GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models

July 14, Jeong Hyun Jae, Journal Club

Hyunjae Jeong (tAILab, CCIDS)

Yonsei University, Medical Life Systems Information Center (TAIL Lab)
Severance Hospital, Center for Clinical Imaging Data Science (CCIDS)
Severance Hospital, Radiology

CCIDS 의료영상데이터사이언스센터 Center for Clinical Imaging Data Science

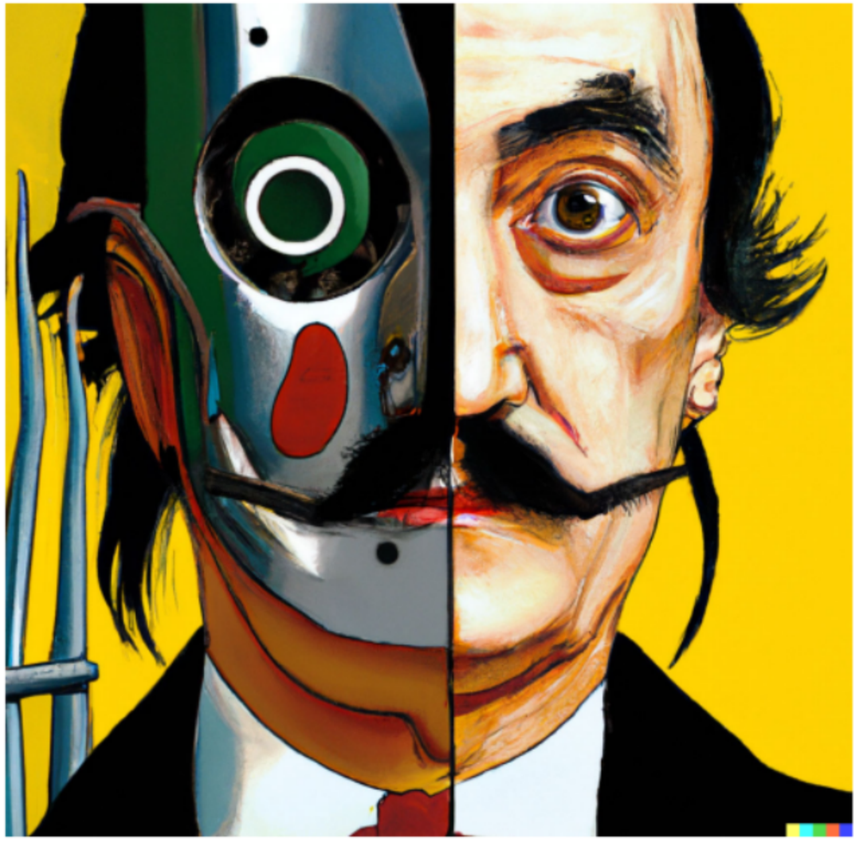
tAILab.

01 Introduction

Text-to-Image Generation

- Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing.

