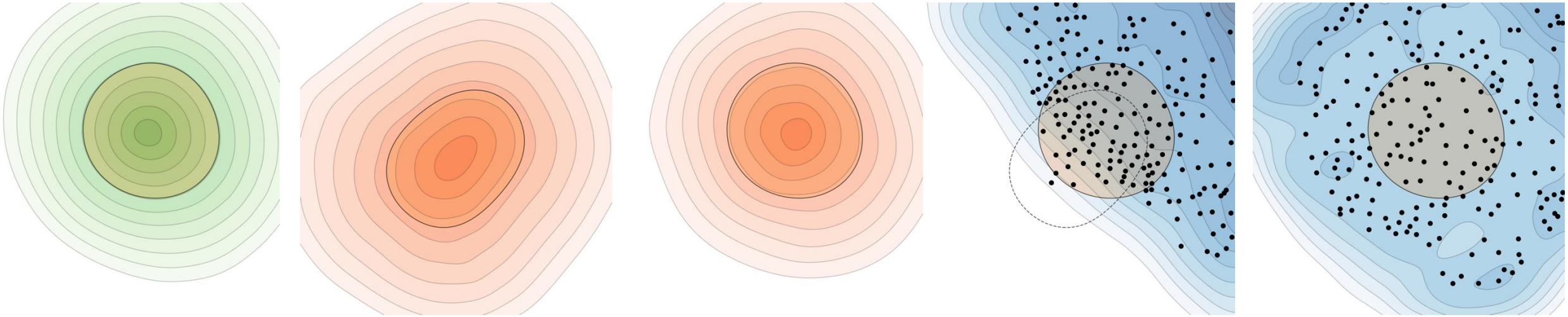
New method - Autoguidance

$$D_w(\mathbf{x};\sigma,\mathbf{c}) = wD_1(\mathbf{x};\sigma,\mathbf{c}) + (1-w)D_0(\mathbf{x};\sigma,\mathbf{c})$$

	D_1	D_0
CFG	Conditional Denoiser	Unconditional Denoiser
Autoguidance	Conditional Denoiser	Worse Conditional Denoiser

CFG

Autoguidance



Conditional

Unconditional

Worse Conditional

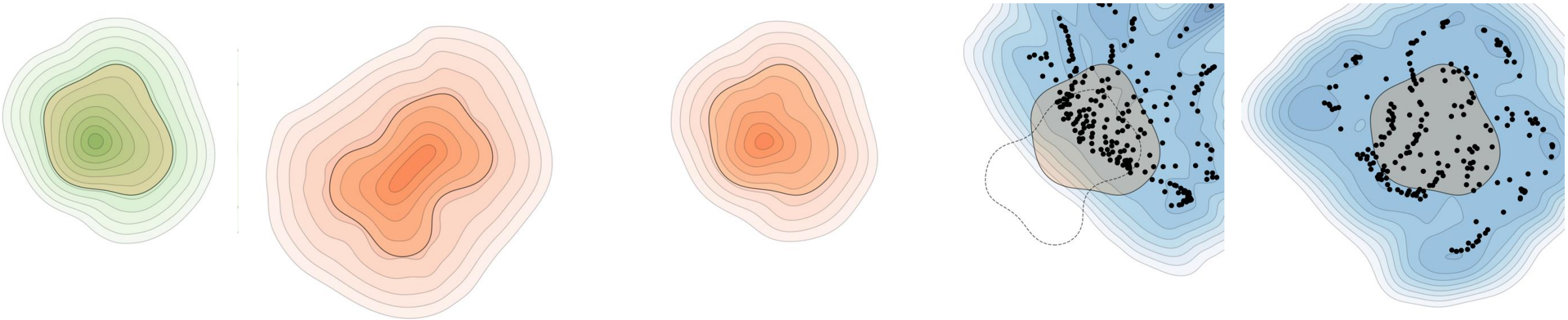
Conditional - Unconditional

Conditional – Worse Conditional

[1]

