HKU MMLAB

# LiT: Delving into a Simple Linear Diffusion Transformer for Image Generation (ICCV 2025)

**Jiahao Wang[1], Ning Kang[3], Lewei Yao[3], Mengzhao Chen[1], Chengyue Wu[1], Songyang Zhang[2], Shuchen Xue[4], Yong Liu[5], Taiqiang Wu[1], Xihui Liu[1], Kaipeng Zhang[2], Shifeng Zhang[3], Wenqi Shao[2], Zhenguo Li[3], Ping Luo[1]**

HKU[1]
Shanghai AI Lab[2]
Huawei Noah's Ark Lab[3]
UCAS[4]
THUsz[5]

# Outline

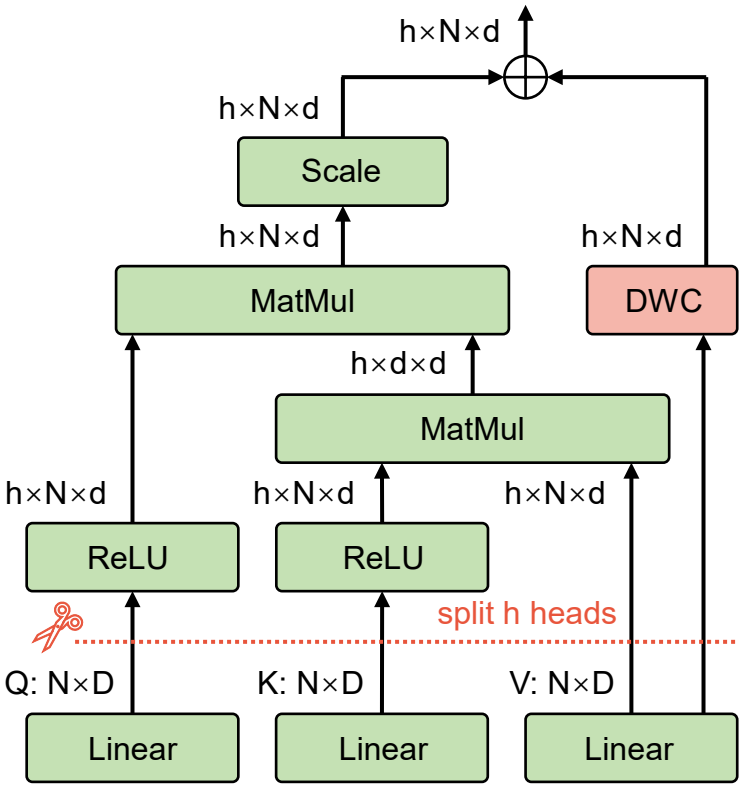
- Overview
- Background
- Motivation
- Exploration Roadmap
- Evaluation
- Conclusion

# Outline

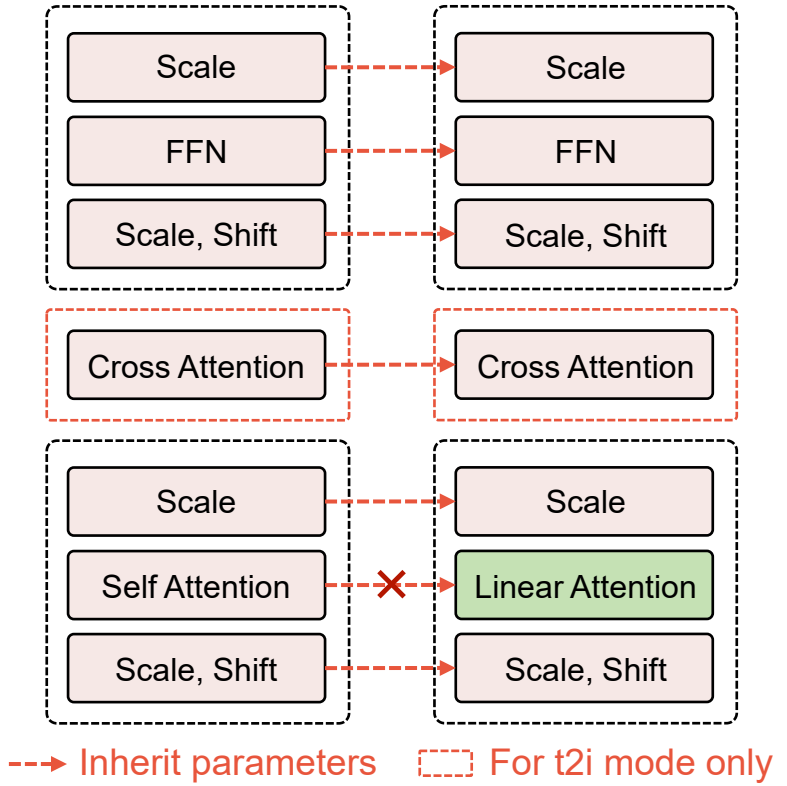
- Overview
- Background
- Motivation
- Exploration Roadmap
- Evaluation
- Conclusion

# Overview

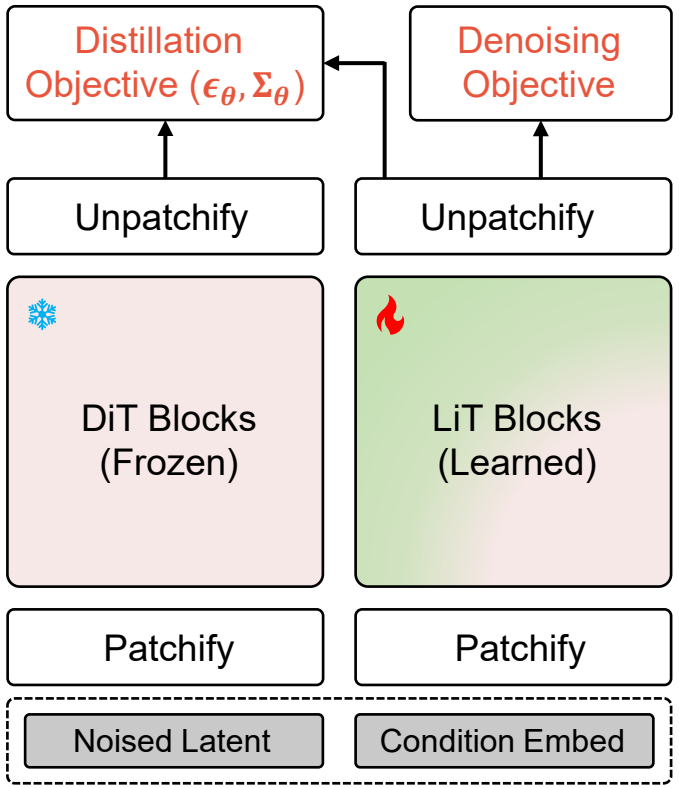
Guideline 1&2: Linear DiT with Few Heads




