Contents

I. Overview

Scalable Diffusion Models with Transformers

- **Problem:** Diffusion models still rely on U-Nets as the standard backbone, but it's unclear whether the U-Net inductive bias is actually necessary.
- **Solution:** Replace U-Net with Vision Transformer (ViT) backbone under latent diffusion
- **Key Findings:**

  - Scaling Law : compute(Gflops) ↑ , FID ↓ (strong negative correlation)
    - B/2 < L/2 < XL/2 : consistent improvement
    - Outperforms U-Net at same compute- Scaling Law
  - Best result: DiT-XL/2 achieves FID 2.27 (SOTA)
    - Smaller patches(=more tokens) → more compute + better quality
    - Simple recipe : AdamW, LR=1e-4, DDPM-1000

[Peebles et al., 2023] Scalable Diffusion Models with Transformers, ICCV

I. Background

Understanding DiT's Scaling

Model	Gflops	FID-50K
DiT – S/2	6.06	68.40
DiT – B/2	23.01	43.47
DiT – L/2	80.71	23.33
DiT – XL/2	118.64	19.47

Data: 400K training iterations, no guidance

DiT – XL/2-G	118.64	**2.27(SOTA)**

Classifier-free guidance, cfg=1.50, 7M steps

- Key: Transformer Gflops are strongly correlated with FID-50K (400K / no guidance)
- Two scaling paths:
  ① Model size: S → B → L → XL (depth/width ↑)
  ② Patch size: Smaller patches → More tokens → Higher Gflops

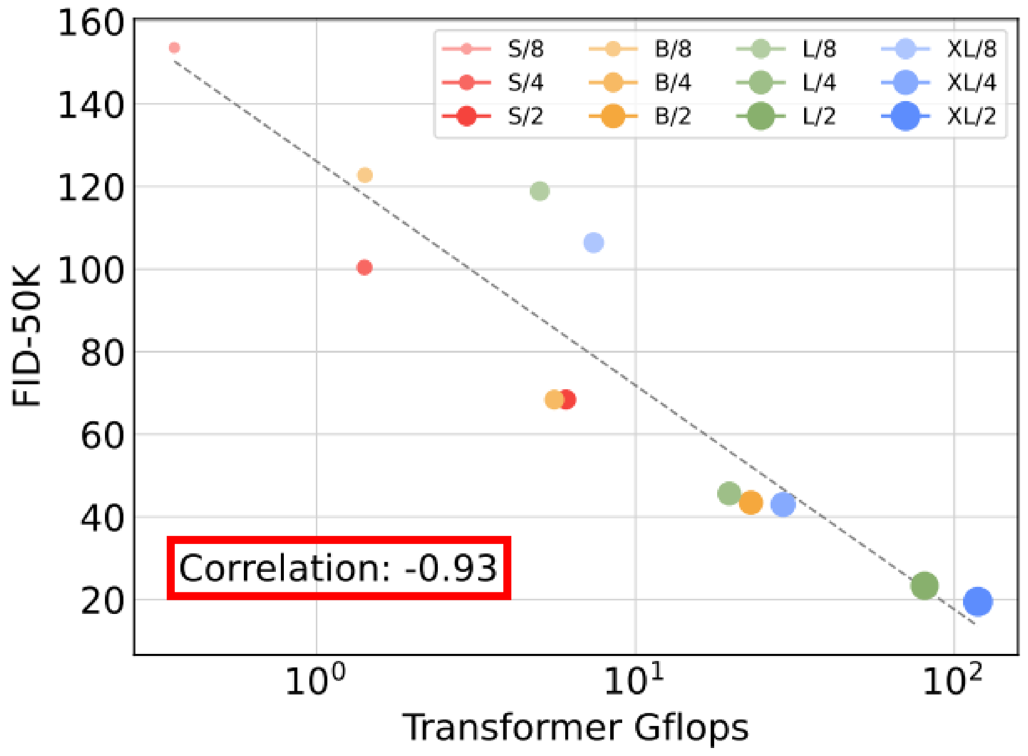
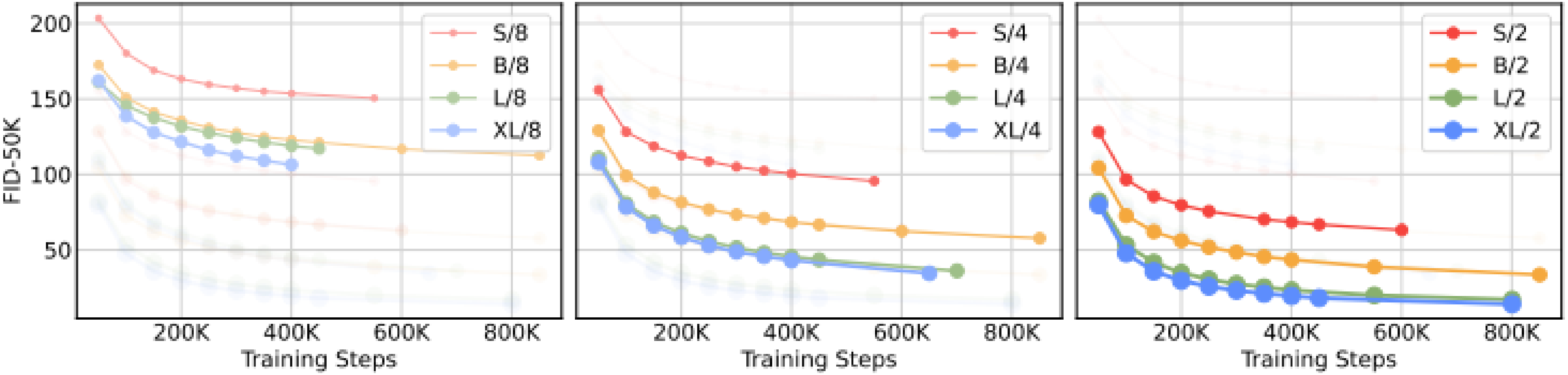