CFG

Autoguidance



Conditional

Unconditional

Worse Conditional

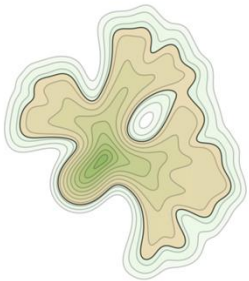
Conditional - Unconditional

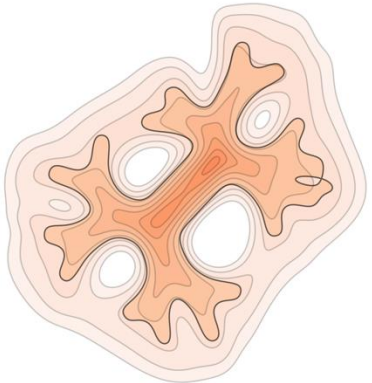
Conditional – Worse Conditional

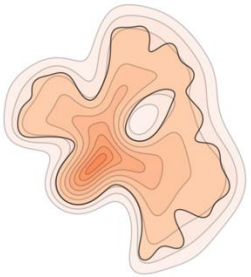
[1]

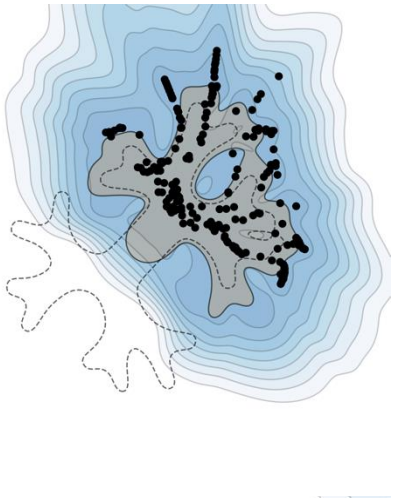
CFG

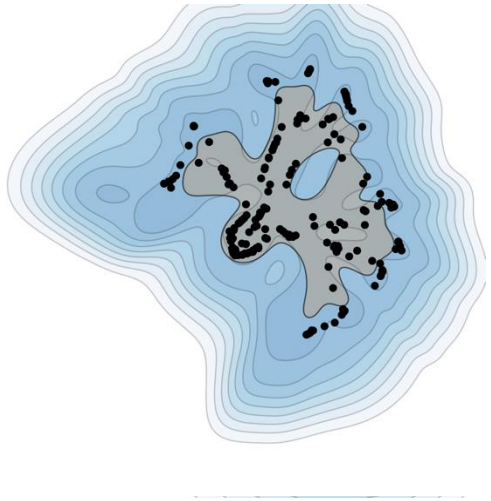
Autoguidance











Conditional

Unconditional

Worse Conditional

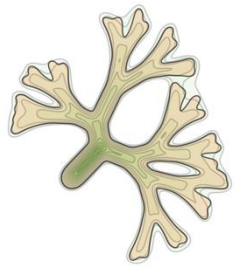
Conditional - Unconditional

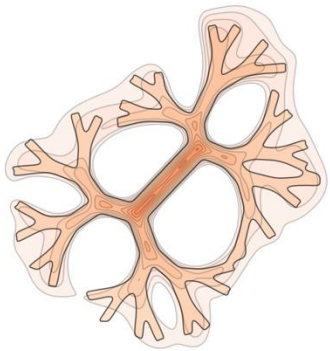
Conditional – Worse Conditional

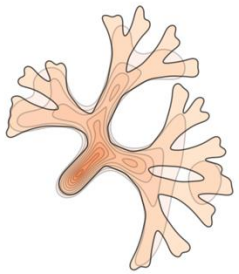
[1]

