Scalable Diffusion Models with Transformers

ft. Stable Diffusion

Jun Hyung Lee

Korea University
Department of
Computer Science and Engineering

IE LAB
Interlligence Engineering

References

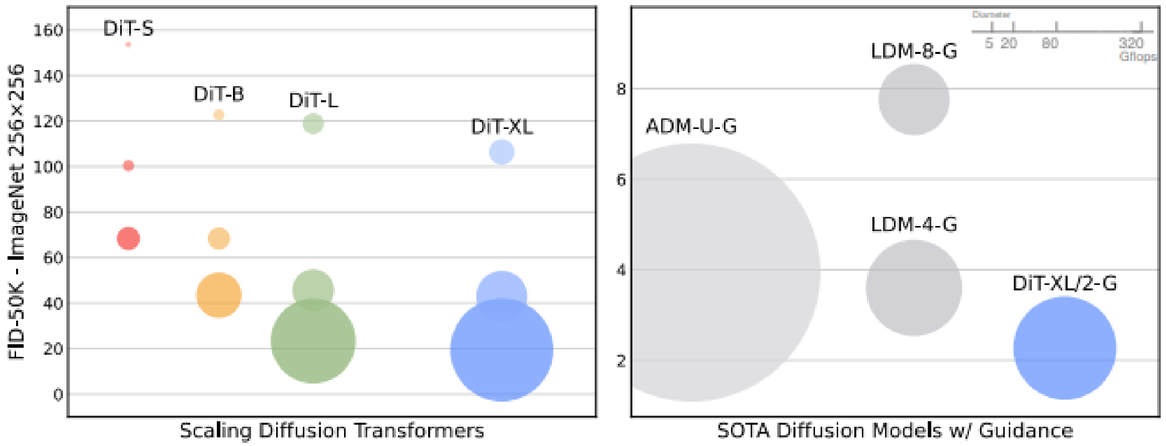
- **Denoising Diffusion Probabilistic Models**, Jonathan Ho (2020)
- **High-Resolution Image Synthesis with Latent Diffusion Models**, Robin Rombach (2022), Alexey Dosovitskiy
- **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,** Alexey Dosovitskiy (2020)
- **Diffusion Models Beat GANs on Image Synthesis,** Prafulla Dhariwall (2021)
- **Improved Denoising Diffusion Probabilistic Models,** Alex Nichol (2021)

Contribution

① U-Net Inductive bias is not crucial -> **replaced backbone with transformer**
② Therefore, it has an advantage of **scalability**, **robustness** and **efficiency**
③ Possible **cross–domain research**
④ Increased transformer depth/width & Input tokens lower FID score -> SOTA **FID of 2.27**

Architecture Complexity

- As network complexity increases sample quality increases
- Bigger bubble -> more flops -> architecture complexity goes up
- What is Flop: floating point operations per second. It's a measure of computer performance
- Flops = cores $\times \frac{\text{cycles}}{\text{second}} \times \frac{\text{Flops}}{\text{cycle}}$



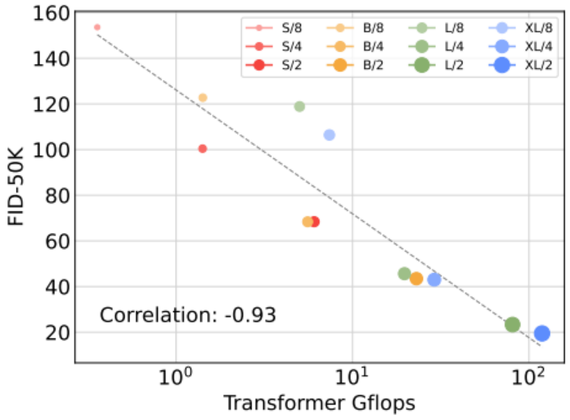
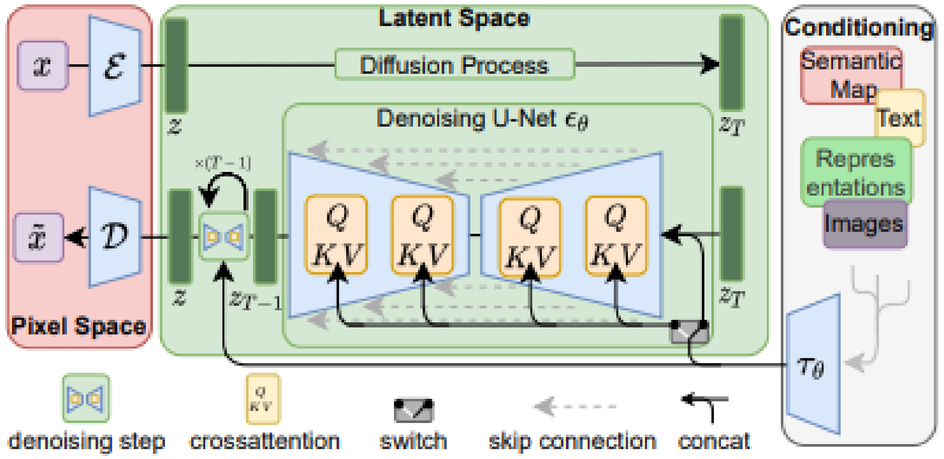