Guideline 3&4: Inheriting Weight w/o SA



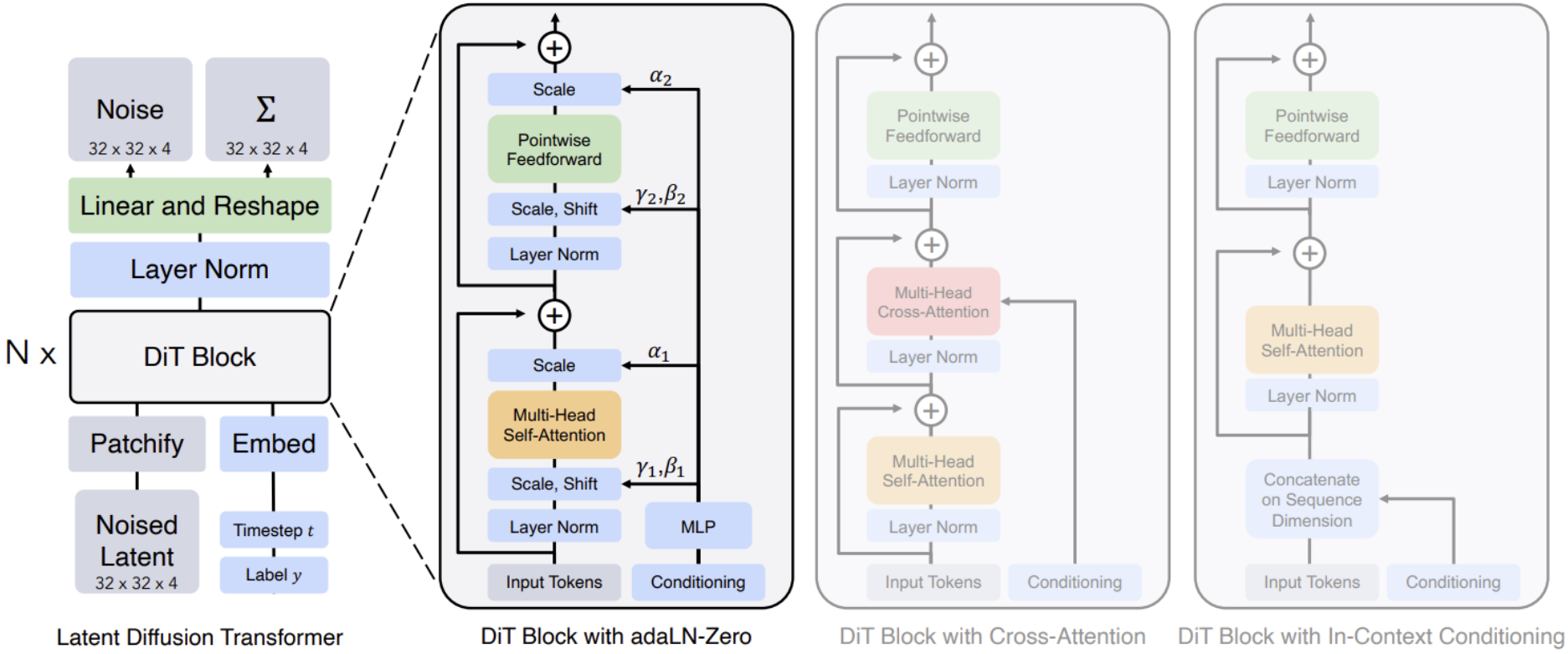

Guideline 5: Distilling Noise and Variance



# Outline

- Overview
- Background
- Motivation
- Exploration Roadmap
- Evaluation
- Conclusion