Figure 8. **Transformer Gflops are strongly correlated with FID.** We plot the Gflops of each of our DiT models and each model's FID-50K after 400K training steps.

[Peebles et al., 2023] Scalable Diffusion Models with Transformers, ICCV

II. Claims to Reproduce

Claim 1: (Model-size scaling, patch fixed) With **patch size held constant**, increasing model size (S/B/L/XL)—and thus FLOPs—consistently improves FID.



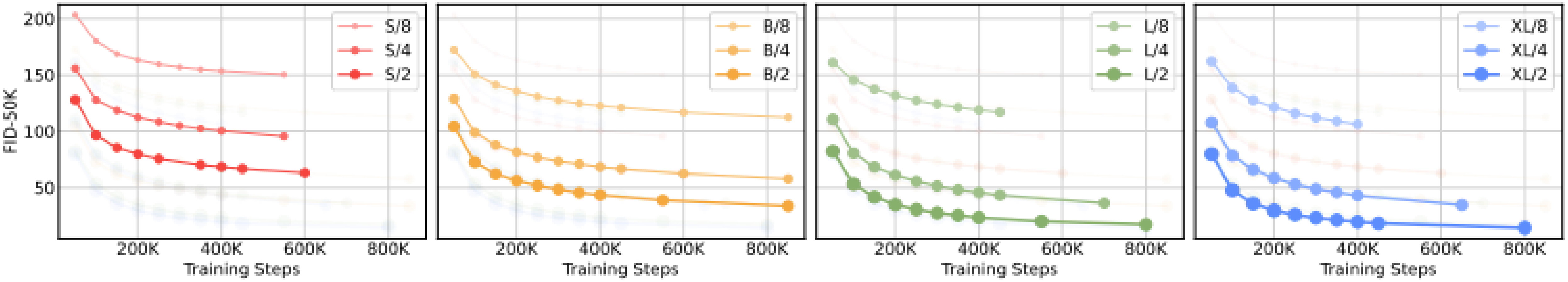
B/2, L/2, XL/2를 동일 레시피로 학습 → 같은 조건으로 샘플링 → FID-50K 계산 (S/2는 제외)