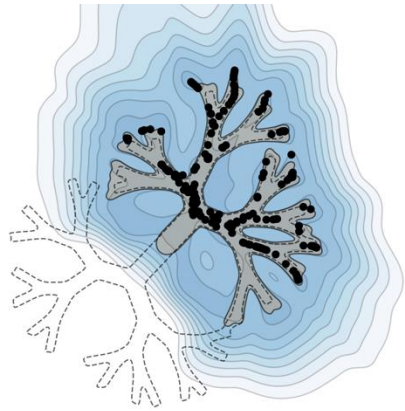
CFG

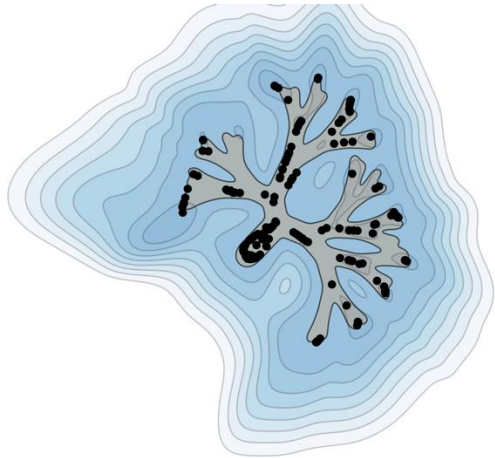
Autoguidance











Conditional

Unconditional

Worse Conditional

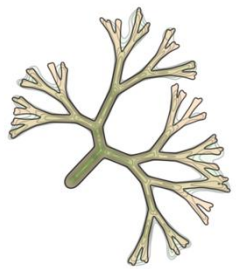
Conditional - Unconditional

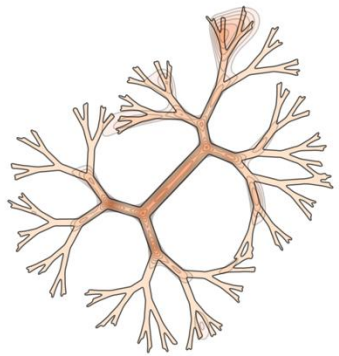
Conditional – Worse Conditional

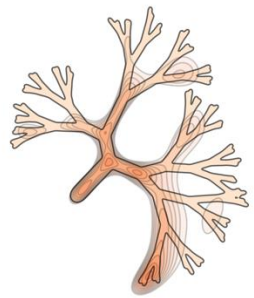
[1]