Figure 8. **Transformer Gflops are strongly correlated with FID.** We plot the Gflops of each of our DiT models and each model's FID-50K after 400K training steps.

Latent Diffusion Model (LDM)

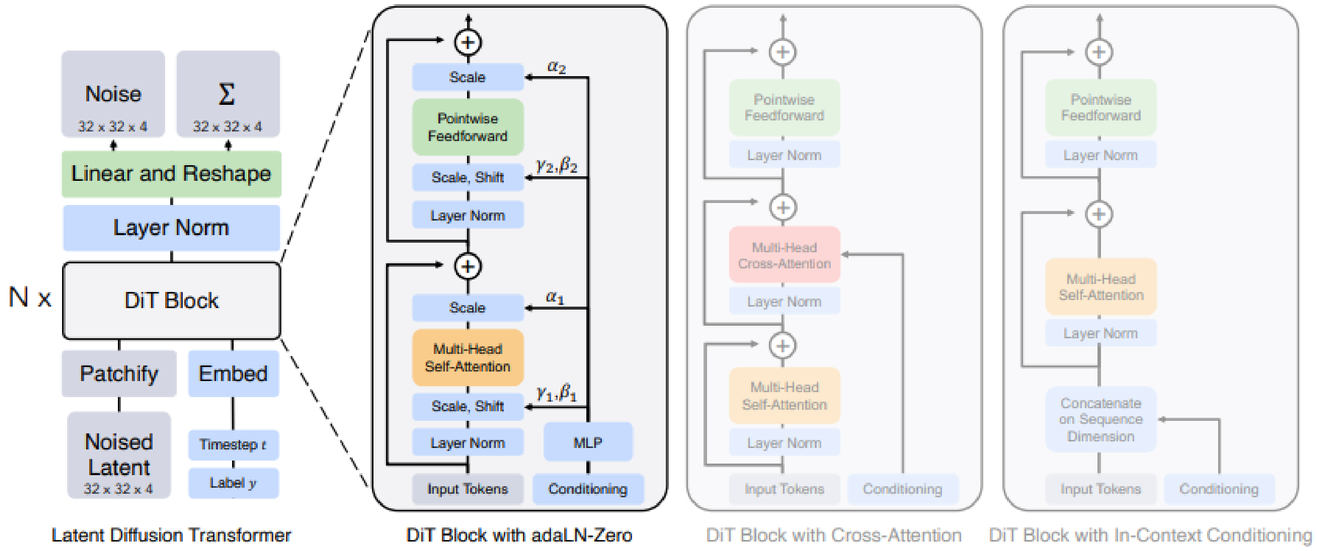
- Previous diffusion models typically operate directly in pixel space causing expensive computation. In this work, to retain the quality and flexibility under the limited computational resources, it applies in the latent space of pretrained autoencoders.
- Inductive bias of Diffusion Models inherited from their U-Net architecture, which makes them particularly effective for data with spatial structure and therefore alleviates the need for aggressive, quality-reducing compression levels as required by previous approaches.