A teddy bear on a skateboard in times square



Vibrant portrait painting of Salvador Dalí with a robotic half face

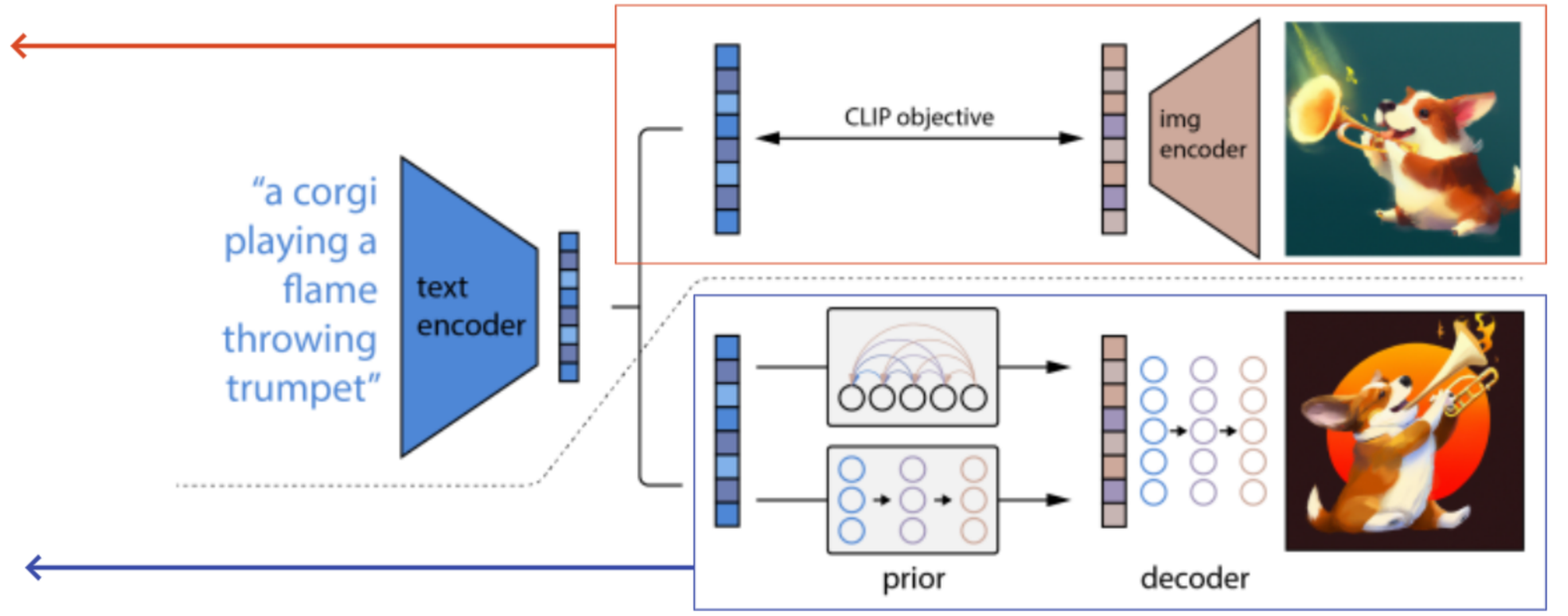
01 Introduction

Previous Presentation - DALLE 2

- unCLIP

  - Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process:

CLIP : Learn the joint representation space of text images



Text-to-image generation process : The image embedding is generated from the text encoder of CLIP into the prior, the embedding is converted into the final image through the decoder.