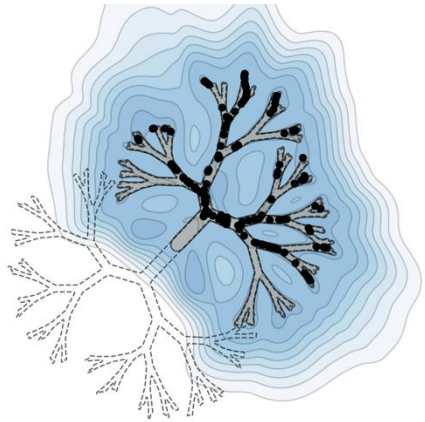
CFG

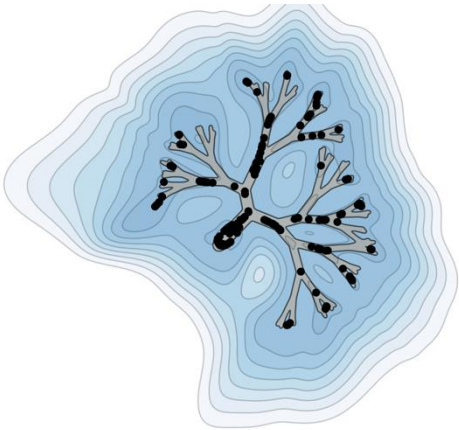
Autoguidance











Conditional

Unconditional

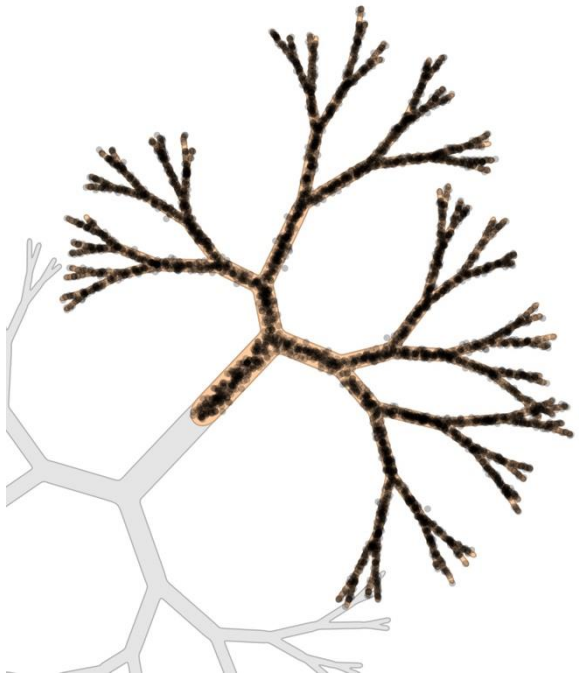
Worse Conditional

Conditional - Unconditional

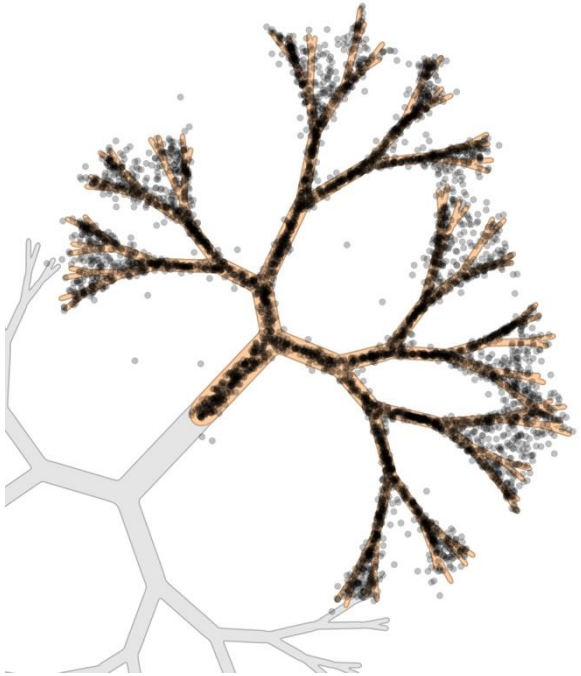
Conditional – Worse Conditional

[1]

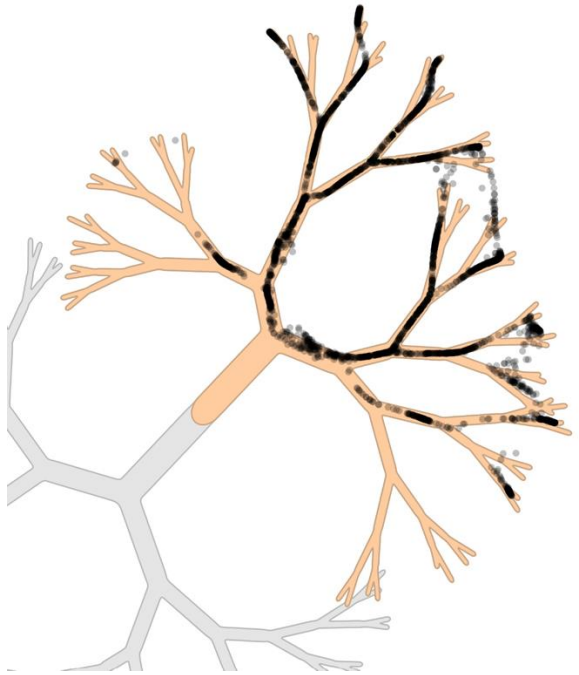
How is it in comparison?