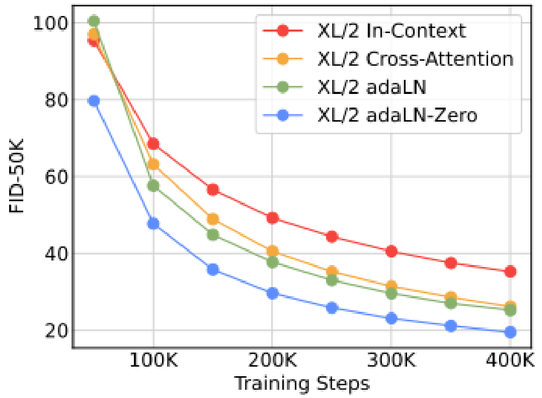
Latent Diffusion Model

DiT Architecture

- Train conditional latent DiT models. The input latent is decomposed into patches and processed by several DiT blocks.

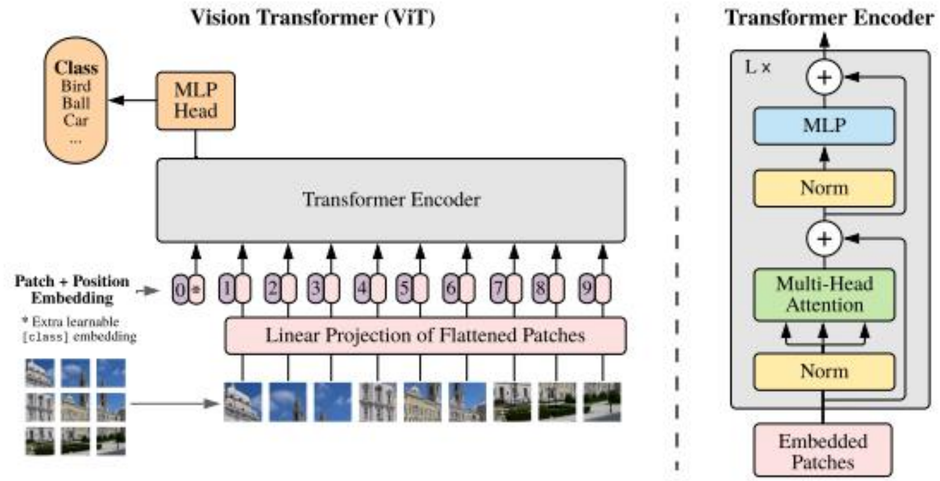


DiT Model

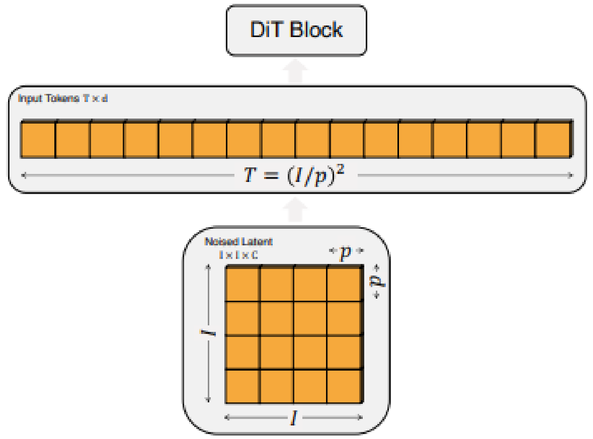


Vision Transformer (ViT)

- Split an image into ($P \times P$) size patches, linearly embed each of them, add position embeddings, and feed the sequence of vectors to a standard Transformer encoder.
- Problem: Transformers lack some of the inductive biases compare to CNN (translation equivariance and locality). Therefore sufficient amounts of data are necessary to generalize well.
- Patch size ($P \times P$), a spatial representation(the noised latent from the VAE) of shape $I \times I \times C$ is "patchified" into a sequence of length $T = (I/P)^2$ with hidden dimension d.
- A smaller patch size P results in a longer sequence length and thus more Gflops.



