

# Diffusion Models Beat GANs on Image Synthesis

Dhariwal & Nichol (2021)

# Background

- At the time, GANs were considered the state-of-the-art method for image generation tasks
  - BUT:
    - GANs captured less diversity in image synthesis than likelihood-based models
    - GANs are difficult to train and prone to collapse
    - Consequently, difficult to scale and apply to new domains
- Competing likelihood-based models captured more diversity and were easier to train
  - BUT:
    - Worse visual fidelity compared to GANs
    - Apart from VAEs, slower to sample than GANs

# Diffusion Models (at the time)

- Pros: high quality images, good distribution coverage, stationary training objective, easy scalability
- At the time, held best performance on CIFAR-10, but lagged behind GANs for more difficult datasets (LSUN, ImageNet)

Why?



**ISSUE:** Less explored architectures = less refinement in performance

**ISSUE:** GANs can trade diversity for visual fidelity, unexplored for diffusion models

# Model Architecture

- UNet Base (as with prior diffusion models)
- Added multi-resolution attention (32 x 32, 16 x 16, 8 x 8) -> Better capture long-range relationships like object shape/symmetry
- Stronger residual blocks for better up/down sampling and stable gradients -> refined details and sharper images
- Adaptive group Normalization (AdaGN) to inject class information and timestep into normalization layers -> network can adapt processing depending on denoising stage
- Attention head tuning -> stable training and finer-grained attention patterns

**ISSUE:** Less explored architectures = less refinement in performance

**FIX:** Finetuned architecture to narrow engineering gap between diffusion and GANs

# Classifier Guidance

**ISSUE:** GANs can trade diversity for fidelity, unexplored in diffusion models

- Train a classifier on noisy images and use its gradients to steer the diffusion process toward a target class
  - Classifier:  $p_\theta(y|x_t, t)$
  - Modify the reverse diffusion step using  $\nabla_{x_t} \log p_\theta(y|x_t, t)$
- Guidance Scaling:
  - Introduce a scaling factor  $s$  to control strength of steering of classifier gradient
    - Low  $s \rightarrow$  higher diversity, lower fidelity
    - High  $s \rightarrow$  higher fidelity, lower diversity
- This is actually an improvement over GANs:
  - Single parameter that can tune trade-off

**FIX:** Implemented classifier guidance for tunable fidelity-diversity trade-off

# **CODE DEMONSTRATION**

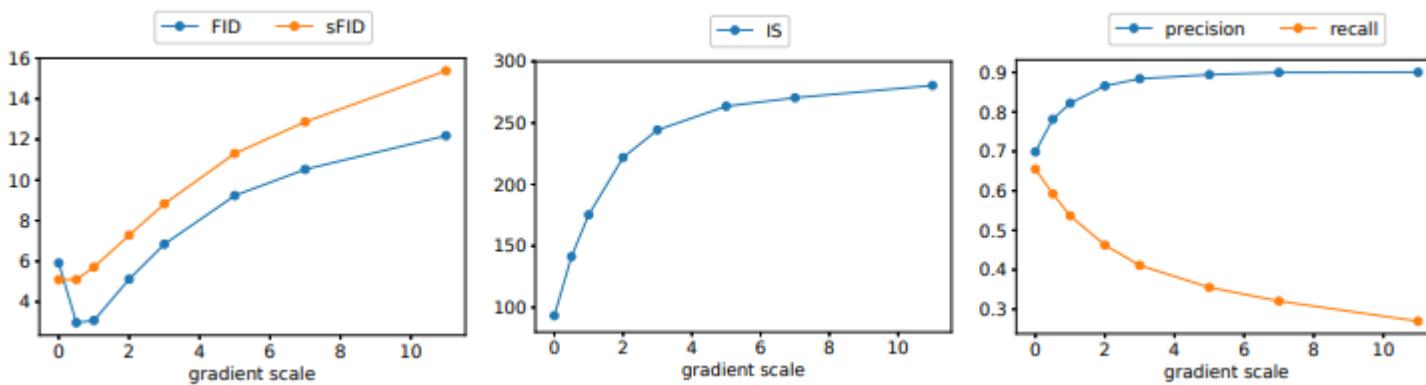
## Prompt: “Pembroke Welsh Corgi”



**Classifier scale 1.0:** More diverse samples, but less convincing



**Classifier scale 10.0:** Less diverse samples, but more convincing



- FID: *captures both fidelity and diversity*
- sFID: *captures spatial relationships within image*
- Precision: *proxy for separate measure of fidelity*
- Inception Score (IS): *proxy for fidelity*
- Recall: *proxy for separate measure of diversity*

# Performance Comparisons

Table 4: Sample quality comparison with state-of-the-art generative models for each task. LSUN diffusion models are sampled using 1000 steps (see Appendix D). ImageNet diffusion models are sampled using 250 steps, except when we use the DDIM sampler with 25 steps. \*No BigGAN-deep model was available at this resolution, so we trained our own. <sup>†</sup>Values are taken from a previous paper, due to lack of public models or samples. <sup>‡</sup>Results use two-resolution stacks. <sup>§</sup>Results use compute-intensive classifier rejection sampling.

