On Distillation of Guided Diffusion Models

Chenlin Meng*
Stanford University
chenlin@cs.stanford.edu

Robin Rombach
Stability AI & LMU Munich
robin@stability.ai

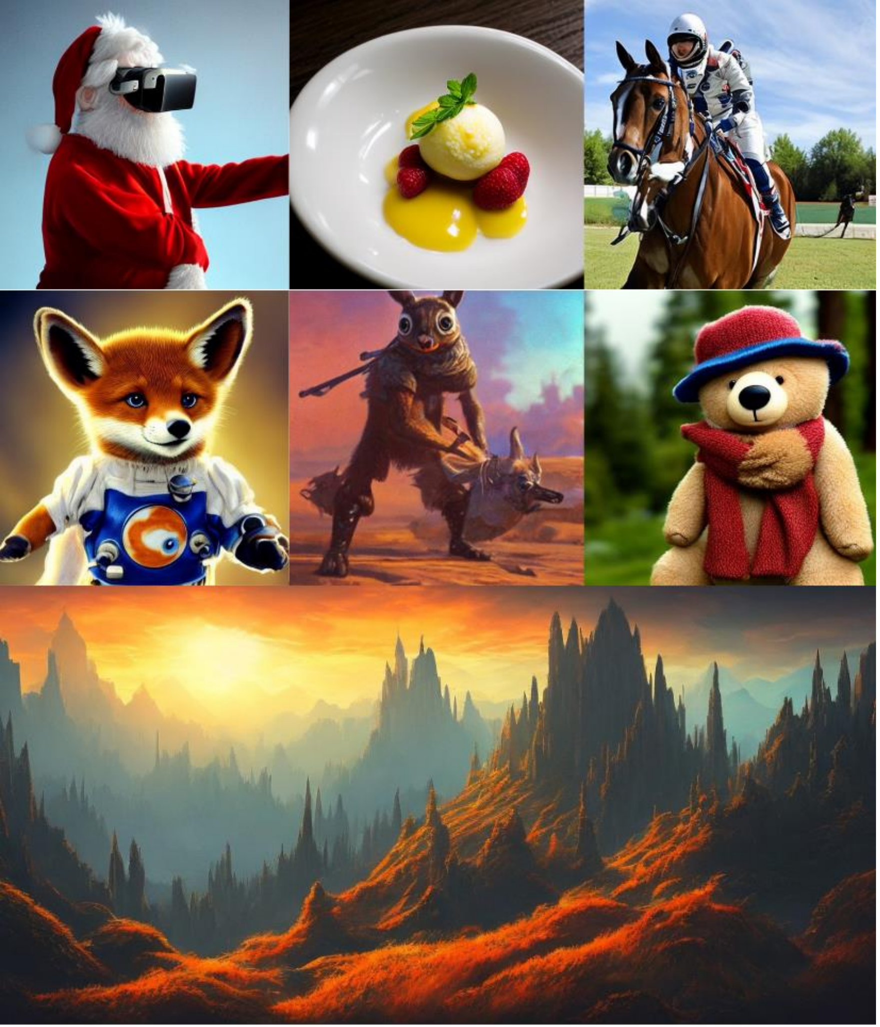
Ruiqi Gao
Google Research, Brain Team
ruiqig@google.com

Diederik P. Kingma
Google Research, Brain Team
durk@google.com

Stefano Ermon
Stanford University
ermon@cs.stanford.edu

Jonathan Ho
Google Research, Brain Team
jonathanho@google.com

Tim Salimans
Google Research, Brain Team
salimans@google.com



Text-guided generation (4 steps)

Input Mask Result 1 Result 2

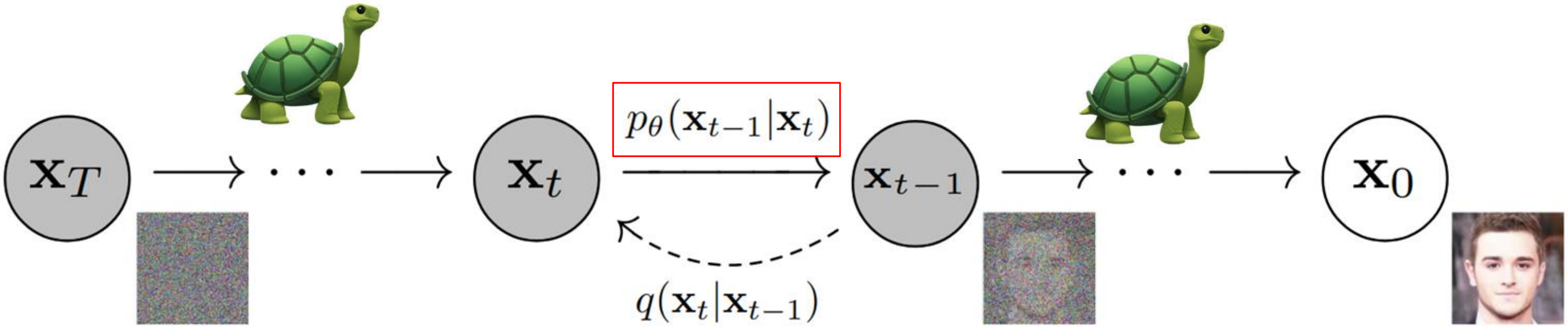
Image inpainting (2 steps)

