REG: Rectified Gradient Guidance for Conditional Diffusion Models

Zhengqi Gao[1], Kaiwen Zha[1], Tianyuan Zhang[1], Zihui Xue[2], Duane S. Boning[1]

[1] Massachusetts Institute of Technology

[2] University of Texas at Austin

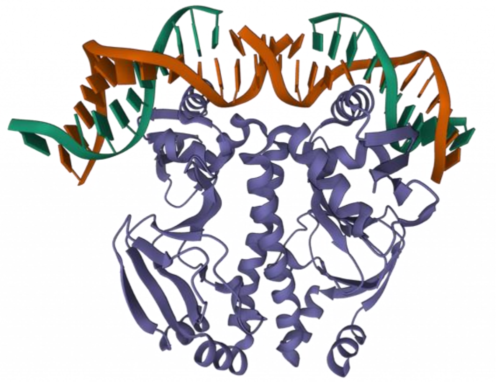
Contact: zhengqi@mit.edu

Introduction

Diffusion models have achieved great success in generative ML tasks.

Image Generation

Audio Synthesis

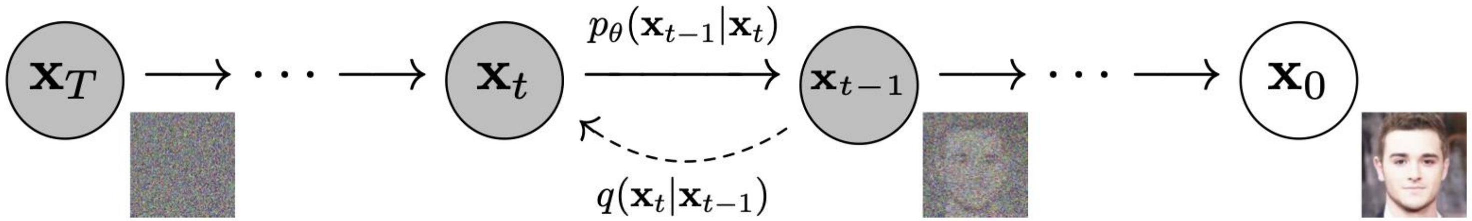


Protein Design

. . .

Preliminary

The math behind diffusion model (DDPM formulation) is a Markov chain:



Forward Noising

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t|\sqrt{\alpha}_t\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I})$$

$$q(\mathbf{x}_{0:T}|\mathbf{y}) = q(\mathbf{x}_0|\mathbf{y})\prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$$

Reverse Denoising

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y}) = \mathcal{N}(\mathbf{x}_{t-1}|\boldsymbol{\mu}_{\theta,t}, \sigma_t^2\mathbf{I})$$

$$p_\theta(\mathbf{x}_{0:T}|\mathbf{y}) = p_\theta(\mathbf{x}_T|\mathbf{y})\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y})$$

where $\boldsymbol{\mu}_{\theta,t} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{y})\right)$

Massachusetts Institute of Technology

ICML International Conference On Machine Learning

Preliminary

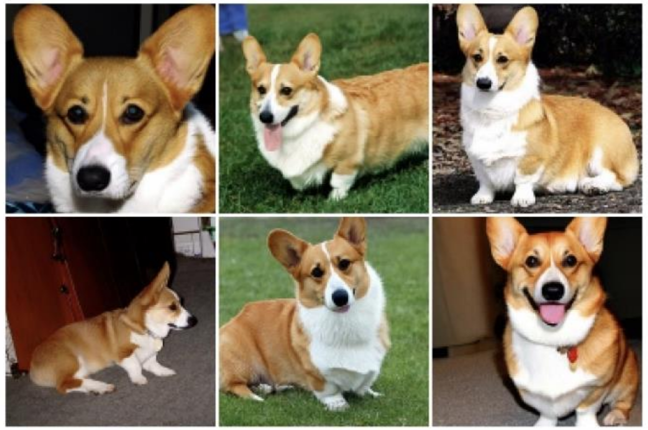
Guidance technique is critical for conditional diffusion models.

Classifier guidance (CG): $\bar{\boldsymbol{\epsilon}}_{\theta,t} = \boldsymbol{\epsilon}_{\theta,t} - w\sqrt{1-\bar{\alpha}_t}\nabla_{\mathbf{x}_t} \log p_\phi(\mathbf{y}|\mathbf{x}_t)$

Classifier free guidance (CFG): $\bar{\boldsymbol{\epsilon}}_{\theta,t} = \boldsymbol{\epsilon}_{\theta,t} + w\left(\boldsymbol{\epsilon}_{\theta,t} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right)$

Autoguidance (AutoG): $\bar{\boldsymbol{\epsilon}}_{\theta,t} = \boldsymbol{\epsilon}_{\theta,t} + w\left(\boldsymbol{\epsilon}_{\theta,t} - \boldsymbol{\epsilon}_{\theta_{\text{bad}}}(\mathbf{x}_t, t, \mathbf{y})\right)$

w/o guidance [1]



w/ CFG [1]

Massachusetts Institute of Technology

[1] Jonanthan Ho & Tim Salimans, ‘Classifier-Free Diffusion Guidance’. Arxiv, 2022.