➔ **400K training steps, no guidance**로 scaling을 봄
➔ **ID-50K는 250 DDPM sampling steps**로 계산(ADM TF eval suite 사용)

[Peebles et al., 2023] Scalable Diffusion Models with Transformers, ICCV

II. Claims to Reproduce

Claim 2: (Patch/token scaling, model fixed) With **model size held constant**, decreasing patch size (more tokens → higher FLOPs) consistently improves FID



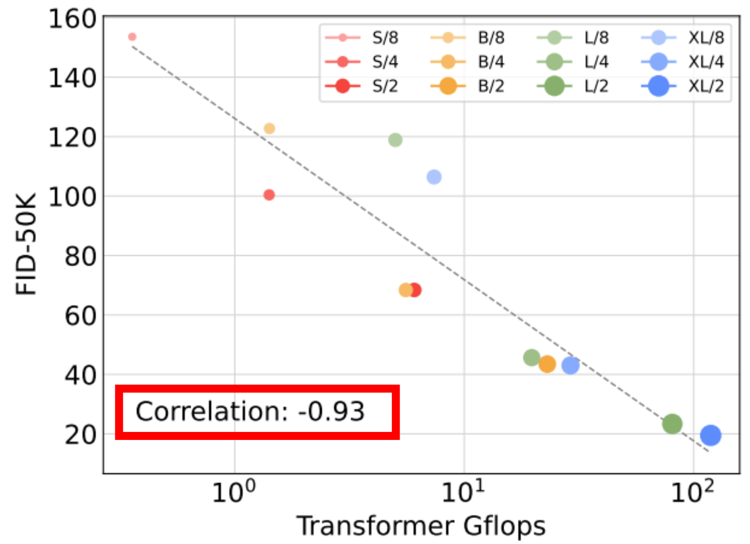
Smaller patch (=more tokens) → higher compute → better FID

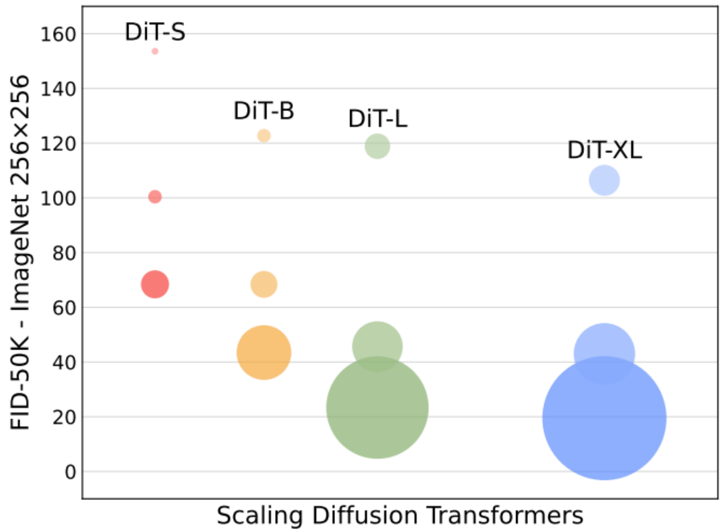
XL/2가 XL/4보다 더 많은 Gflops(=더 작은 patch) → FID가 더 좋아야 claim2 지지.
(Table 4에서도 XL/2가 Gflops 더 큼)

[Peebles et al., 2023] Scalable Diffusion Models with Transformers, ICCV

II. Claims to Reproduce

Claim 3: (Compute-aware scaling) Sample quality cannot be explained by parameter count alone; compute—measured in Gflops—is the key driver of performance improvements (lower FID)





patch만 줄이면 params는 거의 안 변하는데 Gflops만 증가하고, 그때 **FID가 좋아진다 → params가 유일한 설명변수가 아니다**XL/2 vs XL/4: params는 거의 같은데(Table 4에서 둘 다 ~675M), Gflops와 FID가 크게 다름 → “params만으로 설명 불가”에 강함
그리고 FID–Gflops가 더 잘 정렬된다는 걸 표/산점도로 제시하면 됨(논문 Figure 8 방식).