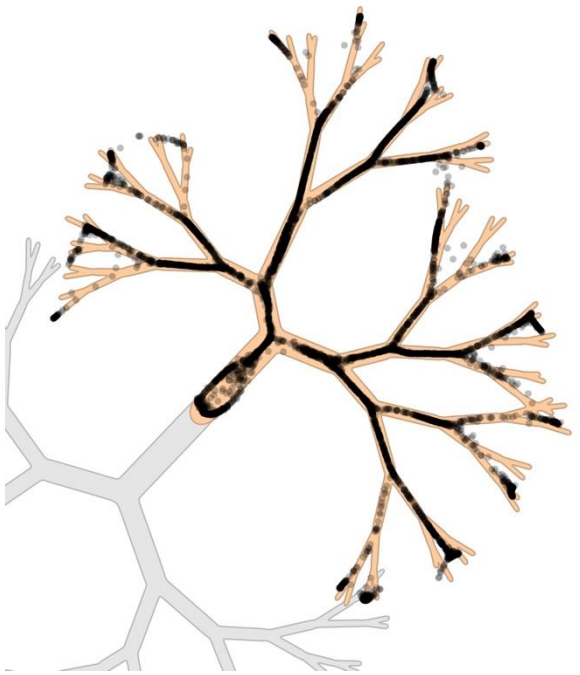
Ground Truth



No Guidance

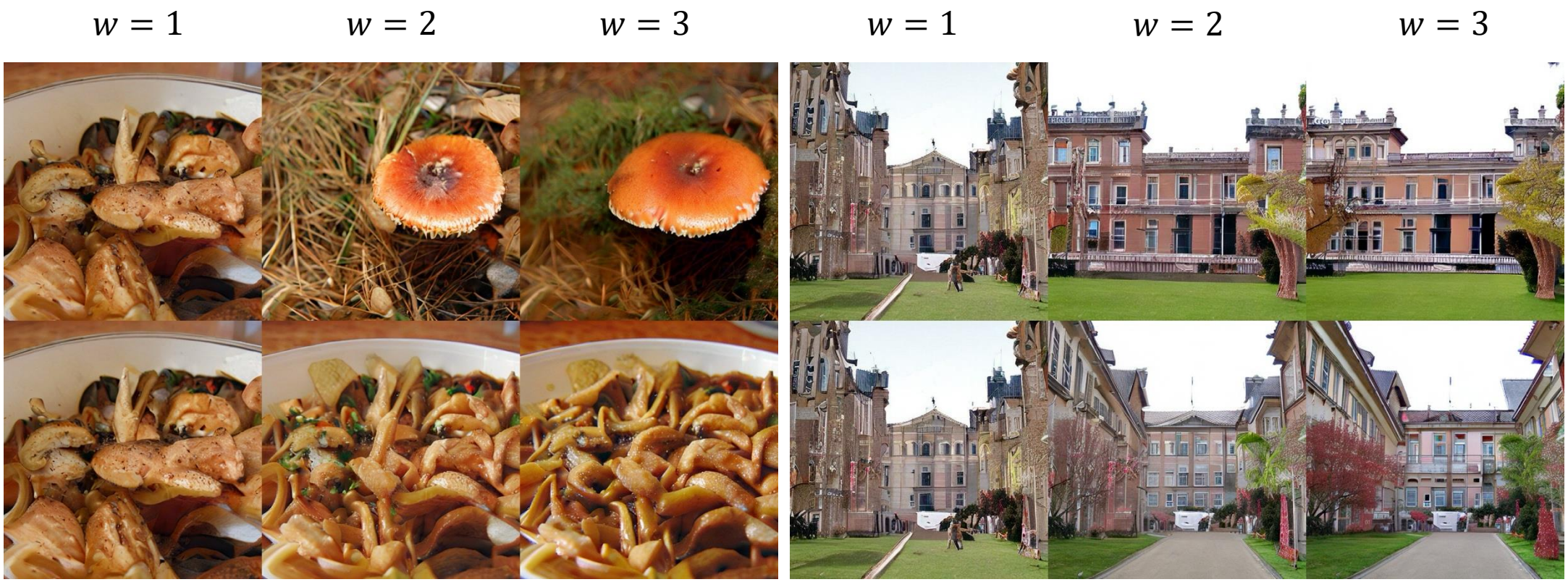


CFG



Autoguidance

[1]



[1]

Fréchet inception distance (FID)

- Metric to assess the quality of images created by a generative model

- Measures the distance between distribution of generated and real images

Autoguidance in practice

- Assumption:

  D_1 and D_0 must suffer from the same kind of degradation.