ICML International Conference On Machine Learning

Preliminary

Original guidance motivation/theory:

Sample from marginal scaled distributions [1]: $\bar{p}_\theta(\mathbf{x}_t|\mathbf{y}) \propto p_\theta(\mathbf{x}_t|\mathbf{y}) \cdot R_t(\mathbf{x}_t, \mathbf{y})$

=> (*) $\bar{\boldsymbol{\epsilon}}_{\theta,t} = \boldsymbol{\epsilon}_{\theta,t} - \sqrt{1-\bar{\alpha}_t}\nabla_{\mathbf{x}_t} \log R_t(\mathbf{x}_t, \mathbf{y})$ using score function: $\nabla_{\mathbf{x}_t} \log p_\theta(\mathbf{x}_t|\mathbf{y}) = -\frac{\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{y})}{\sqrt{1-\bar{\alpha}_t}}$

With the following scale factors, recover the equations in previous page.

CG: $R_t(\mathbf{x}_t, \mathbf{y}) = [p_{\phi, X_t}(\mathbf{y}|\mathbf{x}_t)]^w$

CFG: $R_t(\mathbf{x}_t, \mathbf{y}) = \left[\frac{p_{\theta, X_t}(\mathbf{x}_t|\mathbf{y})}{p_{\theta, X_t}(\mathbf{x}_t)}\right]^w$

AutoG: $R_t(\mathbf{x}_t, \mathbf{y}) = \left[\frac{p_{\theta, X_t}(\mathbf{x}_t|\mathbf{y})}{p_{\theta_{\text{bad}}, X_t}(\mathbf{x}_t|\mathbf{y})}\right]^w$

Problem: Cannot specificy all R_t since $R_{t-1}(\mathbf{x}_{t-1}, \mathbf{y}) \propto \frac{\mathbb{E}\left[\mathcal{N}(\mathbf{x}_{t-1}|\bar{\boldsymbol{\mu}}_{\theta,t}, \sigma_t^2\mathbf{I})R_t(\mathbf{x}_t, \mathbf{y})\right]}{\mathbb{E}\left[\mathcal{N}(\mathbf{x}_{t-1}|\boldsymbol{\mu}_{\theta,t}, \sigma_t^2\mathbf{I})\right]}$

Rectified Gradient Guidance

Correct formulation w/ joint scaling: $\bar{p}_\theta(\mathbf{x}_{0:T}|\mathbf{y}) \propto p_\theta(\mathbf{x}_{0:T}|\mathbf{y}) \cdot R_0(\mathbf{x}_0, \mathbf{y})$

Theorem 1: To satisfy this scaled goal, we must have a unique set of transition kernels:

$$\bar{p}_\theta(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}) = \frac{E_t(\mathbf{x}_t, \mathbf{y})}{E_{t+1}(\mathbf{x}_{t+1}, \mathbf{y})} p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}), \quad E_t(\mathbf{x}_t, \mathbf{y}) = \int p_\theta(\mathbf{x}_0|\mathbf{x}_t, \mathbf{y}) R_0(\mathbf{x}_0, \mathbf{y})\, d\mathbf{x}_0$$

It implies the noise prediction network should be: $\bar{p}_\theta(\mathbf{x}_t|\mathbf{y}) = \frac{E_t(\mathbf{x}_t, \mathbf{y})}{E(\mathbf{y})} p_\theta(\mathbf{x}_t|\mathbf{y})$

where $t = 0, 1, \ldots, T$ and $x_T = \varnothing$, which determines:

$$\bar{\boldsymbol{\epsilon}}^\star_{\theta,t} = \boldsymbol{\epsilon}_{\theta,t} - \sqrt{1 - \bar{\alpha}_t} \nabla_{\mathbf{x}_t} \log E_t(\mathbf{x}_t, \mathbf{y}) \qquad (*)$$

Rectified Gradient Guidance

- Present implementation (*) compared with golden (*): off by one term, R_t should be E_t.
- See our paper for theoretical bounds on the gap between (*) and (*)
- Since (*) is only an approximation to (*), is there an even better approximation?

