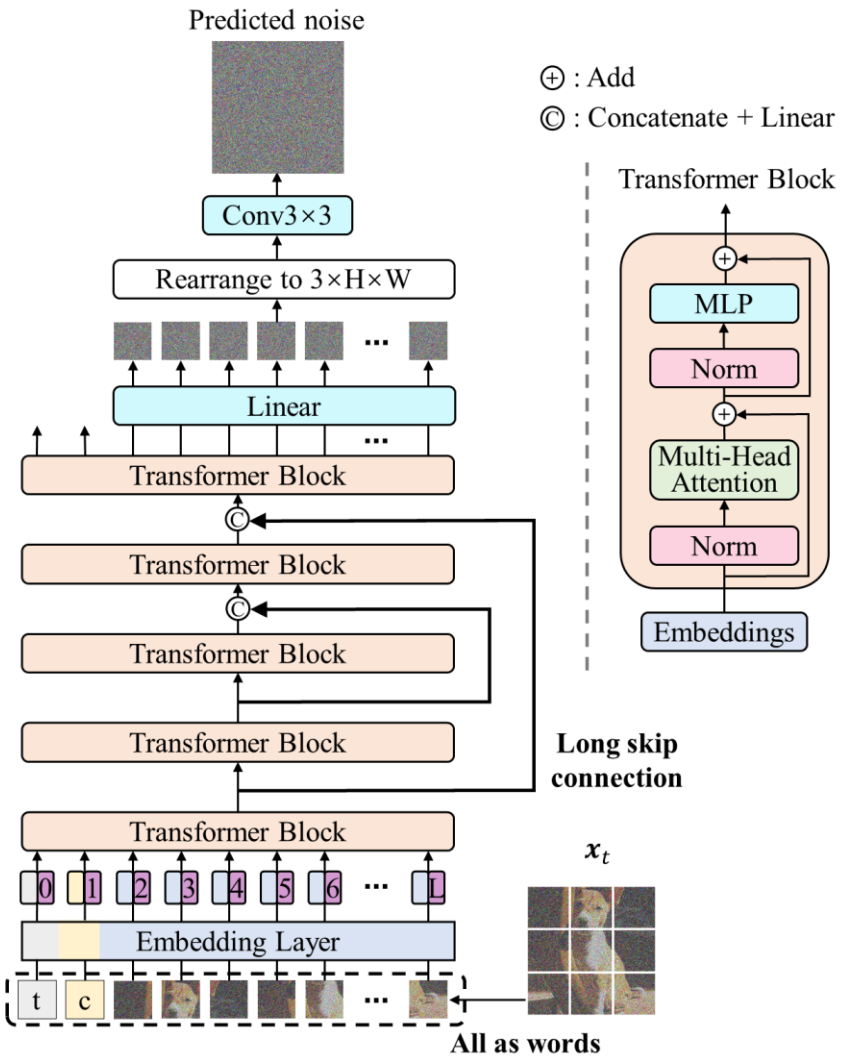
Title: All are Worth Words: A ViT Backbone for Diffusion Models