[Peebles et al., 2023] Scalable Diffusion Models with Transformers, ICCV

III. GPU specification

- **Hardware**
  - GPUs: 8 × A100 (80GB)
  - Interconnect: NVLink
  - CPU: 2 × AMD (128 cores / 256 threads)
  - RAM: 1.0 TiB (Swap 8 GiB)

Item	Value
Dataset	ImageNet-1K (≈1.33M images), 256×256
Global batch size	256
Steps / epoch	≈ 1,331,167 / 256 ≈ **5.2K**
Paper-like scaling regime	**~400K steps ≈ ~80 epochs**
Our initial run	**10 epochs ≈ ~52K steps**

III. Experiments

	Paper	Ours
Dataset / Resolution	ImageNet-1K, 256×256, class-conditional	Same
Latent space	Stable Diffusion VAE latent (32×32×4)	Same
Latent scaling	Official code: encode ×0.18215; decode ÷0.18215	Same
Diffusion	DDPM training horizon: **T=1000**	Same
Optimizer / LR	AdamW, constant LR = 1e-4, weight decay = 0	Same
Augmentation	Random horizontal flip only	Same
EMA	EMA enabled ; results typically reported using EMA weights	Same
Training length	Scaling plots reported at ~400K training steps	**10 epochs ≈ 52K steps**
Sampling steps	Commonly 250 steps for sampling/evaluation	Same
Classifier-Free Guidance	Scaling comparisons typically no-guidance; SOTA uses CFG	**CFG scale = 4.0**
FID evaluation	FID-50K (requires 50K generated samples)	*

- **Claim 1** (Model scaling @ fixed patch): B/2 → L/2 → XL/2
- **Claim 2** (Patch scaling @ fixed model): XL/4 vs XL/2
- **Claim 3** (Compute matters): compare params vs GFLOPs across {B/2, L/2, XL/4, XL/2}

IV. Results & Challenges

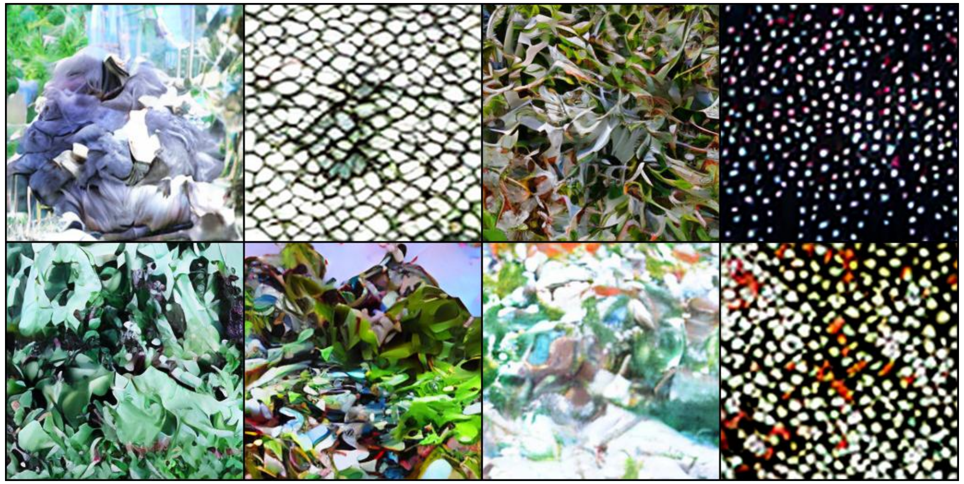
- **Goal:** Reproduce DiT scaling trends via pretrained {B/2, L/2, XL/2, XL/4}
- **Observation:** Samples remained largely noise-like across all variants.
- **Training budget:** 10 epochs ≈ 52K steps (ImageNet-1K, global batch 256).
- **Paper regime:** Scaling plots are reported at ~400K training steps,

so our models were under-trained.