Input Output (different styles)

Image to image translation (3 steps)

Review: DDPM

- T = 1000 !!
    - So slow…



Review: Classifier-Free Guidance

- Motivation: $P(c|x) \propto \boxed{P(x|c)} / \boxed{P(x)}$

- Use a combination of
  1) Conditional model
  2) Unconditional model

$$\hat{\mathbf{x}}_{\boldsymbol{\theta}}^{w} = (1+w)\boxed{\hat{\mathbf{x}}_{c,\boldsymbol{\theta}}} - w\boxed{\hat{\mathbf{x}}_{\boldsymbol{\theta}}}$$

*Note that coefficients are chosen to preserve scaling

Review: Classifier-Free Guidance

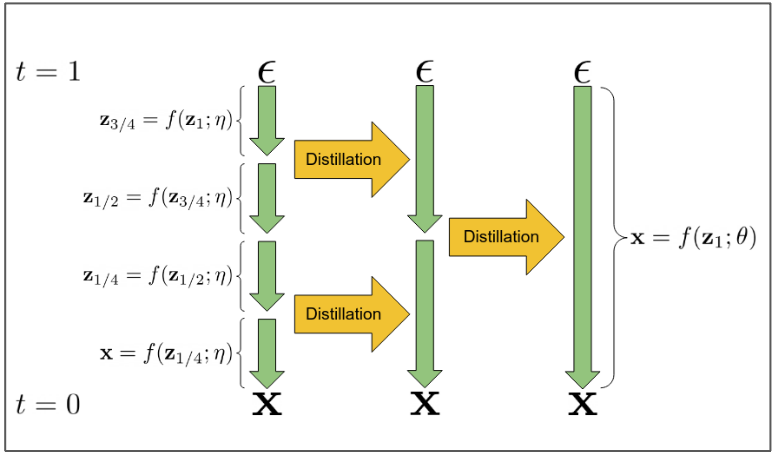
- As *w* in increases:
    - FID performance decreases
    - IS performance increases

- *w* can sacrifice sample fidelity for sample variety!

$$\hat{\mathbf{x}}_{\boldsymbol{\theta}}^{w} = (1 + w)\hat{\mathbf{x}}_{c,\boldsymbol{\theta}} - w\hat{\mathbf{x}}_{\boldsymbol{\theta}}$$

Combining Classifier Free Guidance and Distillation

- Any guesses?



$$+ \boxed{\hat{\mathbf{x}}^{w}_{\boldsymbol{\theta}} = (1 + w)\hat{\mathbf{x}}_{c,\boldsymbol{\theta}} - w\hat{\mathbf{x}}_{\boldsymbol{\theta}}} = ?$$

Combining Classifier Free Guidance and Distillation

1) Distill conditional and unconditional models into single model
2) Distill many times according to *progressive distillation*

Math Notation / DDPM Review

$$\mathbb{E}_{t\sim U[0,1],\boxed{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})},\mathbf{z}_t\sim q(\mathbf{z}_t|\mathbf{x})}[\omega(\lambda_t)||\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t)-\boxed{\mathbf{x}}||_2^2],$$

where $\lambda_t = \log[\alpha_t^2/\sigma_t^2]$ is a signal-to-noise ratio $q(\mathbf{z}_t|\mathbf{x}) = \mathcal{N}(\mathbf{z}_t; \alpha_t\mathbf{x}, \sigma_t^2\mathbf{I})$ and $\omega(\lambda_t)$ is a pre-specified weighting function

Math Notation / DDPM Review

$$\mathbb{E}_{t\sim U[0,1],\mathbf{x}\sim p_{\text{data}}(\mathbf{x}),\mathbf{z}_t\sim q(\mathbf{z}_t|\mathbf{x})}[\omega(\lambda_t)||\boxed{\hat{\mathbf{x}}_{\boldsymbol{\theta}}}(\mathbf{z}_t)-\mathbf{x}||_2^2],$$

where $\lambda_t = \log[\alpha_t^2/\sigma_t^2]$ is a signal-to-noise ratio $q(\mathbf{z}_t|\mathbf{x}) = \mathcal{N}(\mathbf{z}_t;\alpha_t\mathbf{x},\sigma_t^2\mathbf{I})$ and $\omega(\lambda_t)$ is a pre-specified weighting function