Author: Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, Jun Zhu

Papertag: THU-PM-197

OverView

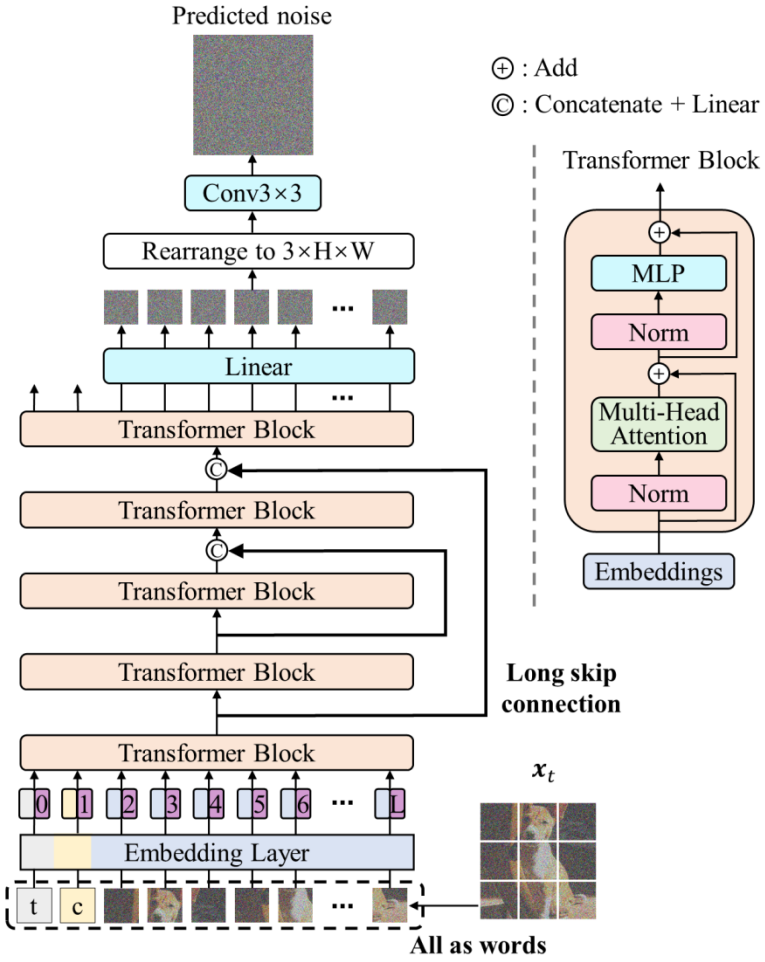


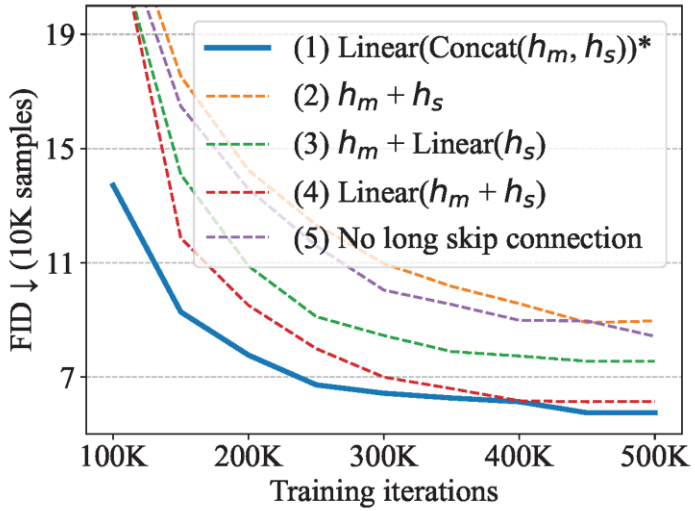
We design a simple and general ViT-based architecture (named U-ViT) for image generation with diffusion models

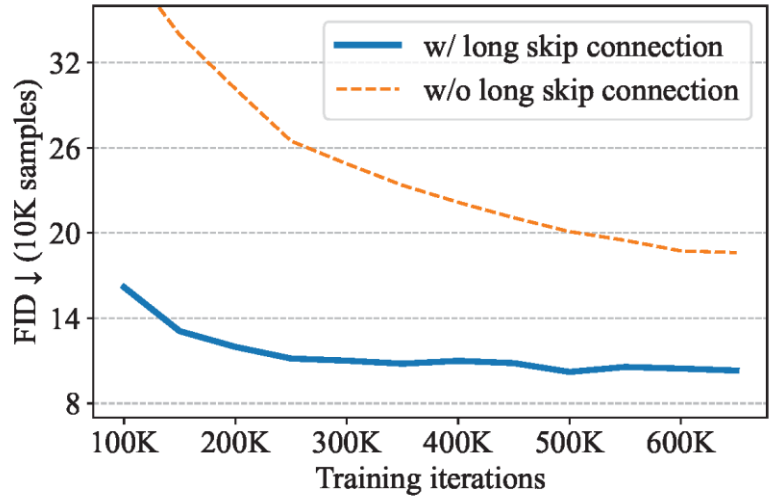
- **Employe long skip connection between shallow and deep layers**
- **Treat all input as tokens**

U-ViT achieve record-breaking FID scores of 2.29 in class-conditional image generation on ImageNet 256×256, and 5.48 in text-to-image generation on MS-COCO

The way to combine the long skip connection.