Experiment: Undoing the damage from synthetic degradations

- **Base model:** EDM2-S trained on ImageNet-512
- **Dropout:** Added in a post-hoc fashion
- **Input noise:** Increased noise level of input images

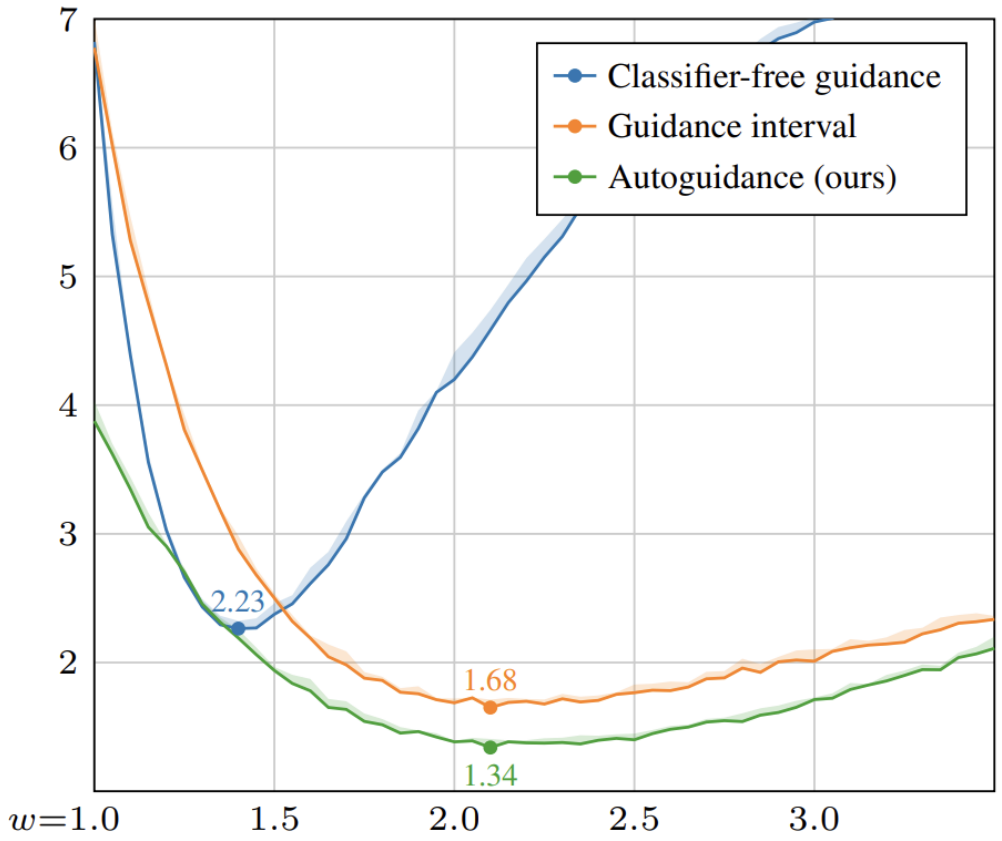
Base Model FID: 2.56

D_1	D_0	FID D_1	FID D_0	FID guided
Dropout 5%	Dropout 10%	4.98	15.00	2.55
Input noise 10%	Input noise 20%	3.96	9.73	2.56
Dropout 5%	Input noise 20%	4.98	9.73	20.00

Which guiding models help?

- Same task and data, significant quality gap
- These worked:
  - Fewer layers and/or features
  - Less training
- These didn‘t:
  - Manual degradations (dropout, input noise, …)
  - Weight quantization
  - Smaller dataset
  - Fundametally different generations of models, e.g. SD3 + SD2

ImageNet-512 FID



[1]

ImageNet-512 performance

Method	FID	w
EDM2-S	2.56	-
+ Classifier-free guidance	2.23	1.40
+ Guidance interval	1.68	2.10
+ Autoguidance (XS, T/16)	**1.34**	2.10
- Reduce training only	1.51	2.20
- Reduce capacity only	2.13	1.80
EDM2-XXL	1.91	-
+ Classifier-free guidance	1.81	1.20
+ Guidance interval	1.40	2.00
+ Autoguidance (M, T/3.5)	**1.25**	2.05
EDM2-S, unconditional	11.67	-
+ Autoguidance (XS, T/16)	**3.86**	2.85