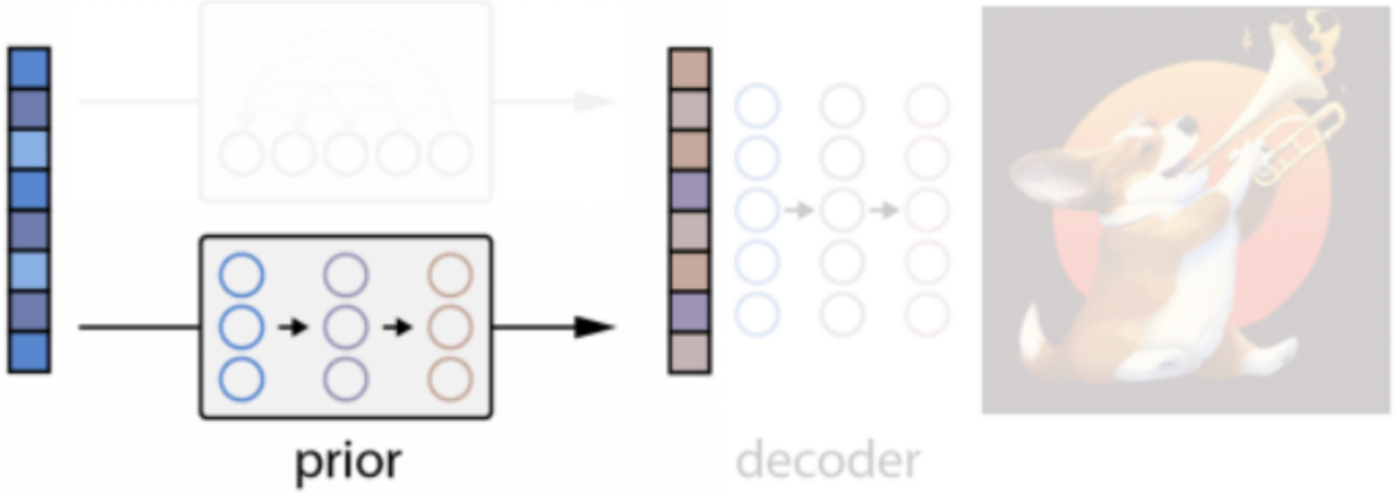
Proposed Method : unCLIP

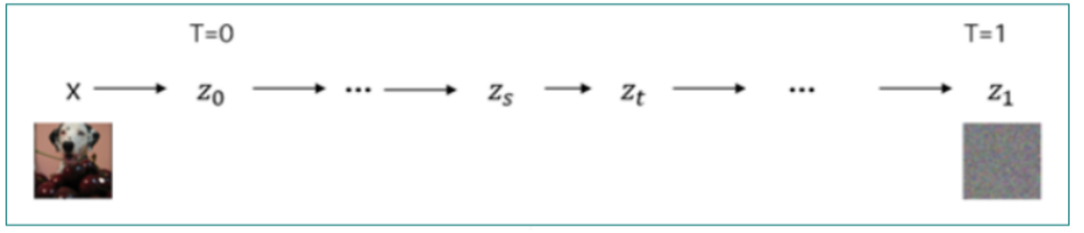
01 Introduction

Previous Presentation - Diffusion model

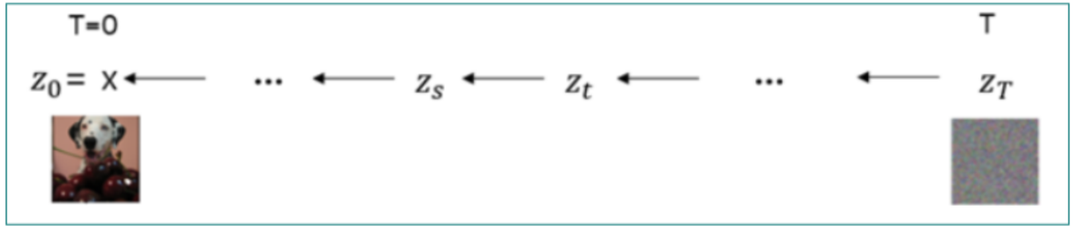
- A continuous vector z(i) is constructed through a Gaussian diffusion model in which the caption y is given as a condition. (decoder-only Transformer)
- In order : the encoded text, the CLIP text embedding, an embedding for the diffusion timestep, the noised CLIP image embedding, and a final embedding whose output from the Transformer is used to predict the unnoised CLIP image embedding.

$$L_{\text{prior}} = \mathbb{E}_{t\sim[1,T],z_i^{(t)}\sim q_t}\left[\|f_\theta(z_i^{(t)}, t, y) - z_i\|^2\right]$$





Diffusion Process



Reverse Diffusion Process

01 Introduction

Related Journal

Diffusion Models Beat GANs on Image Synthesis !