$$\bar{\epsilon}_{\theta,t}^{\text{REG}} = \epsilon_{\theta,t} - \sqrt{1-\bar{\alpha}_t}\nabla_{\mathbf{x}_t} \log R_t(\mathbf{x}_t, \mathbf{y}) \odot \underbrace{\left(1 - \sqrt{1-\bar{\alpha}_t}\frac{\partial(\mathbf{1}^T \cdot \epsilon_{\theta,t})}{\partial \mathbf{x}_t}\right)}_{\text{REG correction term}}$$

Numerical Results

Class-conditional ImageNet generation

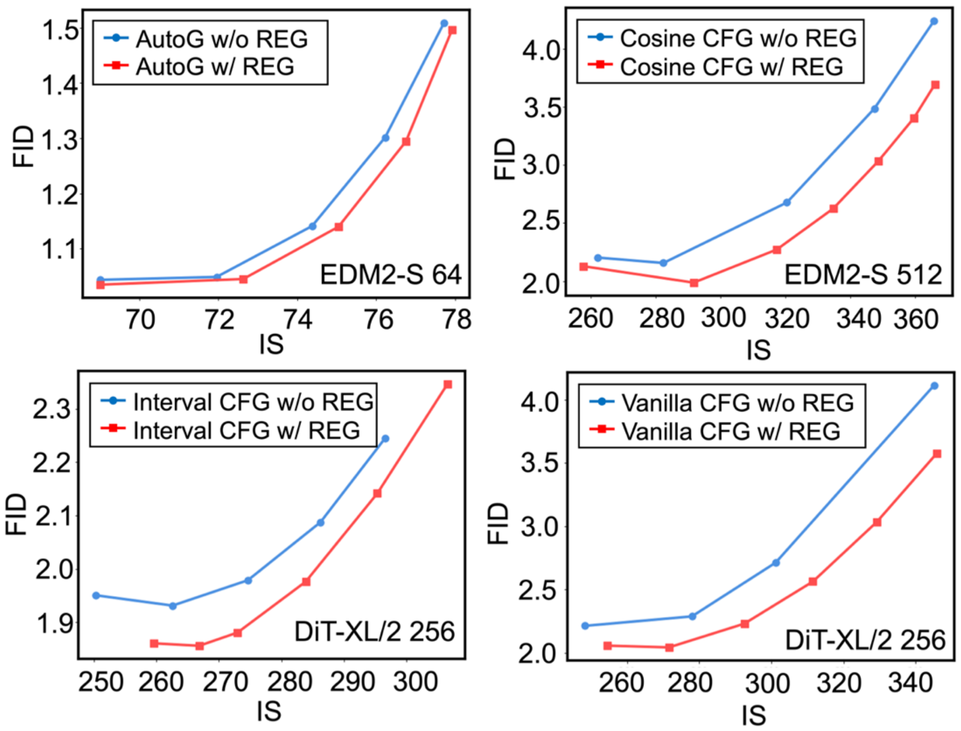


Figure. Pareto Front of FID v.s. IS when sweeping guidance weight.

Resolution	Benchmark	FID ↓	IS ↑
64×64	**EDM2-S**	1.580	——
	+ AutoG	1.044	69.01
	+ REG (ours)	1.035 ↓	69.01
256×256	**DiT-XL/2**	9.62	121.50
	+ Vanilla CFG	2.21	248.36
	+ REG (ours)	2.04 ↓	276.26 ↑
	+ Cosine CFG	2.30	300.73
	+ REG (ours)	1.76 ↓	287.48
	+ Linear CFG	2.23	268.69
	+ REG (ours)	2.18 ↓	284.20 ↑
	+ Interval CFG	1.95	250.44
	+ REG (ours)	1.86 ↓	259.57 ↑
512×512	**EDM2-S**	2.56	——
	+ Vanilla CFG	2.29	268.56
	+ REG (ours)	2.02 ↓	275.30 ↑
	+ Cosine CFG	2.16	282.46
	+ REG (ours)	1.99 ↓	291.77 ↑
	+ Linear CFG	2.21	282.89
	+ REG (ours)	1.99 ↓	291.04 ↑
	+ Interval CFG	1.67	287.45
	+ REG (ours)	1.67	288.43 ↑
	EDM2-XXL	1.91	——
	+ Vanilla CFG	1.83	265.76
	+ REG (ours)	1.74 ↓	289.24 ↑
	+ Cosine CFG	1.80	261.94
	+ REG (ours)	1.69 ↓	268.84 ↑
	+ Linear CFG	1.81	262.03
	+ REG (ours)	1.69 ↓	268.30 ↑
	+ Interval CFG	1.45	283.26
	+ REG (ours)	1.45	288.72 ↑