# Diffusion Transformer (DiT)

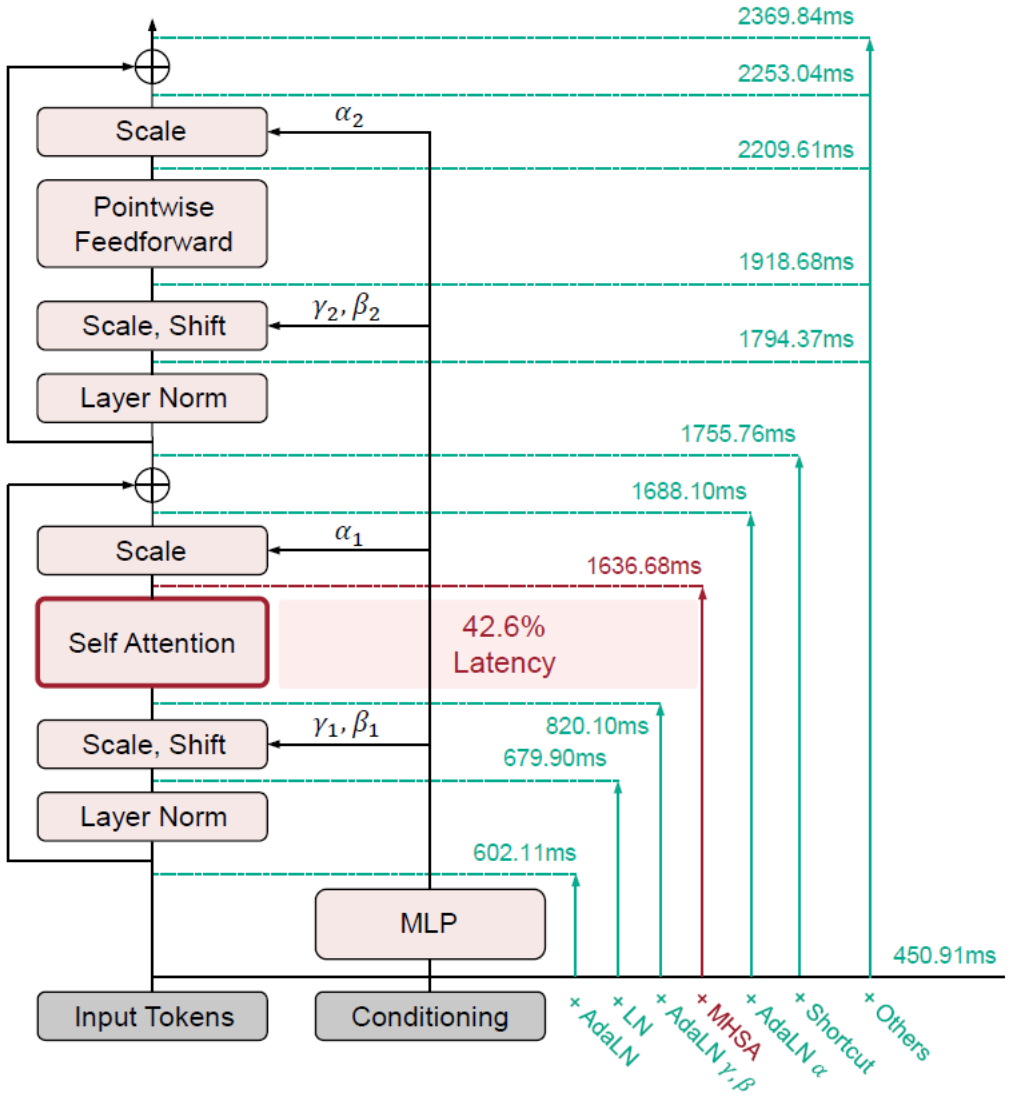



- Replace U-Net backbone with pure transformer for latent diffusion model

Scalable Diffusion Models with Transformers

# Self-attention is Slow in DiT

- DiT-B/4 with a batch size of 8 using NVIDIA A100 GPU
- **42.6%** Latency in a DiT block



# Linear Attention is Conceptually Simple

$$O_i = \sum_{j=1}^{N} \frac{\mathrm{Sim}(Q_i, K_j)}{\sum_{j=1}^{N} \mathrm{Sim}(Q_i, K_j)} V_j$$

$$\mathrm{Sim}(Q, K) = \exp(QK^T/\sqrt{d})$$

(a) Self-attention: $\mathcal{O}(N^2dh)$

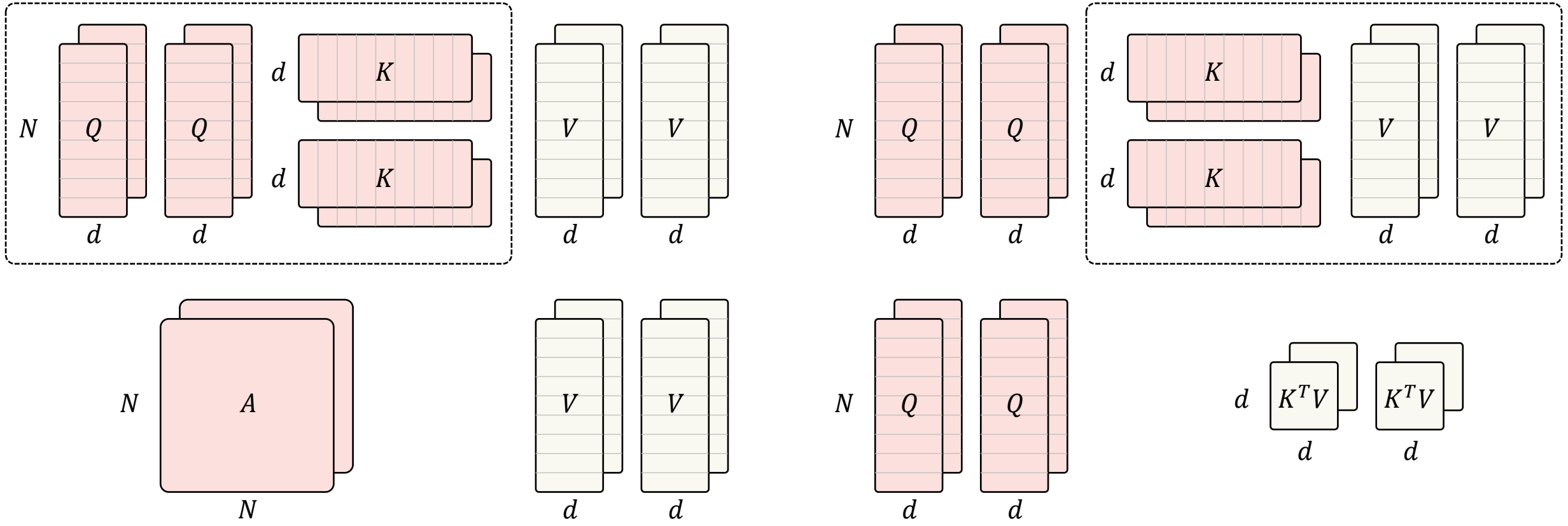
$$O_i = \sum_{j=1}^{N} \frac{\phi(Q_i)\phi(K_j)^T}{\sum_{j=1}^{N} \phi(Q_i)\phi(K_j)^T} V_j = \frac{\phi(Q_i)\left(\sum_{j=1}^{N} \phi(K_j)^T V_j\right)}{\phi(Q_i)\left(\sum_{j=1}^{N} \phi(K_j)^T\right)}$$

$$\mathrm{Sim}(Q, K) = \phi(Q)\phi(K)^T$$

(b) Linear attention: $\mathcal{O}(Nd^2h)$

FLatten Transformer: Vision Transformer using Focused Linear Attention

# Linear Attention is Conceptually Simple




(a) Self-attention: $\mathcal{O}(N^2dh)$

(b) Linear attention: $\mathcal{O}(Nd^2h)$

FLatten Transformer: Vision Transformer using Focused Linear Attention

# Outline

- Overview
- Background
- Motivation
- Exploration Roadmap
- Evaluation
- Conclusion

# How to Convert Pre-trained DiTs into Linear DiTs?

## Architectural Design

- Add convolution? (feature diversity in linear attention)
- Linear attention: how many heads?