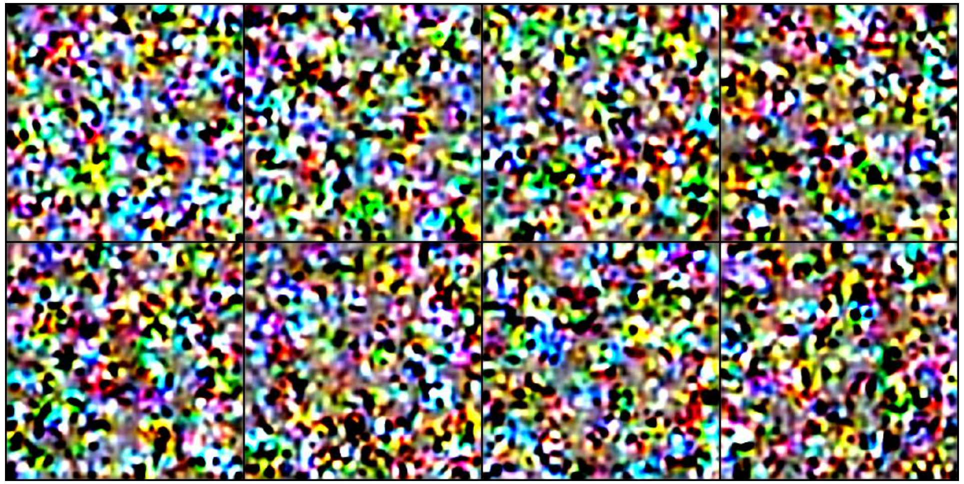
[Peebles et al., 2023] Scalable Diffusion Models with Transformers, ICCV

IV. Results

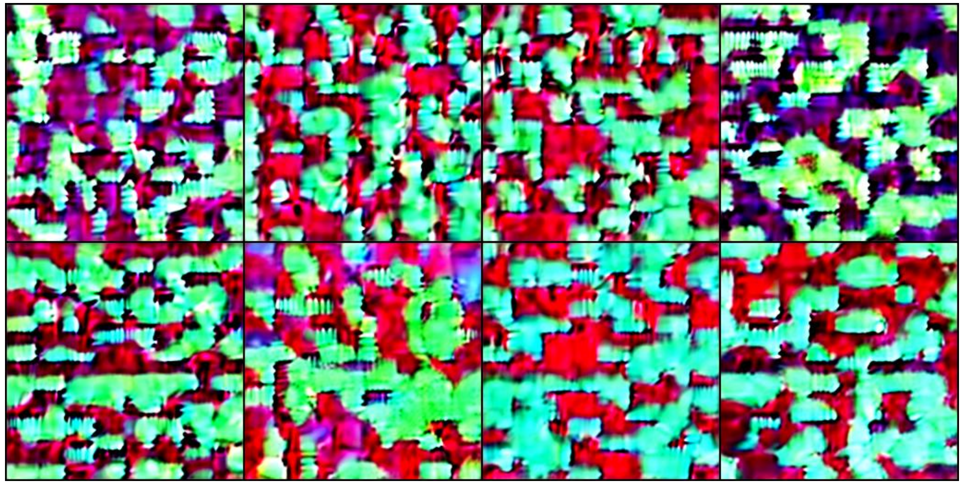
DiT-B/2 Sampling



DiT-L/2 Sampling



DiT-XL/2 Sampling



DiT-XL/4 Sampling

V. Conclusion : Why failed?

Root Cause: Severe Training Budget mismatch

- Paper regime: scaling comparisons reported at 400K training steps (batch=256)
- Paper SOTA model: DiT-XL/2 (256×256) trained for 7M steps
- **Our run**: ~52K steps (10 epochs) → far below the paper's scaling regime Evidence
- All models produced noise-like samples → FID not meaningful at this stage

What we learned

1. A minimum training budget is required

- With ~52K steps, none of the models reached a quality level where size/patch effects are observable.
- Meaningful scaling comparisons require hundreds of thousands of steps. **(paper reports at 400K)**

2. Protocol alignment matters

- For valid comparison, keep training steps, EMA, and sampling protocol (steps/CFG) consistent with the paper.