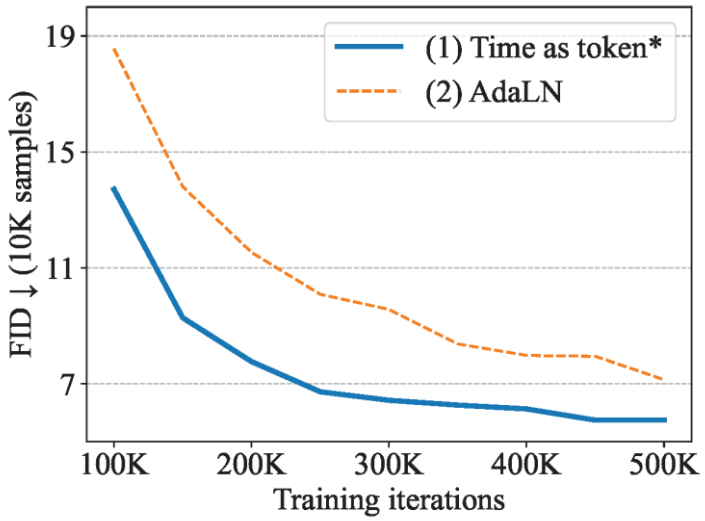


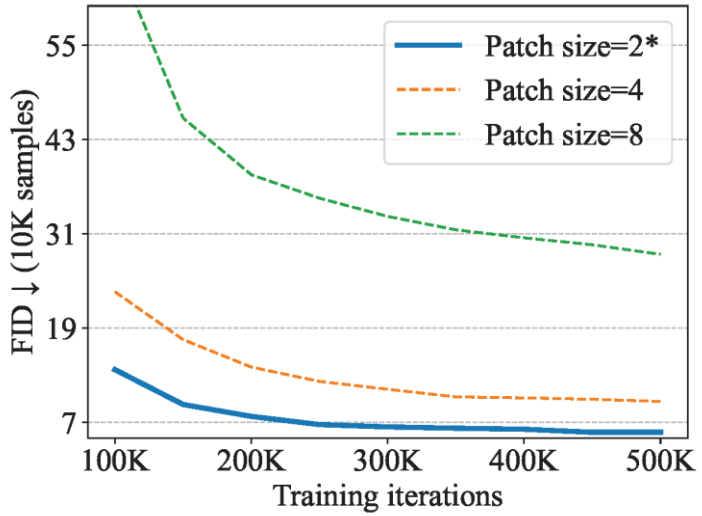


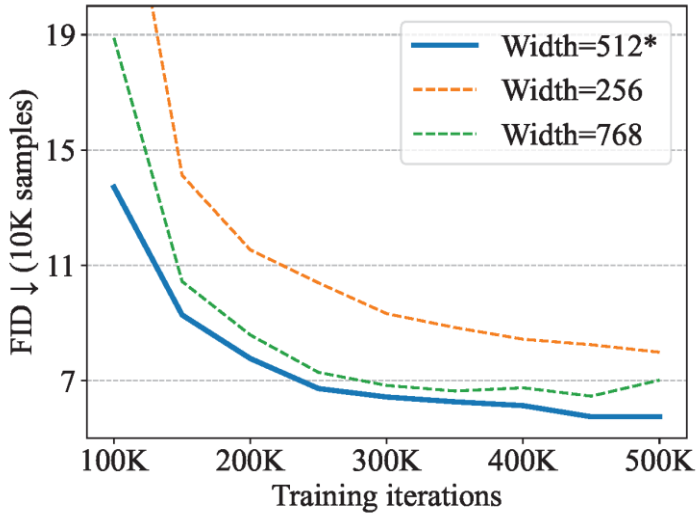
$h_m, h_s \in R^{L,D}$ are the embeddings from the main branch and the long skip branch respectively

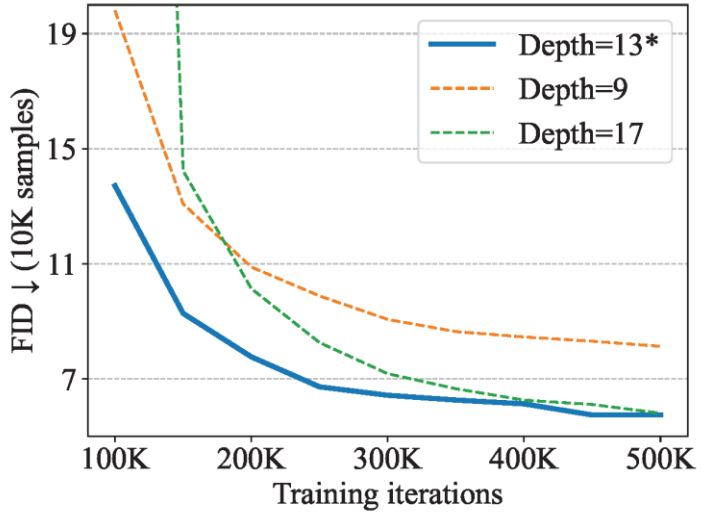
1. $Linear(Concate(h_m, h_s))$
2. $h_m + h_s$
3. $h_m + Linear(h_s)$
4. $Linear(h_m + h_s)$
5. No long skip connection