## Training Strategy

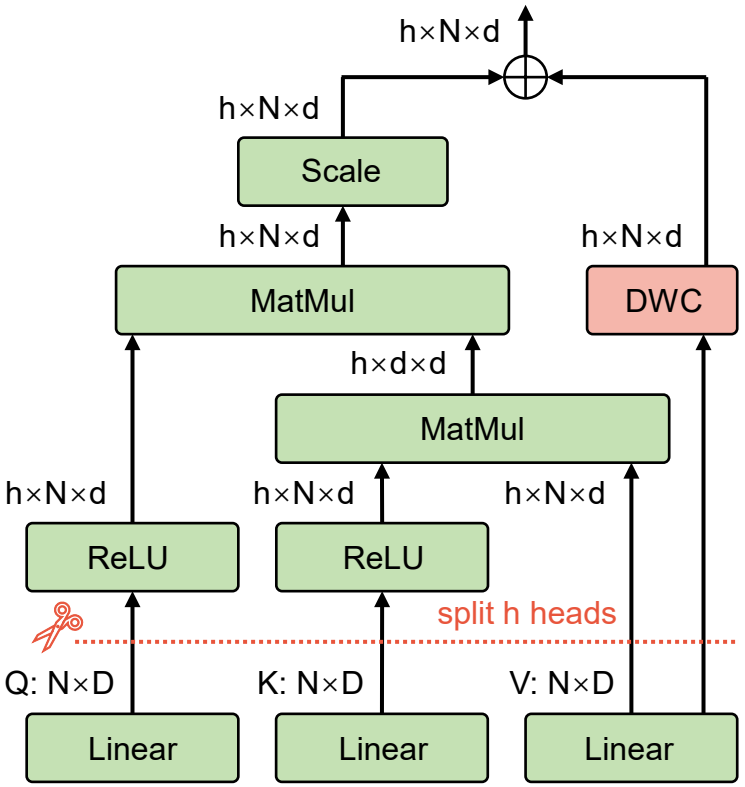
- Should linear DiT be initialized from a converged DiT?
- If so, which parameters should we inherit?
- How to apply knowledge distillation in DiT?

# Outline

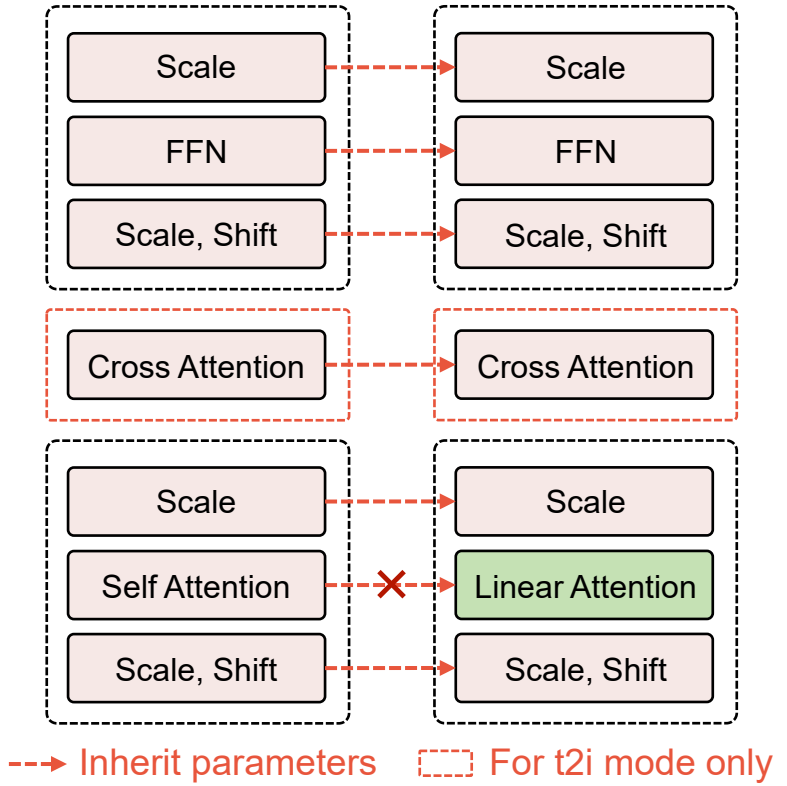
- Overview
- Background
- Motivation
- Exploration Roadmap
- Evaluation
- Conclusion

# Exploration Roadmap

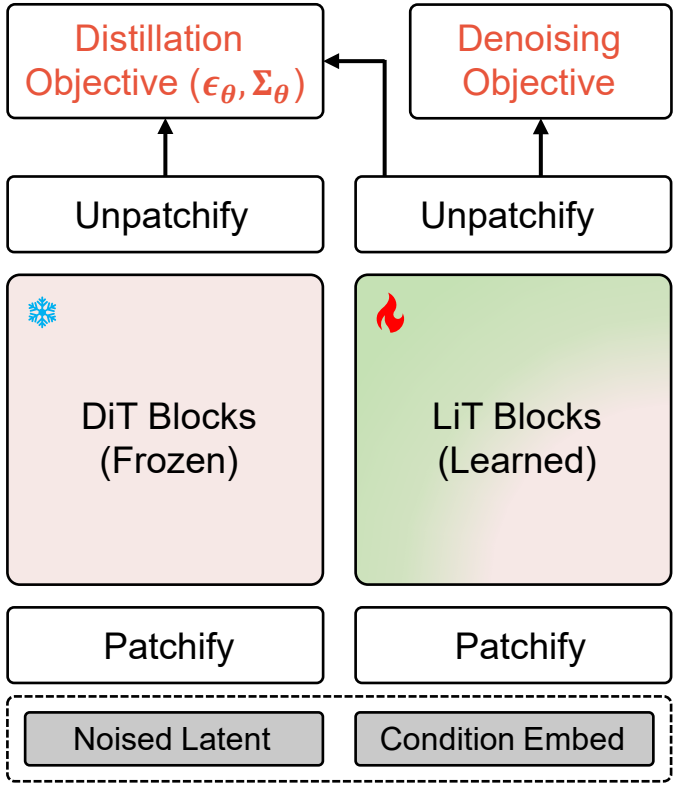
Guideline 1&2: Linear DiT with Few Heads