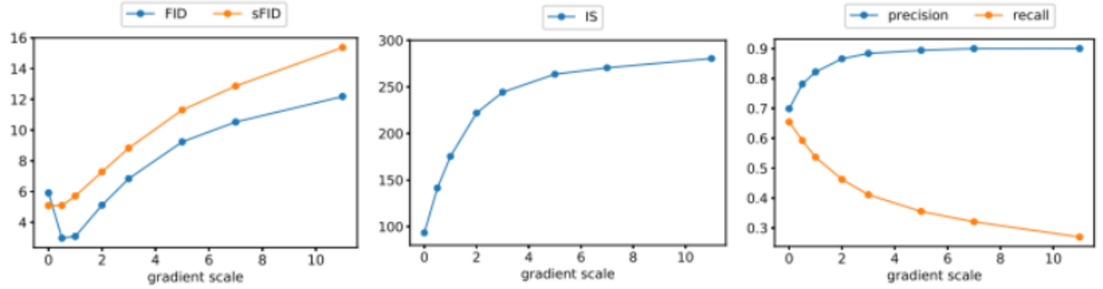
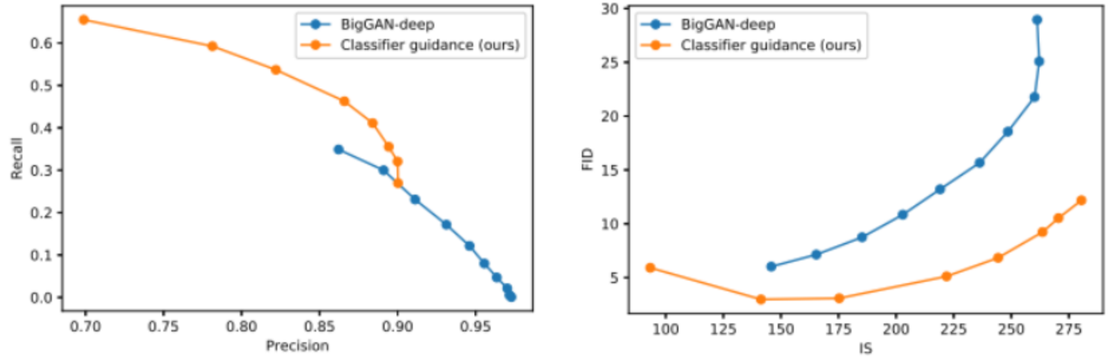


Figure 4: Change in sample quality as we vary scale of the classifier gradients for a class-conditional ImageNet 128×128 model.



Diffusion Models Beat GANs on Image Synthesis

Prafulla Dhariwal*
OpenAI
prafulla@openai.com

Alex Nichol*
OpenAI
alex@openai.com

Abstract

We show that diffusion models can achieve image sample quality superior to the current state-of-the-art generative models. We achieve this on unconditional image synthesis by finding a better architecture through a series of ablations. For conditional image synthesis, we further improve sample quality with classifier guidance: a simple, compute-efficient method for trading off diversity for fidelity using gradients from a classifier. We achieve an FID of 2.97 on ImageNet 128×128, 4.59 on ImageNet 256×256, and 7.72 on ImageNet 512×512, and we match BigGAN-deep even with as few as 25 forward passes per sample, all while maintaining better coverage of the distribution. Finally, we find that classifier guidance combines well with upsampling diffusion models, further improving FID to 3.94 on ImageNet 256×256 and 3.85 on ImageNet 512×512. We release our code at https://github.com/openai/guided-diffusion.

1 Introduction

Figure 1: Selected samples from our best ImageNet 512×512 model (FID 3.85)