Other Method









Experiments

- Unconditional generation: CIFAR10, CelebA
- Class Conditional generation: ImageNet at $64 \times 64, 256 \times 256, 512 \times 512$ resolutions
- Text-to-Image generation: MSCOCO at 256×256 resolutions

Model on **CIFAR10**		FID ↓
GAN		
StyleGAN2-ADA [31]		2.92
Diff. based on U-Net	#Params	
DDPM [25]	36M	3.17
IDDPM [49]	53M	2.90
DDPM++ cont. [67]	62M	2.55
EDM† [30]	56M	**1.97**
Diff. based on ViT	#Params	
GenViT [76]	11M	20.20
U-ViT-S/2	44M	3.11

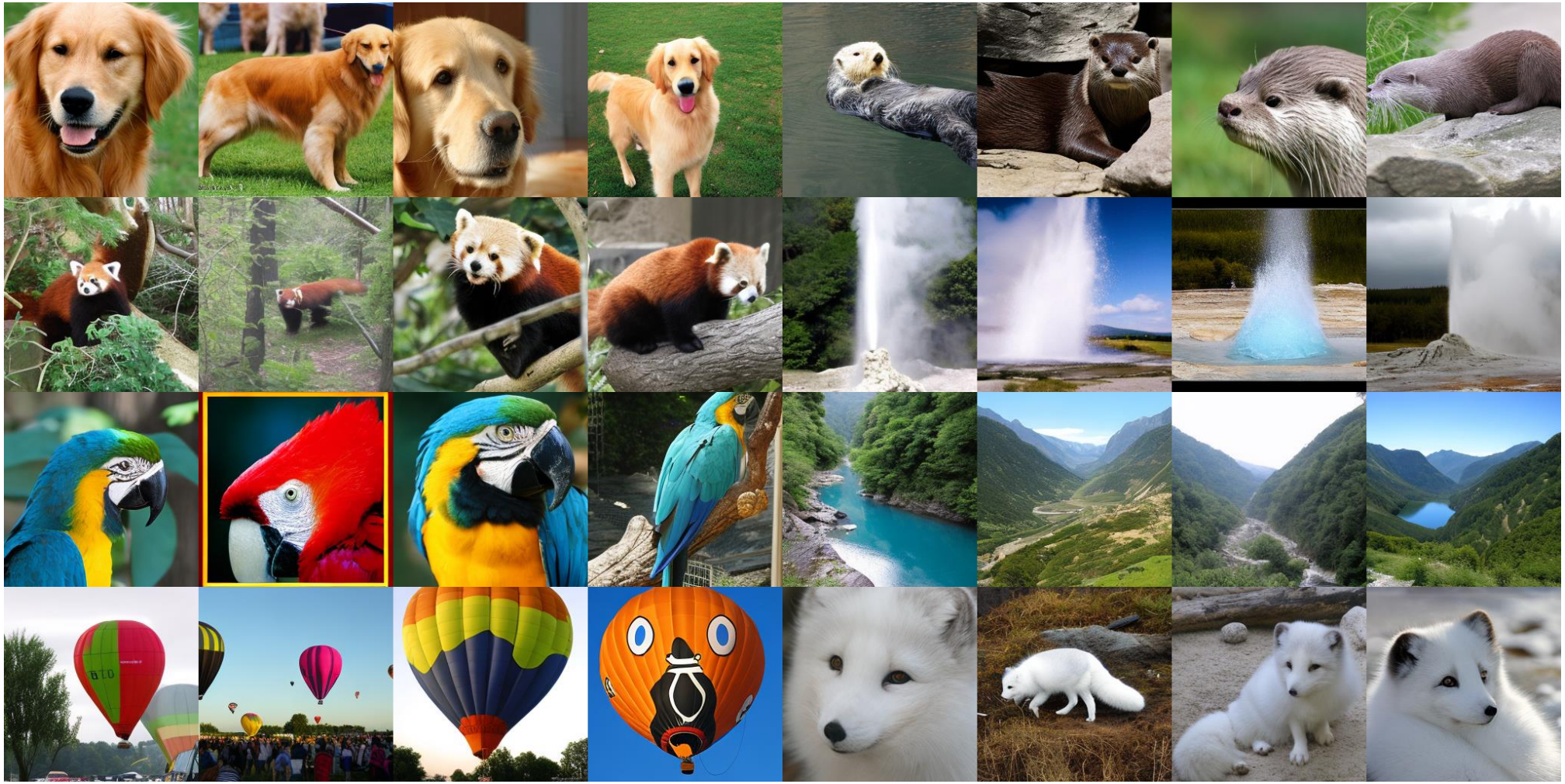
Model on **CelebA 64×64**		FID ↓
Diff. based on U-Net	#Params	
DDIM [63]	79M	3.26
Soft Truncation† [32]	62M	**1.90**
Diff. model based on ViT	#Params	
U-ViT-S/4	44M	2.87