Model	FID	sFID	Prec	Rec
<b>LSUN Bedrooms 256×256</b>				
DCTransformer <sup>†</sup> [48]	6.40	6.66	0.44	<b>0.56</b>
DDPM [31]	4.89	9.07	0.60	0.45
IDDPD [49]	4.24	8.21	0.62	0.46
StyleGAN [33]	2.35	6.62	0.59	0.48
ADM (dropout)	<b>1.90</b>	<b>5.59</b>	<b>0.66</b>	0.51
<b>LSUN Horses 256×256</b>				
StyleGAN2 [34]	3.84	6.46	0.63	0.48
ADM	2.95	<b>5.94</b>	0.69	<b>0.55</b>
ADM (dropout)	<b>2.57</b>	6.81	<b>0.71</b>	<b>0.55</b>
<b>LSUN Cats 256×256</b>				
DDPM [31]	17.1	12.4	0.53	0.48
StyleGAN2 [34]	7.25	<b>6.33</b>	0.58	0.43
ADM (dropout)	<b>5.57</b>	6.69	<b>0.63</b>	<b>0.52</b>

Model	FID	sFID	Prec	Rec
<b>ImageNet 128×128</b>				
BigGAN-deep [8]	6.02	7.18	<b>0.86</b>	0.35
LOGAN <sup>†</sup> [74]	3.36			
ADM	5.91	<b>5.09</b>	0.70	<b>0.65</b>
ADM-G (25 steps)	5.98	7.04	0.78	0.51
ADM-G	<b>2.97</b>	<b>5.09</b>	0.78	0.59
<b>ImageNet 256×256</b>				
DCTransformer <sup>†</sup> [48]	36.51	8.24	0.36	<b>0.67</b>
VQ-VAE-2 <sup>†‡</sup> [57]	31.11	17.38	0.36	0.57
VQ-VAE-2 (RS) <sup>†‡§</sup> [57]	~ 10			
VQ-GAN <sup>†</sup> [21]	15.97	19.05	0.63	0.58
VQ-GAN (RS) <sup>†‡</sup> [21]	5.06	7.34	0.79	0.48
IDDPD <sup>†</sup> [49]	12.26	5.42	0.70	0.62
SR3 <sup>†‡</sup> [60]	11.30			
BigGAN-deep [8]	6.95	7.36	<b>0.87</b>	0.28
ADM	10.94	6.02	0.69	0.63
ADM-G (25 steps)	5.44	5.32	0.81	0.49
ADM-G	<b>4.59</b>	<b>5.25</b>	0.82	0.52
<b>ImageNet 512×512</b>				
BigGAN-deep [8]	8.43	8.13	<b>0.88</b>	0.29
ADM	23.24	10.19	0.73	<b>0.60</b>
ADM-G (25 steps)	8.41	9.67	0.83	0.47
ADM-G	<b>7.72</b>	<b>6.57</b>	0.87	0.42

**Unconditional image generation:** Trained 3 separate diffusion models on bedroom, horse, and cat LSUN classes

**Classifier guidance:** trained conditional diffusion models on ImageNet 128x128, 256x256, and 512x512 resolutions

**Guidance comparisons:** trained two-stage diffusion models by combining a low-resolution diffusion model with a corresponding up-sampling diffusion model

Table 5: Comparing our single, upsampling and classifier guided models. The upsamplers are 64→256 and 128→512. When combining guidance with upsampling, we only guide the lower resolution model. All models are sampled using 250 sampling steps.

Model	FID	sFID	IS	Prec	Rec	Model	FID	sFID	IS	Prec	Rec
<b>ImageNet 256×256</b>						<b>ImageNet 512×512</b>					
ADM	10.94	6.02	100.98	0.69	<b>0.63</b>	ADM	23.24	10.19	58.06	0.73	0.60
ADM, ADM-U	7.49	<b>5.13</b>	127.49	0.72	<b>0.63</b>	ADM, ADM-U	9.96	<b>5.62</b>	121.78	0.75	<b>0.64</b>
ADM-G	4.59	5.25	186.70	0.82	0.52	ADM-G	7.72	6.57	172.71	<b>0.87</b>	0.42
ADM-G, ADM-U	<b>3.94</b>	6.14	<b>215.84</b>	<b>0.83</b>	0.53	ADM-G, ADM-U	<b>3.85</b>	5.86	<b>221.72</b>	0.84	0.53

# Results

Samples speak for themselves:



Figure 4: Samples from BigGAN-deep with truncation 1.0 (FID 6.95, left) vs samples from our diffusion model with guidance (FID 4.59, middle) and samples from the training set (right).

# Critical Reflection

## Limitations:

- Still much slower than GANs
- Implicit latent representation => unclear semantic meaning
  - Difficult to guide these models for user-prompted generation like image editing
  - Limited to labelled datasets

## Impact:

- HUGE!!! Pivotal paper for image generation and predecessor of classifier-free guidance (e.g., Dall-E 2)
- Created more accessible, cost-efficient generation model that outperforms GANs on complex datasets
- Increasing danger of fake news and doctored photos

guided diffusion no longer samples from the true data distribution

- we are biasing the reverse diffusion process towards a class, and that bias was not present in the actual dataset or during training.
- This isn't such a problem for image synthesis, as you can see that it is able to generate very high quality images regardless
- but we can't use this method if we actually want to learn and reproduce the underlying data structure

# References

- Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, 6840-6851.
- **Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34, 8780-8794.**
- Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.
- Many thanks to this Kaggle Notebook:  
<https://www.kaggle.com/code/vikramssandu/guided-diffusion-by-openai-from-scratch/notebook#Generate>