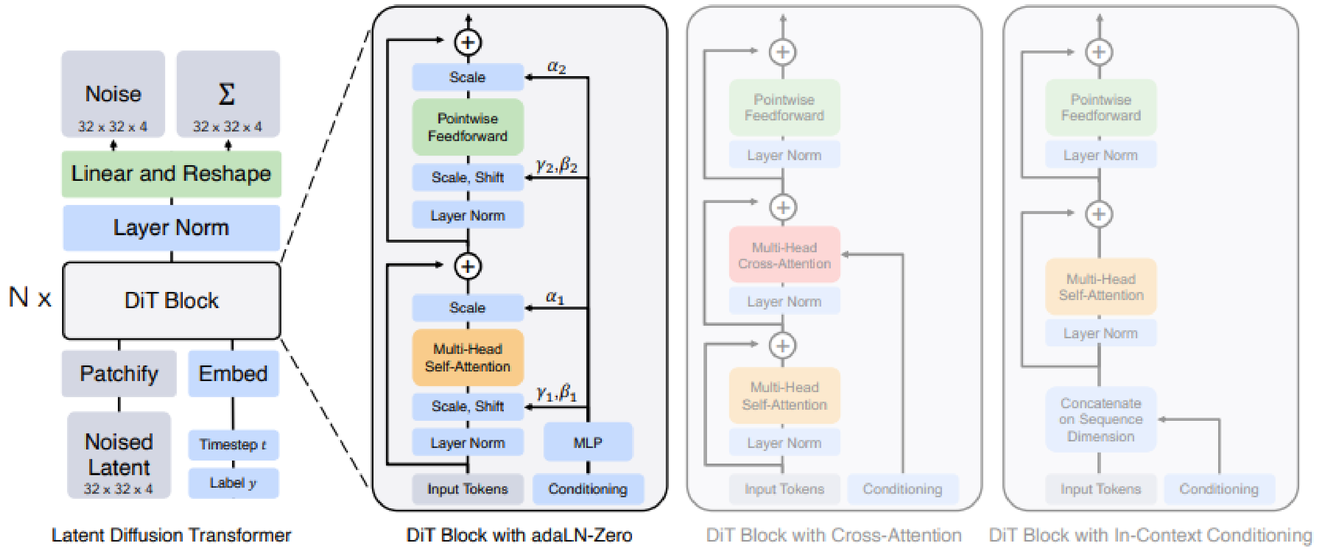
ViT (2020)



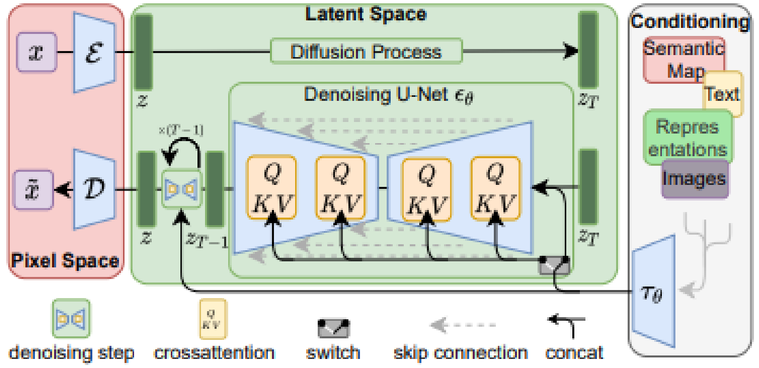
DiT(2022)

DiT Architecture

- Train conditional latent DiT models. The input latent is decomposed into patches and processed by several DiT blocks.