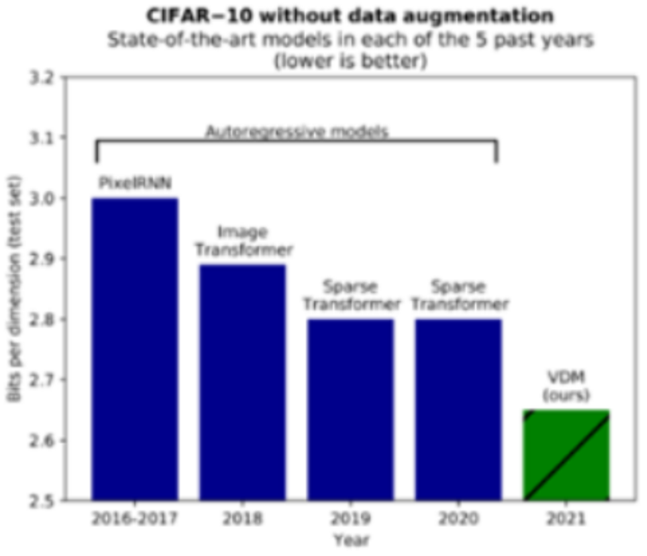
NeurIPS 2021 Spotlight

01 Introduction

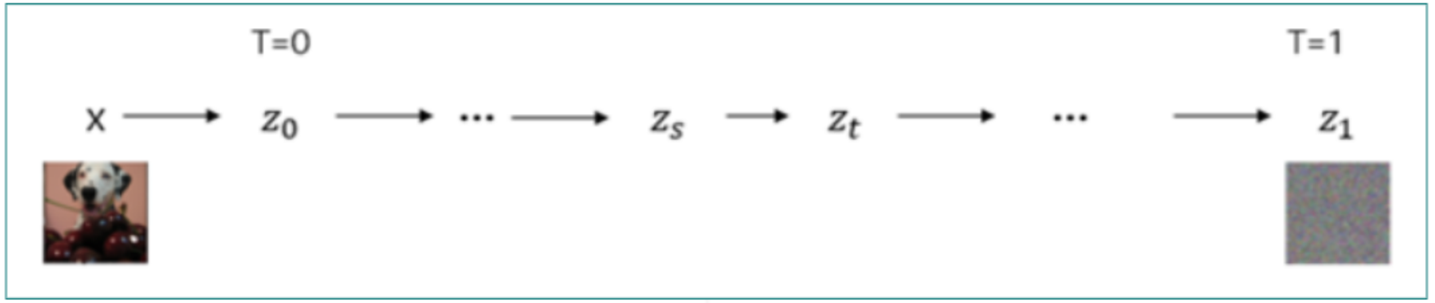
Diffusion Model

- Diffusion Model

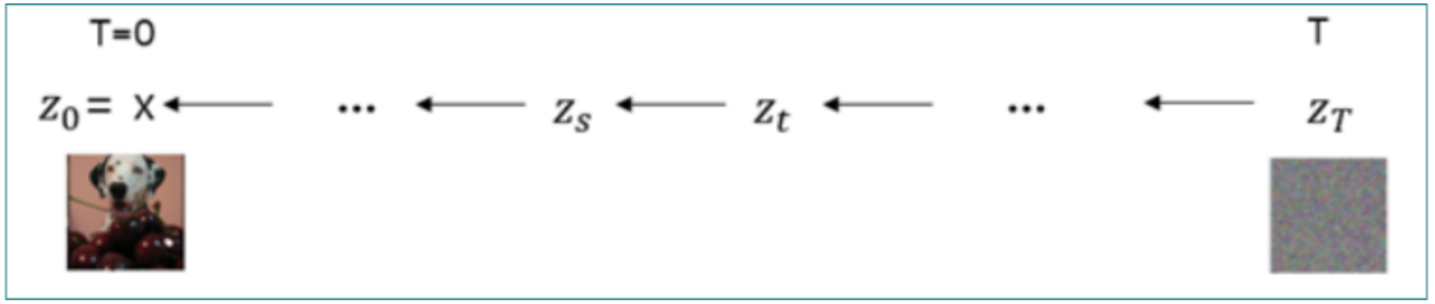
  - Diffusion model is a model inspired by thermodynamics and is largely divided into two stages. First, we gradually ad d noise to the given data x. This process is called the diffusion process. Then, the process that reverses the diffusion process defined earlier is calculated. This process is the process of gradually removing noise from noise data.