Guideline 3&4: Inheriting Weight w/o SA



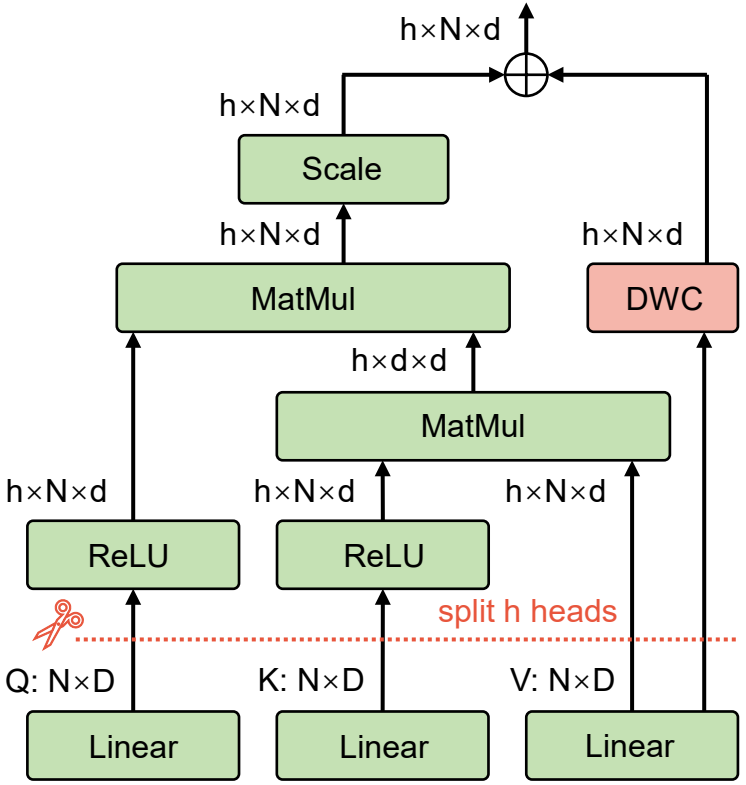

Guideline 5: Distilling Noise and Variance



# Simply Adding a Depth-wise Convolution

**Guideline 1:**

Simply adding a 5×5 depth-wise convolution in linear attention is sufficient for DiT-based image generation.

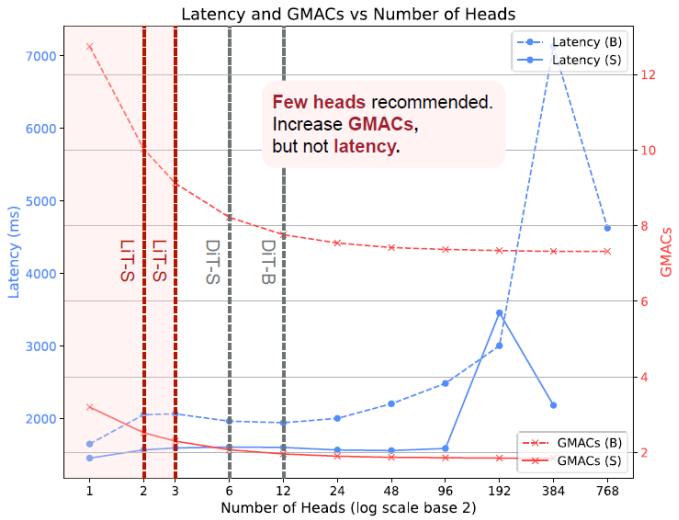


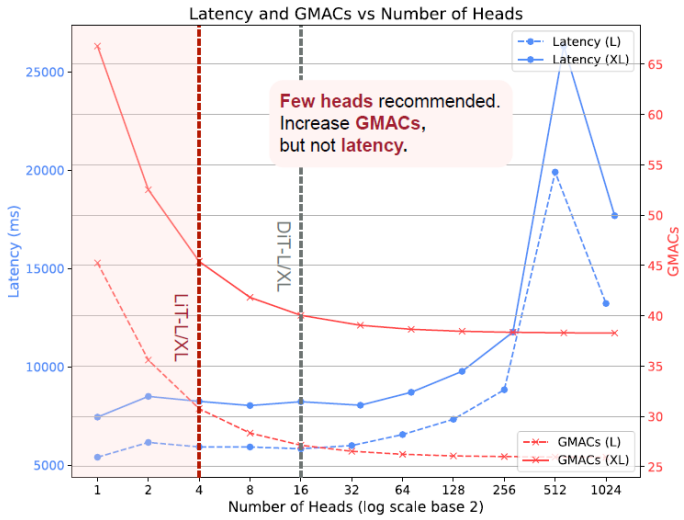

| DiT | Attention | FID-50K (↓) | IS (↑) |
|---|---|---|---|
| S/2 | Softmax | 68.40 | - |
| S/2 | ReLU Linear Baseline | 88.46 | 15.11 |
| S/2 | + Depth-wise Conv. (ReLU) | 63.66 | 22.16 |
| S/2 | + Focused Linear (ReLU) | **63.05** | **22.49** |
| S/2 | + Focused Linear (GELU) | 70.83 | 19.41 |
| B/2 | softmax | 43.47 | - |
| B/2 | ReLU Linear Baseline | 56.92 | 25.80 |
| B/2 | + Depth-wise Conv. (ReLU) | 42.11 | 34.60 |
| B/2 | + Focused Linear (ReLU) | **40.58** | **35.98** |
| B/2 | + Focused Linear (GELU) | 58.86 | 24.23 |

# “Free Lunch” in Linear Attention

**Guideline 2:**

Using few heads in the linear attention increases computation but not latency.