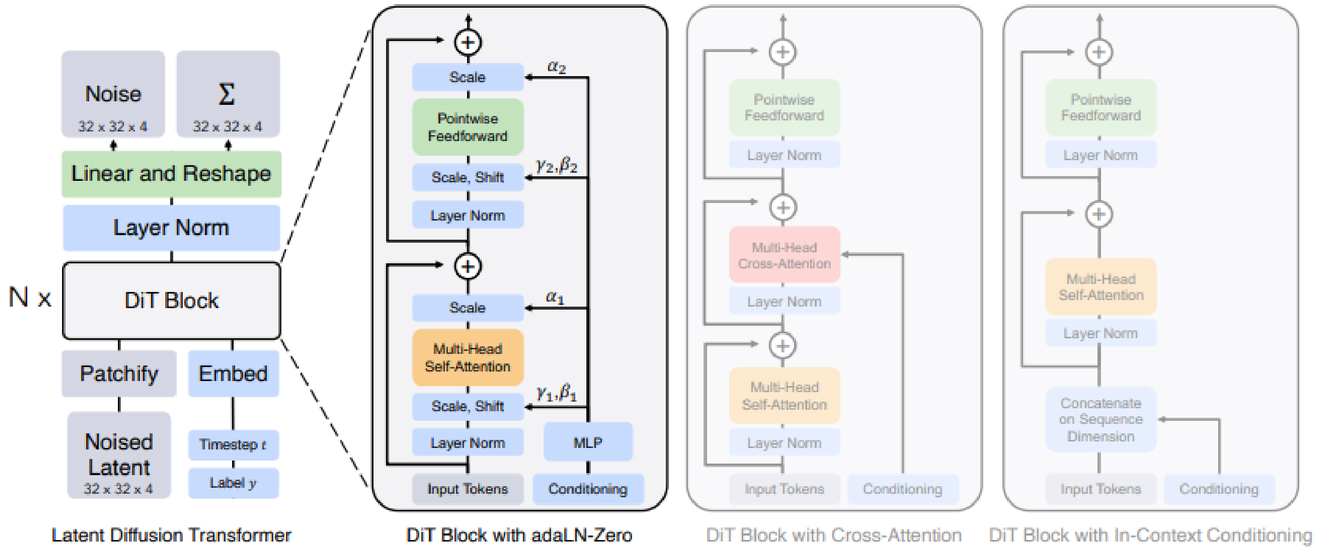
DiT Model



Latent Diffusion Model

Adaptive Layer Norm

- Train conditional latent DiT models. The input latent is decomposed into patches and processed by several DiT blocks.



DiT Model

Operation	FID
AdaGN	13.06
Addition + GroupNorm	15.08

Diffusion Models Beat GANs on Image Synthesis (Dhariwal,2021)

Results

Class-Conditional ImageNet 256×256					
Model	FID↓	sFID↓	IS↑	Precision↑	Recall↑
BigGAN-deep [2]	6.95	7.36	171.4	0.87	0.28
StyleGAN-XL [50]	2.30	4.02	265.12	0.78	0.53
ADM [9]	10.94	6.02	100.98	0.69	0.63
ADM-U	7.49	5.13	127.49	0.72	0.63
ADM-G	4.59	5.25	186.70	0.82	0.52
ADM-G, ADM-U	3.94	6.14	215.84	0.83	0.53
CDM [20]	4.88	-	158.71	-	-
LDM-8 [45]	15.51	-	79.03	0.65	0.63
LDM-8-G	7.76	-	209.52	0.84	0.35
LDM-4	10.56	-	103.49	0.71	0.62
LDM-4-G (cfg=1.25)	3.95	-	178.22	0.81	0.55
LDM-4-G (cfg=1.50)	3.60	-	247.67	**0.87**	0.48
DiT-XL/2	9.62	6.85	121.50	0.67	**0.67**
DiT-XL/2-G (cfg=1.25)	3.22	5.28	201.77	0.76	0.62
DiT-XL/2-G (cfg=1.50)	**2.27**	**4.60**	**278.24**	0.83	0.57

Table 2. **Benchmarking class-conditional image generation on ImageNet 256×256.** DiT-XL/2 achieves state-of-the-art FID.

Class-Conditional ImageNet 512×512					
Model	FID↓	sFID↓	IS↑	Precision↑	Recall↑
BigGAN-deep [2]	8.43	8.13	177.90	0.88	0.29
StyleGAN-XL [50]	2.41	4.06	267.75	0.77	0.52
ADM [9]	23.24	10.19	58.06	0.73	0.60
ADM-U	9.96	5.62	121.78	0.75	**0.64**
ADM-G	7.72	6.57	172.71	**0.87**	0.42
ADM-G, ADM-U	3.85	5.86	221.72	0.84	0.53
DiT-XL/2	12.03	7.12	105.25	0.75	**0.64**
DiT-XL/2-G (cfg=1.25)	4.64	5.77	174.77	0.81	0.57
DiT-XL/2-G (cfg=1.50)	**3.04**	**5.02**	**240.82**	0.84	0.54

Table 3. **Benchmarking class-conditional image generation on ImageNet 512×512.** Note that prior work [9] measures Precision and Recall using 1000 real samples for 512 × 512 resolution; for consistency, we do the same.