(a) CIFAR-10 without data augmentation



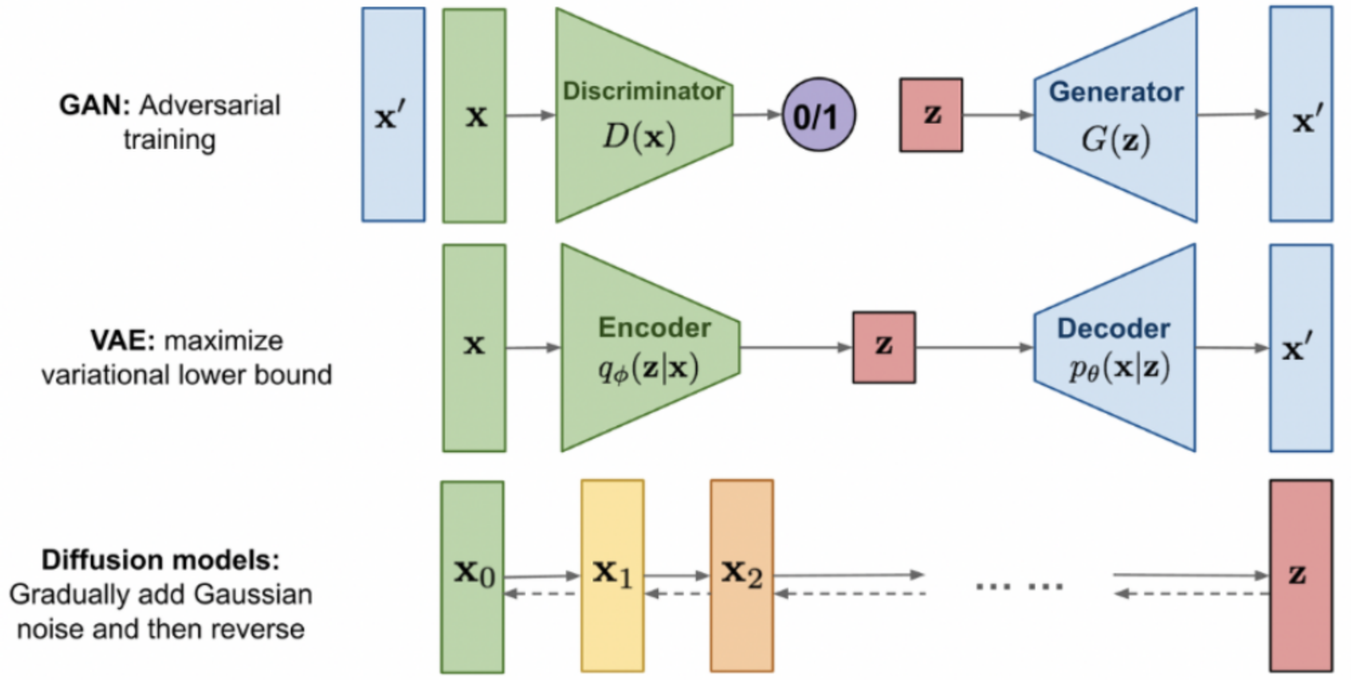
Diffusion Process



Reverse Diffusion Process

01 Introduction

Diffusion Model



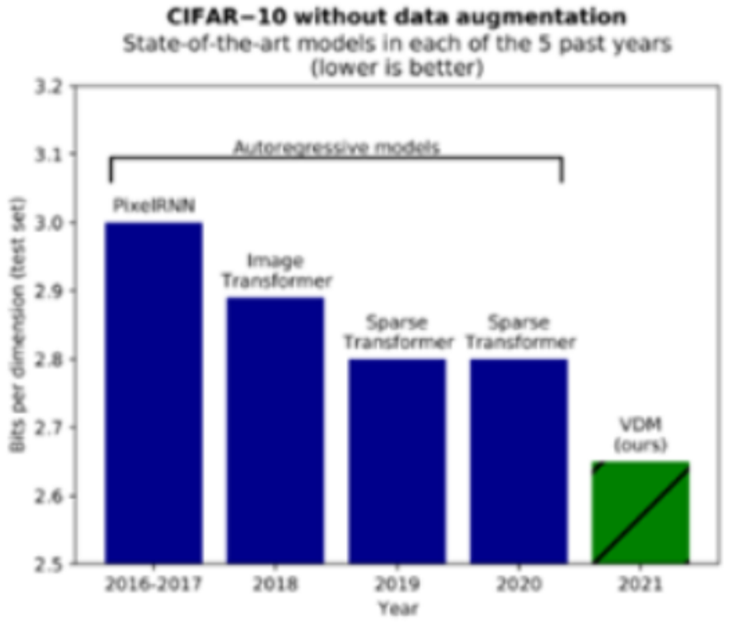
$$q(x_t|x_{t-1}) \coloneqq \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathcal{I})$$

01 Introduction

Diffusion Journals at a glance

- 1. Generative model - Diffusion model Review (2021 NeurIPS)

  - Diffusion model is a model inspired by thermodynamics and is largely divided into two stages. First, we gradually ad d noise to the given data x. This process is called the diffusion process. Then, the process that reverses the diffusion process defined earlier is calculated. This process is the process of gradually removing noise from noise data.