Additional training cost

D_1	D_0	Mparams D_1	Mparams D_0	n iterations	Additional cost %
EDM2-XXL	EDM2-M	1523	498	1/3.5	11
EDM2-S	EDM2-XS	280	125	1/16	3.6

[3}

“A blue jay standing on a large basket of rainbow macarons”

Guidance strength

CFG

Interpolation between CFG and Autoguidance

Autoguidance

Conclusion & Discussion

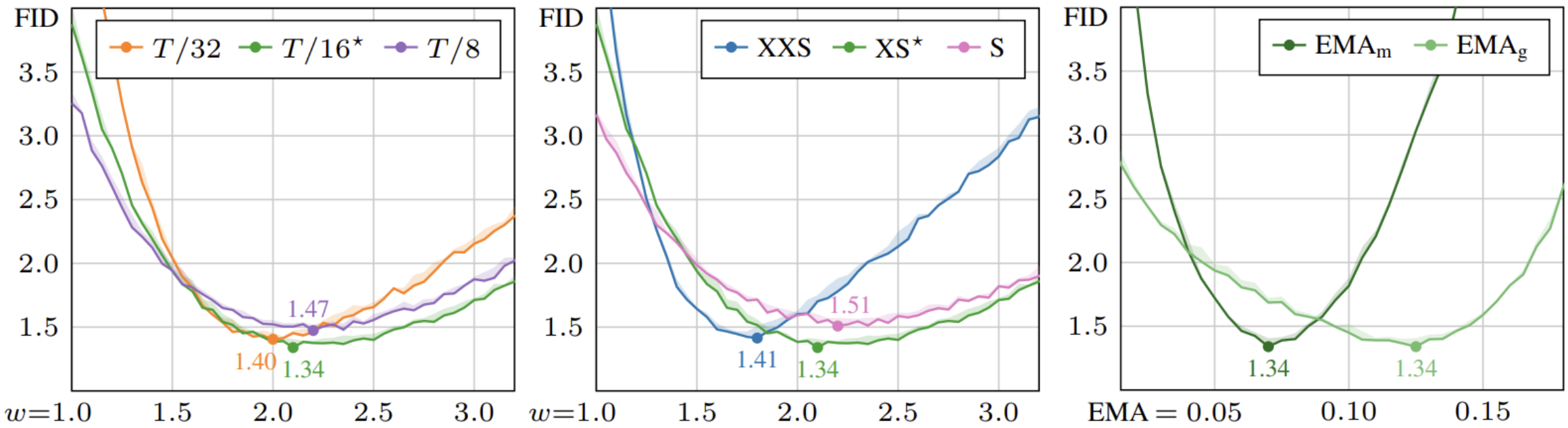
- Control over quality/variation tradoff
- Broader applicability than CFG
- Good rules of thumb for guiding models
- Explore noise level-dependent guidance weight

References

[1] Karras, Tero, et al. "Guiding a diffusion model with a bad version of itself." *Advances in Neural Information Processing Systems* 37 (2025): 52996-53021.

[2] Ho, Jonathan, and Tim Salimans. "Classifier-free diffusion guidance." arXiv preprint arXiv:2207.12598 (2022).

[3] Karras, Tero, et al. "Analyzing and improving the training dynamics of diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.


FID
3.5
3.0
2.5
2.0
1.5
$T/32$
$T/16^\star$
$T/8$
1.47
1.40
1.34
w=1.0
1.5
2.0
2.5
3.0
FID
3.5
3.0
2.5
2.0
1.5
XXS
XS*
S
1.51
1.41
1.34
w=1.0
1.5
2.0
2.5
3.0
FID
3.5
3.0
2.5
2.0
1.5
EMA$_\text{m}$
EMA$_\text{g}$
1.34
1.34
EMA = 0.05
0.10
0.15

CFG

Autoguidance