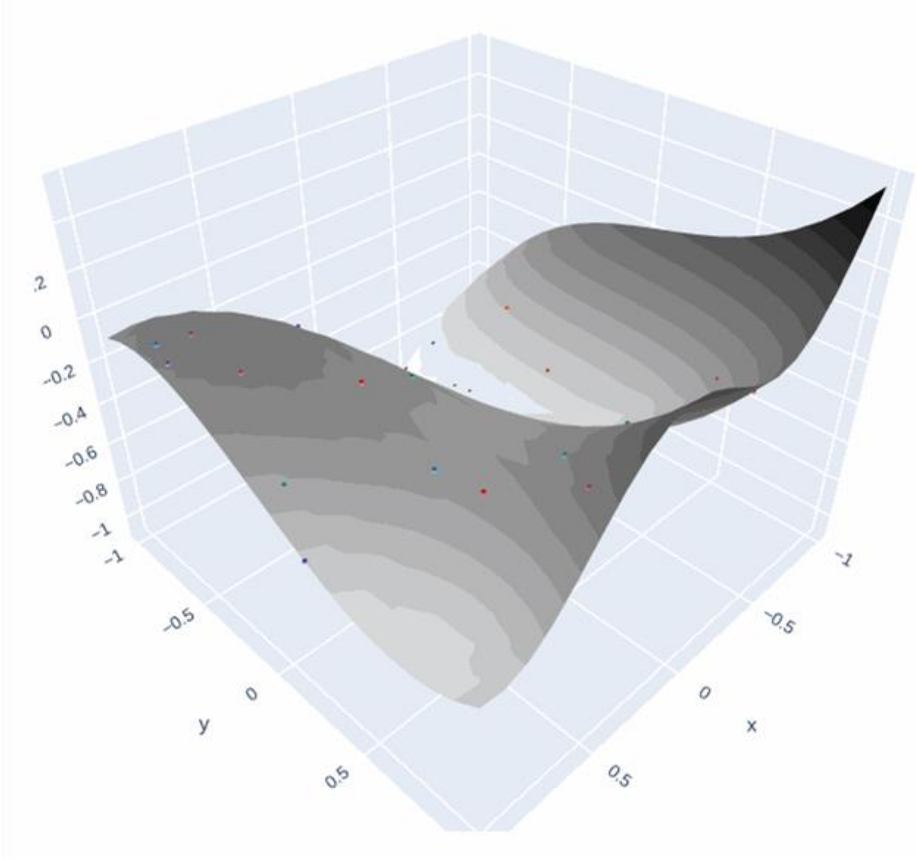
Math Notation / DDPM Review

$$\mathbb{E}_{t\sim U[0,1],\mathbf{x}\sim p_{\text{data}}(\mathbf{x}),\boxed{\mathbf{z}_t\sim q(\mathbf{z}_t|\mathbf{x})}}[\omega(\lambda_t)||\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\boxed{\mathbf{z}_t}) - \mathbf{x}||_2^2],$$

where $\lambda_t = \log[\alpha_t^2/\sigma_t^2]$ is a signal-to-noise ratio $\boxed{q(\mathbf{z}_t|\mathbf{x}) = \mathcal{N}(\mathbf{z}_t; \alpha_t\mathbf{x}, \sigma_t^2\mathbf{I})}$ and $\omega(\lambda_t)$ is a pre-specified weighting function

Note: α_t goes from 1 →0 as t passes

$$q(\mathbf{z}_t|\mathbf{x}) \;=\; \mathcal{N}(\mathbf{z}_t;\alpha_t\mathbf{x},\sigma_t^2\mathbf{I})$$



Math Notation / DDPM Review

$$\mathbb{E}_{t\sim U[0,1],\mathbf{x}\sim p_{\text{data}}(\mathbf{x}),\mathbf{z}_t\sim q(\mathbf{z}_t|\mathbf{x})}\left[\boxed{\omega(\lambda_t)}||\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t)-\mathbf{x}||_2^2\right],$$

where $\boxed{\lambda_t = \log[\alpha_t^2/\sigma_t^2]}$ is a signal-to-noise ratio $q(\mathbf{z}_t|\mathbf{x}) = \mathcal{N}(\mathbf{z}_t;\alpha_t\mathbf{x},\sigma_t^2\mathbf{I})$ and $\omega(\lambda_t)$ is a pre-specified weighting function

Combining Methods: Step 1

$$\mathbb{E}_{w \sim p_w, t \sim U[0,1], \mathbf{x} \sim p_{\text{data}}(\mathbf{x})} \left[\omega(\lambda_t) \| \boxed{\hat{\mathbf{x}}_{\boldsymbol{\eta}_1}(\mathbf{z}_t, w)} - \boxed{\hat{\mathbf{x}}^w_{\boldsymbol{\theta}}(\mathbf{z}_t)} \|_2^2 \right]$$

where $\boxed{\hat{\mathbf{x}}^w_{\boldsymbol{\theta}}(\mathbf{z}_t) = (1+w)\hat{\mathbf{x}}_{c,\boldsymbol{\theta}}(\mathbf{z}_t) - w\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t)}$, $\mathbf{z}_t \sim q(\mathbf{z}_t|\mathbf{x})$
and $p_w(w) = U[w_{\min}, w_{\max}]$.