(a) CIFAR-10 without data augmentation

- It is proposed to use Fourier features to improve the performance of Likelihood.
- Theoretically, it is proved that the more the number of steps co mposing the diffusion process, the better the performance.

$$\mathcal{L}_\infty(\mathbf{x}) = -\frac{1}{2}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})}\int_0^1 \mathrm{SNR}'(t)\,\|\mathbf{x}-\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t;t)\|_2^2\,dt,$$

$$= -\frac{1}{2}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I}),t\sim\mathcal{U}(0,1)}\left[\mathrm{SNR}'(t)\,\|\mathbf{x}-\hat{\mathbf{x}}_{\boldsymbol{\theta}}(\mathbf{z}_t;t)\|_2^2\right]$$

01 Introduction

Diffusion Journals at a glance

- 2. Diffusion Models Beat GANs on Image Synthesis (NeurIPS 2021)

  - we find that classifier guidance combines well with upsampling diffusion models, further improving FID to 3.94 on ImageNet 256×256 and 3.85 on ImageNet 512×512.

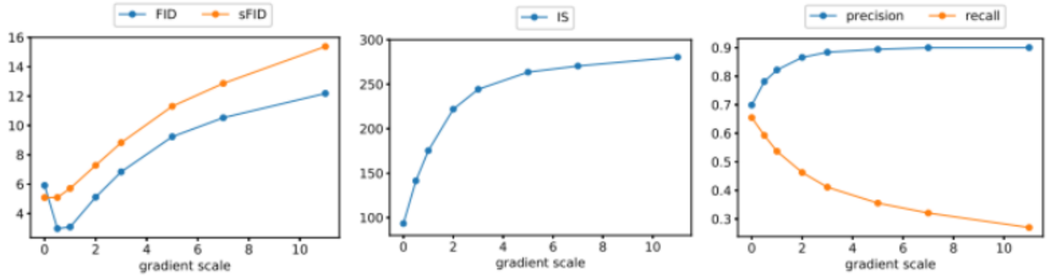
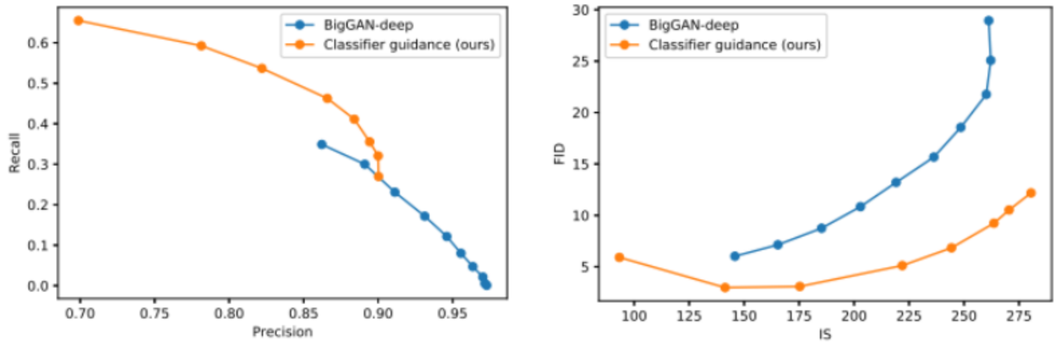


Figure 4: Change in sample quality as we vary scale of the classifier gradients for a class-conditional ImageNet 128×128 model.



GAN vs Diffusion Model vs Training Dataset

01 Introduction

GLIDE : Classifier Guidance for Photorealistic image

- Main Goal : Photorealistic Image and text condition trade-off

  - The unconditional image model can create a photorealistic image, but it cannot create an image that reflects text, and the conditional image such as text-to-image can reflect text but cannot make it photorealistic.