Model on **ImageNet 64×64**		FID ↓
GAN		
BigGAN-deep [5]		4.06
StyleGAN-XL [59]		1.51
Diff. based on U-Net	#Params	
IDDPM (small) [49]	100M	6.92
IDDPM (large) [49]	270M	2.92
CDM [26]	Unknown	1.48
ADM [13]	296M	2.07
EDM† [30]	296M	**1.36**
Diff. based on ViT	#Params	
U-ViT-M/4	131M	5.85
U-ViT-L/4	287M	4.26

Model on **ImageNet 256×256**		FID ↓
GAN		
BigGAN-deep [5]		6.95
StyleGAN-XL [59]		2.30
Discrete diff. based on transformer		
VQ-Diffusion [20]		11.89
VQ-Diffusion (acc0.05) [20]		5.32
Diff. based on U-Net	#Params	
IDDPM [49]	270M + 280M (SR)	12.26
CDM [26]	Unknown	4.88
ADM [13]	554M	10.94
ADM-U [13]	296M + 312M (SR)	7.49
ADM-G [13]	554M + 54M (Cls)	4.59
ADM-G, ADM-U [13]	296M + 65M (Cls) + 312M (SR)	3.94
LDM‡ [54]	400M + 55M (AE)	3.60
Diff. based on ViT	#Params	
U-ViT-H/2‡	501M + 84M (AE)	**2.29**

Model on **ImageNet 512×512**		FID ↓
GAN		
BigGAN-deep [5]		8.43
StyleGAN-XL [59]		2.41
Diff. based on U-Net	#Params	
ADM [13]	559M	23.24
ADM-U [13]	422M + 309M (SR)	9.96
ADM-G [13]	559M + 54M (Cls)	7.72
ADM-G, ADM-U [13]	422M + 43M (Cls) + 309M (SR)	**3.85**
Diff. based on ViT	#Params	
U-ViT-H/4‡	501M + 84M (AE)	4.05