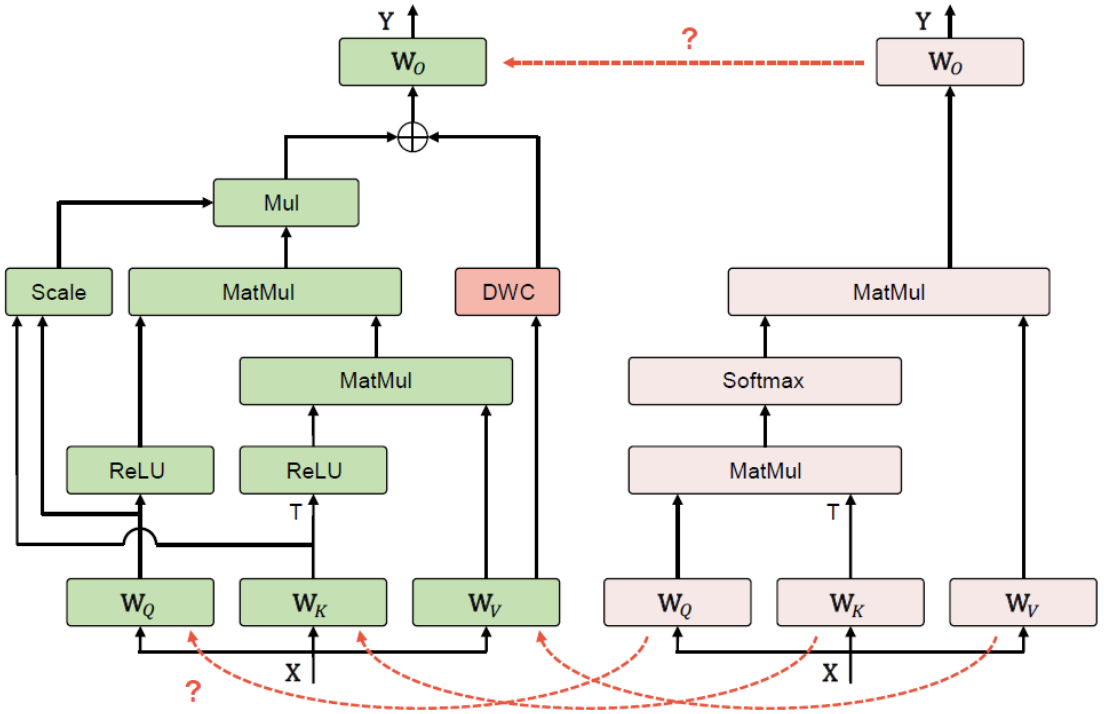

| DiT | Head | FID-50K (↓) | IS (↑) | Prec. (↑) | Rec. (↑) |
|---|---|---|---|---|---|
| S/2 | 1 | 64.42 | 21.54 | 0.380 | 0.574 |
| S/2 | 2 | 63.24 | 22.07 | 0.385 | 0.570 |
| S/2 | 3 | **63.21** | **22.08** | **0.386** | **0.583** |
| S/2 | 6 | 63.66 | 22.16 | 0.383 | 0.580 |
| S/2 | 48 | 78.76 | 17.46 | 0.322 | 0.482 |
| S/2 | 96 | 116.00 | 11.49 | 0.224 | 0.261 |
| B/2 | 1 | 41.77 | 34.78 | 0.487 | 0.631 |
| B/2 | 2 | 41.39 | 35.59 | 0.494 | 0.631 |
| B/2 | 3 | **40.86** | **35.79** | **0.497** | **0.629** |
| B/2 | 12 | 42.11 | 34.60 | 0.484 | 0.631 |
| B/2 | 96 | 68.30 | 20.45 | 0.375 | 0.531 |
| B/2 | 192 | 112.39 | 12.07 | 0.240 | 0.282 |
| L/2 | 1 | 24.46 | 57.36 | 0.600 | 0.637 |
| L/2 | 2 | 24.37 | 57.02 | 0.599 | 0.622 |
| L/2 | 4 | **24.04** | **59.02** | **0.597** | **0.636** |
| L/2 | 16 | 25.25 | 54.67 | 0.587 | 0.632 |
| XL/2 | 1 | 21.13 | 65.06 | 0.619 | 0.632 |
| XL/2 | 2 | **20.66** | **65.39** | **0.624** | **0.636** |
| XL/2 | 4 | 20.82 | 65.52 | 0.619 | 0.632 |
| XL/2 | 16 | 21.69 | 63.06 | 0.617 | 0.628 |

# Weight Inheritance

**Guideline 3:**
Linear diffusion Transformer should be initialized from a converged DiT.

| Load | Iter. | FFN | Modu. | Attention | FID-50K (↓) |
|---|---|---|---|---|---|
| model | 400K | ✓ | ✓ | ✗ | 56.07 |
| ema | 400K | ✓ | ✓ | ✗ | 56.07 |
| model | 200K | ✓ | ✓ | ✗ | 57.84 |
| model | 300K | ✓ | ✓ | ✗ | 56.95 |
| model | 400K | ✓ | ✓ | ✗ | 56.07 |
| model | 600K | ✓ | ✓ | ✗ | 54.80 |
| model | 800K | ✓ | ✓ | ✗ | **53.83** |
| model | 600K | ✓ | ✓ | Q, K, V | 55.29 |
| model | 600K | ✓ | ✓ | K, V | 55.07 |
| model | 600K | ✓ | ✓ | V | 54.93 |
| model | 600K | ✓ | ✓ | Q | 54.82 |
| model | 600K | ✓ | ✓ | O | 54.84 |