  - Recently, text-conditional image models that take a free-form text prompt as input and generate related images have been proposed. However, it is not yet possible to create a photorealistic image that fully reflects the information of the text. ⟶ Proposed model : guided diffusion model

(a) DALL-E (Temp 0.85, CLIP reranked top 16 out of 512)

(d) GLIDE (Classifier-free guidance, scale 3.0)

02 Methods

GLIDE : A way to improve performance at conditional image

- Guided Diffusions

  -Given the image xt, when the classifier predicting class y is called p(ylxt), the gradient of the classifier is pr eturbated in addition to the mean and variance of the diffusion model. Also, s is called the guidance scale, a nd as this value is increased, the diversity of the generated image decreases, but the quality of the image in creases.

$$\hat{\epsilon}_\theta(x_t|c) = \epsilon_\theta(x_t|\emptyset) + s \cdot (\epsilon_\theta(x_t|c) - \epsilon_\theta(x_t|\emptyset))$$

- Classifier-Free Guidance

  - Proposed classifier-free guidance, a technique for guiding diffusion models that does not require a separate classifier model to be trained

- CLIP Guidance

  - We perturb the reverse-process mean with the gradient of the dot product of the image and caption encodings with respect to the image

03 Experiments

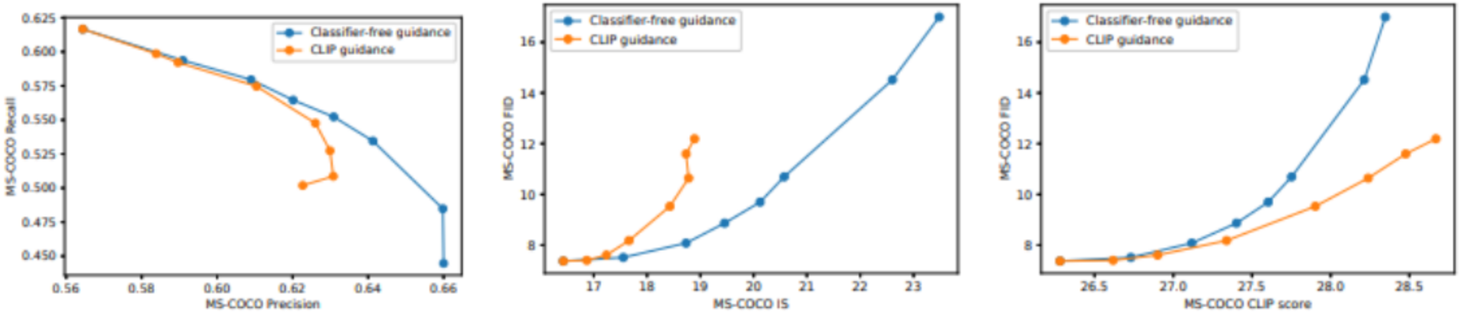
Qualitative Results

03 Experiments

Quantitative Results

Guidance	Photorealism	Caption
Unguided	-88.6	-106.2
CLIP guidance	-73.2	29.3
Classifier-free guidance	**82.7**	**110.9**

Model	FID	Zero-shot FID
AttnGAN (Xu et al., 2017)	35.49	
DM-GAN (Zhu et al., 2019)	32.64	
DF-GAN (Tao et al., 2020)	21.42	
DM-GAN + CL (Ye et al., 2021)	20.79	
XMC-GAN (Zhang et al., 2021)	9.33	
LAFITE (Zhou et al., 2021)	**8.12**	
DALL-E (Ramesh et al., 2021)		~ 28
LAFITE (Zhou et al., 2021)		26.94
GLIDE		**12.24**
GLIDE (Validation filtered)		**12.89**



(a) Precision/Recall (b) IS/FID (c) CLIP score/FID

	DALL-E Temp.	Photo-realism	Caption Similarity
No reranking	1.0	91%	83%
	0.85	84%	80%
DALL-E reranked	1.0	89%	71%
	0.85	87%	69%
DALL-E reranked + GLIDE blurred	1.0	72%	63%
	0.85	66%	61%