Experiments

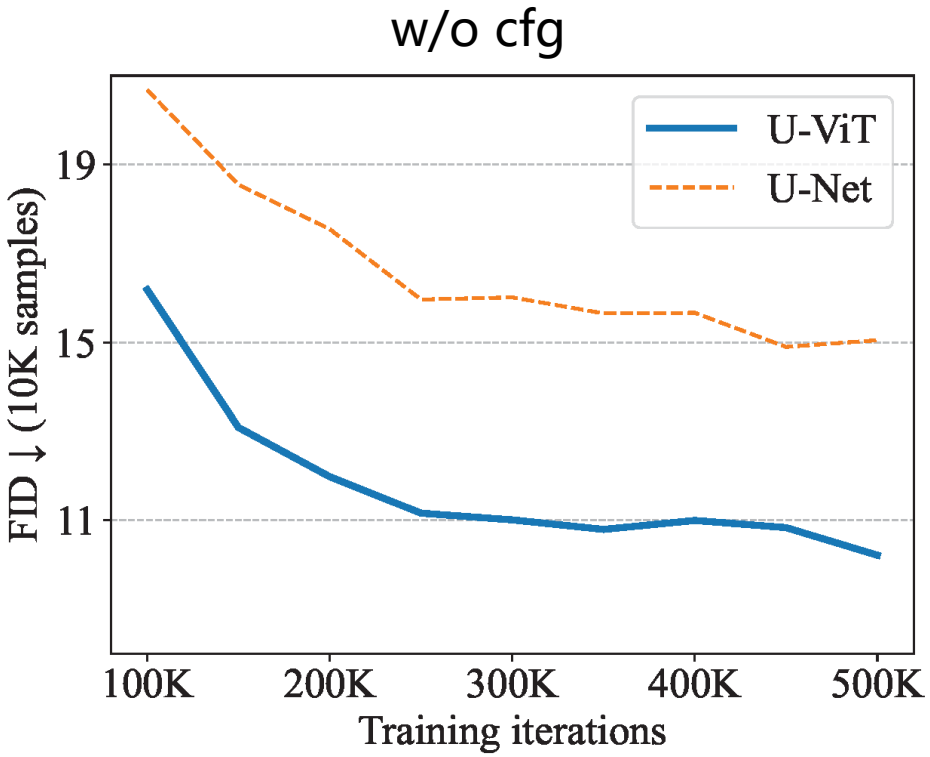


Experiments

Model	FID	Type	Training datasets	#Params
Generative model trained on external large dataset (zero-shot)				
DALL-E [53]	∼ 28	Autoregressive	DALL-E dataset (250M)	12B
CogView [14]	27.1	Autoregressive	Internal dataset (30M)	4B
LAFITE [82]	26.94	GAN	CC3M (3M)	75M + 151M (TE)
GLIDE [48]	12.24	Diffusion	DALL-E dataset (250M)	3.5B + 1.5B (SR)
Make-A-Scene [19]	11.84	Autoregressive	Union datasets (without MS-COCO) (35M)	4B
DALL-E 2 [52]	10.39	Diffusion	DALL-E dataset (250M)	4.5B + 700M (SR)
Imagen [56]	7.27	Diffusion	Internal dataset (460M) + LAION (400M)	2B + 4.6B (TE) + 600M (SR)
Parti [77]	7.23	Autoregressive	LAION (400M) + FIT (400M) + JFT (4B)	20B + 630M (AE)
Re-Imagen [8]	**6.88**	Diffusion	KNN-ImageText (50M)	2.5B + 750M (SR)
Generative model trained on external large dataset with access to MS-COCO				
VQ-Diffusion† [20]	13.86	Discrete diffusion	Conceptual Caption Subset (7M)	370M
Make-A-Scene [19]	7.55	Autoregressive	Union datasets (with MS-COCO) (35M)	4B
Re-Imagen‡ [8]	5.25	Diffusion	KNN-ImageText (50M)	2.5B + 750M (SR)
Parti† [77]	**3.22**	Autoregressive	LAION (400M) + FIT (400M) + JFT (4B)	20B + 630M (AE)
Generative model trained on MS-COCO				
AttnGAN [75]	35.49	GAN	MS-COCO (83K)	230M
DM-GAN [83]	32.64	GAN	MS-COCO (83K)	46M
VQ-Diffusion [20]	19.75	Discrete diffusion	MS-COCO (83K)	370M
DF-GAN [70]	19.32	GAN	MS-COCO (83K)	19M
XMC-GAN [79]	9.33	GAN	MS-COCO (83K)	166M
Friro [18]	8.97	Diffusion	MS-COCO (83K)	512M + 186M (TE) + 68M (AE)
LAFITE [82]	8.12	GAN	MS-COCO (83K)	75M + 151M (TE)
U-Net*	7.32	Latent diffusion	MS-COCO (83K)	53M + 123M (TE) + 84M (AE)
U-ViT-S/2	5.95	Latent diffusion	MS-COCO (83K)	45M + 123M (TE) + 84M (AE)
U-ViT-S/2 (Deep)	**5.48**	Latent diffusion	MS-COCO (83K)	58M + 123M (TE) + 84M (AE)

Experiments

UViT vs UNeT in “Latent Diffusion” setting



U-Net employs implementation from ADM

Same optimizer configuration as ADM.

	Params	GFLOPs	FID (w/ cfg)
UNeT	646M + 84M(AE)	135 + 312(AE)	2.66
UViT-H/2	501M + 84M(AE)	133 + 312(AE)	2.29

Thanks