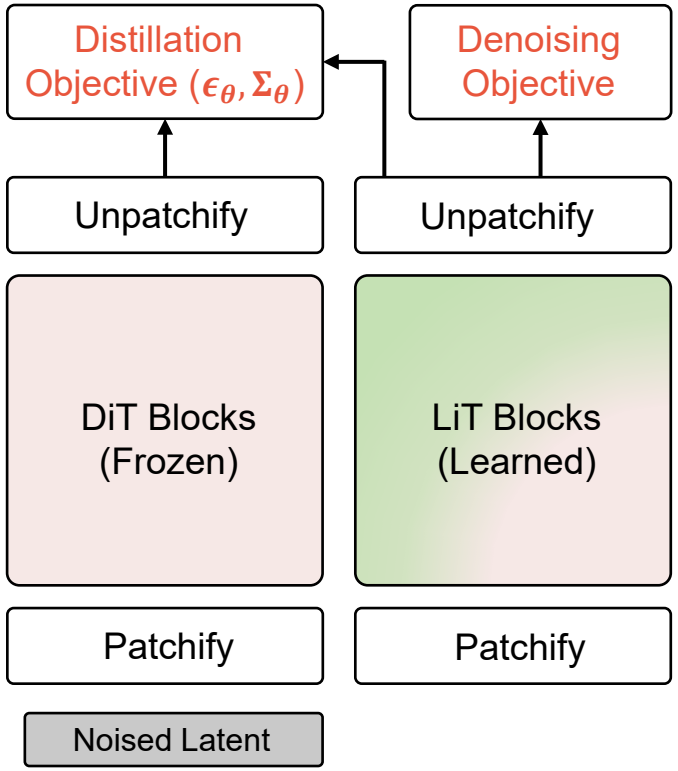
**Guideline 4:**
Projection matrices of query, key, value, and output in linear attention should be initialized randomly.



# Knowledge Distillation

**Guideline 5:**

Hybrid distillation is neccessary for student linear diffusion Transformer. We distill not only the predicted noise but also variances of the reverse diffusion process, but in a moderate way.




| Iter. | Teacher | $\lambda_1$ | $\lambda_2$ | FID-50K (↓) | IS (↑) |
|---|---|---|---|---|---|
| 800K | DiT-S/2 | 0.1 | 0.0 | 55.11 | 26.28 |
| 800K | DiT-XL/2 | 0.0 | 0.0 | <u>53.83</u> | <u>27.16</u> |
| 800K | DiT-XL/2 | 0.1 | 0.0 | 53.05 | 27.43 |
| 800K | DiT-XL/2 | 0.05 | 0.0 | 53.41 | 27.26 |
| 800K | DiT-XL/2 | 0.5 | 0.0 | 51.13 | 28.89 |
| 800K | DiT-XL/2 | 0.1 | 0.05 | 52.76 | 27.70 |
| 800K | DiT-XL/2 | 0.0 | 0.05 | 53.49 | 27.26 |
| 800K | DiT-XL/2 | 0.05 | 0.05 | 53.14 | 27.46 |
| 800K | DiT-XL/2 | 0.5 | 0.05 | **50.79** | **29.17** |

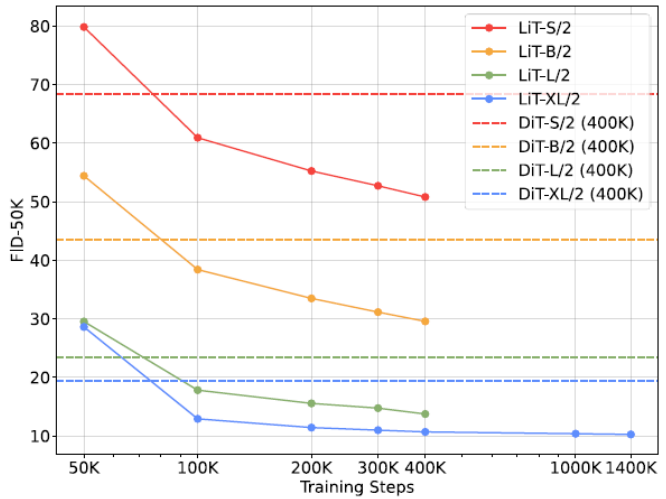
# Outline

- Overview
- Background
- Motivation
- Exploration Roadmap
- Evaluation
- Conclusion

# Class-Conditional ImageNet Results

| Class-Conditional ImageNet 256×256 | | | | |
|---|---|---|---|---|
| Model | FID↓ | IS↑ | Precision↑ | Recall↑ |
| BigGAN-deep [2] | 6.95 | 171.4 | 0.87 | 0.28 |
| StyleGAN-XL [49] | 2.30 | 265.12 | 0.78 | 0.53 |
| ADM [12] | 10.94 | 100.98 | 0.69 | 0.63 |
| ADM-G | 4.59 | 186.70 | 0.82 | 0.52 |
| ADM-G, ADM-U | 3.94 | 215.84 | 0.83 | 0.53 |
| CDM [25] | 4.88 | 158.71 | - | - |
| RIN [29] | 3.42 | 182.0 | - | - |
| LDM-4-G (cfg=1.25) [45] | 3.95 | 178.22 | 0.81 | 0.55 |
| LDM-4-G (cfg=1.50) | 3.60 | 247.67 | 0.87 | 0.48 |
| Simple Diffusion (U-Net) [27] | 3.76 | 171.6 | - | - |
| Mask-GIT [4] | 6.18 | 182.1 | - | - |
| Simple Diffusion (U-ViT, L) | 2.77 | 211.8 | - | - |
| DiT-XL/2 [41] | 9.62 | 121.50 | 0.67 | 0.67 |
| DiT-XL/2-G (cfg=1.25) | 3.22 | 201.77 | 0.76 | 0.62 |
| DiT-XL/2-G (cfg=1.50) | 2.27 | 278.24 | 0.83 | 0.57 |
| SiT-XL [38] (cfg=1.50) | 2.06 | 277.50 | 0.83 | 0.59 |
| DiM-L [53] | 2.64 | - | - | - |
| DiM-H [53] | 2.40 | - | - | - |
| DiffuSSM-XL-G [61] | 2.28 | 259.13 | 0.86 | 0.56 |
| **LiT-XL/2** | 10.24 | 114.79 | 0.666 | 0.674 |
| **LiT-XL/2-G** (cfg=1.25) | 3.60 | 191.06 | 0.758 | 0.623 |
| **LiT-XL/2-G** (cfg=1.50) | 2.32 | 265.20 | 0.824 | 0.574 |