Table. Class-conditional ImageNet generation results

Massachusetts Institute of Technology

ICML International Conference On Machine Learning

Numerical Results

Text-to-Image generation on COCO 2017-5k

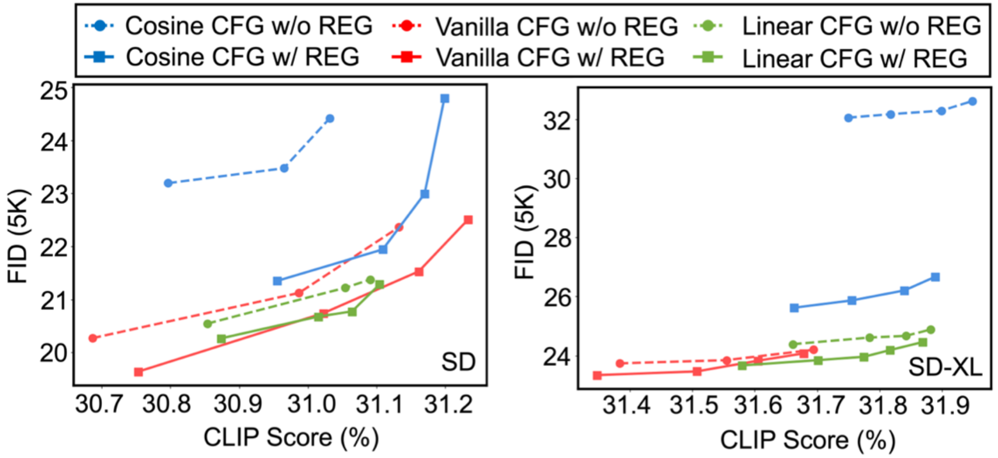


Figure. Pareto Front of FID v.s. CLIP when sweeping guidance weight.

Model	Benchmark	FID ↓	CLIP (%) ↑
SD-v1-4 512×512	+ Vanilla CFG	20.27	30.68
	+ REG (ours)	19.63 ↓	30.75 ↑
	+ Cosine CFG	23.19	30.80
	+ REG (ours)	21.35 ↓	30.96 ↑
	+ Linear CFG	20.55	30.85
	+ REG (ours)	20.27 ↓	30.87 ↑
SD-XL 1024×1024	+ Vanilla CFG	23.73	31.38
	+ REG (ours)	23.46 ↓	31.51 ↑
	+ Cosine CFG	32.14	31.58
	+ REG (ours)	25.62 ↓	31.66 ↑
	+ Linear CFG	24.43	31.55
	+ REG (ours)	23.67 ↓	31.58 ↑

Table. Text-to-Image generation results

Massachusetts Institute of Technology

ICML International Conference On Machine Learning

Conclusions

- We identify the flaw in present guidance theory for conditional diffusion models.
- We propose the correct guidance theory from scaling the joint distribution.
- The theory inspired REG method can consistently boost existing guidance methods
  - At the cost of memory and runtime increasing

Model	# Param	Sampler	Prediction
DiT-XL/2	675 M	250-step DDPM	ϵ-prediction
EDM2-S	280 M	2nd Heun	$\mathbf{x}_0$-prediction
EDM2-XXL	1.5 B	2nd Heun	$\mathbf{x}_0$-prediction
SD-v1-4	860 M	PNDM	ϵ-prediction
SD-XL	2.6 B	Euler Discrete	ϵ-prediction

Table. Summary of models used in our experiment.

Model	Resolution / BS	CFG / REG Runtime (sec)
EDM2-S	64 / 8	25.96 / 42.99 (1.66×)
DiT-XL/2	256 / 8	59.79 / 94.23 (1.58×)
EDM2-S	512 / 8	46.14 / 62.87 (1.36×)
EDM2-XXL	512 / 8	49.21 / 92.60 (1.88×)
SD-v1-4	512 / 4	32.63 / 39.54 (1.21×)
SD-XL	1024 / 2	47.48 / 74.52 (1.57×)
Model	Resolution / BS	CFG / REG Memory (GB)
EDM2-S	64 / 1	0.87 / 1.49 (1.71×)
DiT-XL/2	256 / 1	4.15 / 5.01 (1.21×)
EDM2-S	512 / 1	1.19 / 1.81 (1.52×)
EDM2-XXL	512 / 1	4.59 / 7.31 (1.59×)
SD-v1-4	512 / 1	2.73 / 4.39 (1.61×)
SD-XL	1024 / 1	6.91 / 19.49 (2.82×)

Table. Memory and runtime overhead.