- Remember:
  - *w* can sacrifice sample fidelity for sample variety!
  - We would like to keep this functionality

Additional Information

- Student network is initialized with teacher parameters

- How is *w* integrated?
    - Similar to how *t* is normally integrated with DDPM paper (not covered here)
    - Fourier features used instead of transformer positional embeddings
        - $\sin(2^n \pi w)$
        - $\cos(2^n \pi w)$
        - *n* runs over some range of integers

Combining Methods: Step 2

- Same as progressive distillation

- NOTE:
  - The distilled model is *deterministic*
  - But what if I want stochastic sampling??

N - Step Stochastic Sampling

- Two steps forward, one step back
- Take a distilled model
    - Eg. an N=8 step distilled model

1) Take a sampling step equal to 2 steps
    a) (or one step of an N/2 model)
    b) Add noise according to step size N

Results

- Distilled model can MATCH performance of teacher model
  - 2 or 4 sampling steps !!!

(a) 2 denoising steps (b) 4 denoising steps (c) 8 denoising steps

Figure 4. Text-guided generation on LAION (512x512) using our distilled Stable Diffusion model. Our model is able to generate high-quality image samples using 2, 4 or 8 denoising steps, significantly improving the inference efficiency of Stable Diffusion.

Distilled Text-to-Image samples (4 steps)

Native Text-to-Image samples (4 steps)

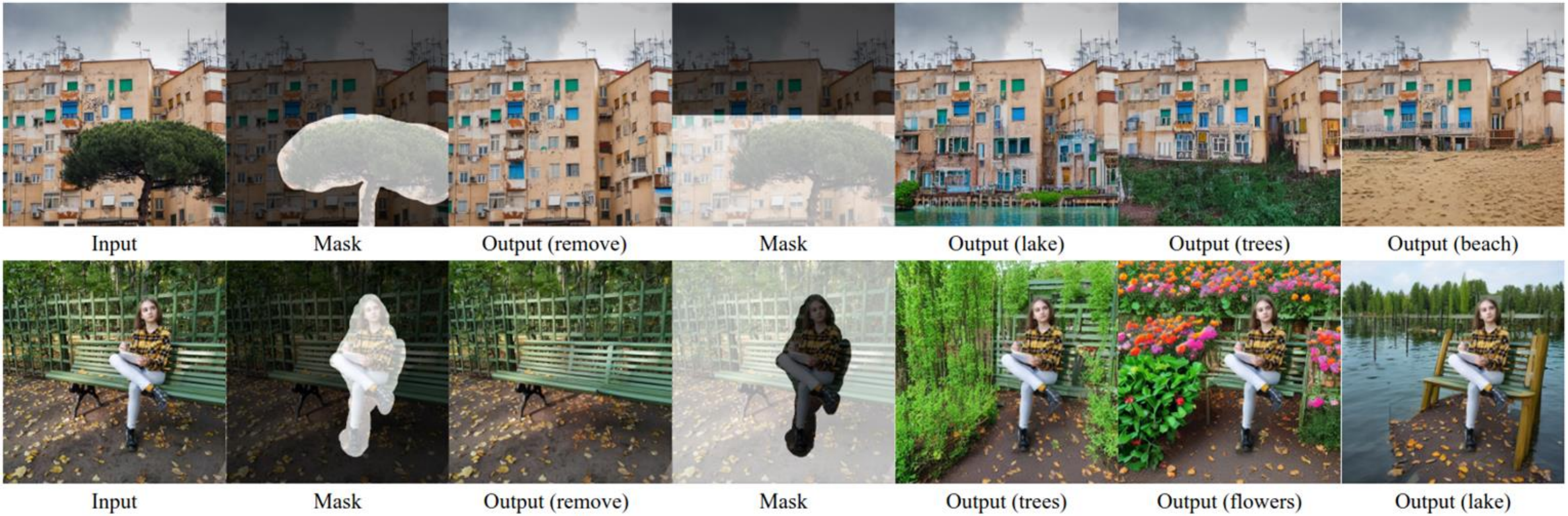


Figure 8. Image inpainting with our distilled Stable Diffusion model (4 denoising steps). Our model is able to generate high-quality image inpainting results using 4 denoising steps on unseen data.