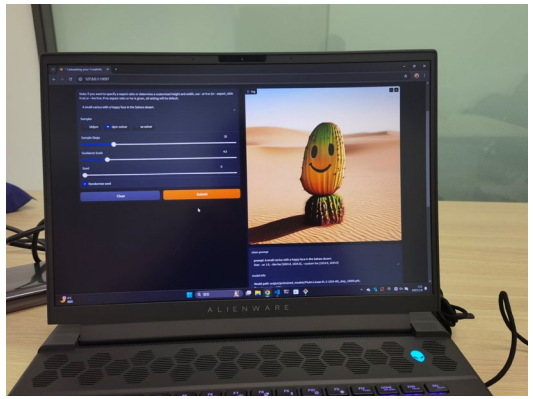
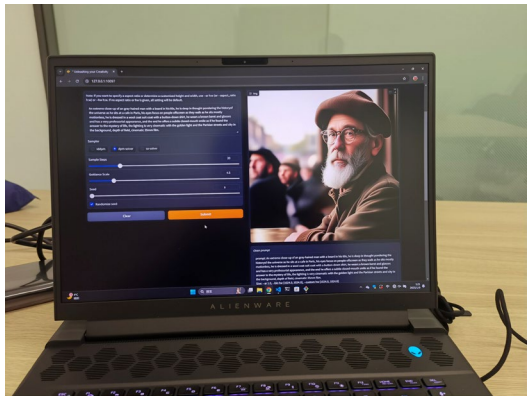
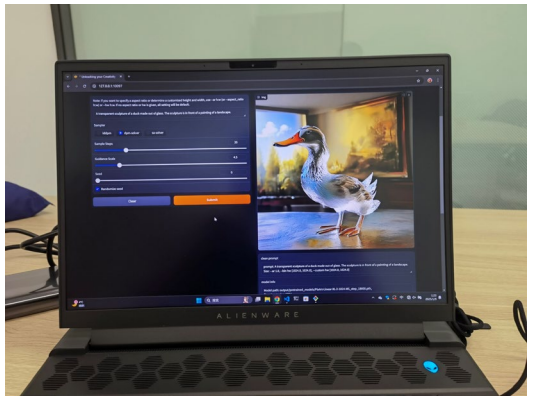
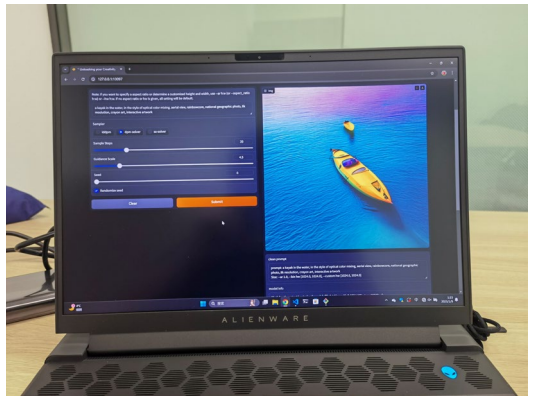
| Class-Conditional ImageNet 512×512 | | | | |
|---|---|---|---|---|
| Model | FID↓ | IS↑ | Precision↑ | Recall↑ |
| BigGAN-deep [2] | 8.43 | 177.90 | 0.88 | 0.29 |
| StyleGAN-XL [49] | 2.41 | 267.75 | 0.77 | 0.52 |
| ADM [12] | 23.24 | 58.06 | 0.73 | 0.60 |
| ADM-U | 9.96 | 121.78 | 0.75 | 0.64 |
| ADM-G | 7.72 | 172.71 | 0.87 | 0.42 |
| ADM-G, ADM-U | 3.85 | 221.72 | 0.84 | 0.53 |
| Simple Diffusion (U-Net) [27] | 4.28 | 171.0 | - | - |
| Mask-GIT [4] | 7.32 | 156.0 | - | - |
| Simple Diffusion (U-ViT, L) | 4.53 | 205.3 | - | - |
| DiT-XL/2 [41] | 12.03 | 105.25 | 0.75 | 0.64 |
| DiT-XL/2-G (cfg=1.25) | 4.64 | 174.77 | 0.81 | 0.57 |
| DiT-XL/2-G (cfg=1.50) | 3.04 | 240.82 | 0.84 | 0.54 |
| SiT-XL [38] (cfg=1.50) | 2.62 | 252.21 | 0.84 | 0.57 |
| **LiT-XL/2** | 14.00 | 92.84 | 0.76 | 0.62 |
| **LiT-XL/2-G** (cfg=1.50) | 3.69 | 207.97 | 0.85 | 0.53 |



# Text-to-Image Generation Results

| Model | #Params. | Single | Two | Count. | Colors | Pos. | Attri. | Overall |
|---|---|---|---|---|---|---|---|---|
| ▲ LDM [45] | 1.4B | 0.92 | 0.29 | 0.23 | 0.70 | 0.02 | 0.05 | 0.37 |
| ▲ SDv1.5 [45] | 0.9B | 0.97 | 0.38 | 0.35 | 0.76 | 0.04 | 0.06 | 0.43 |
| ▲ SDv2.1 [45] | 0.9B | 0.98 | 0.51 | 0.44 | 0.85 | 0.07 | 0.17 | 0.50 |
| ● LlamaGen [51] | 0.8B | 0.71 | 0.34 | 0.21 | 0.58 | 0.07 | 0.04 | 0.32 |
| ● PixArt-$\alpha$ [6] | 0.6B | 0.98 | 0.50 | 0.44 | 0.80 | 0.08 | 0.07 | 0.48 |
| ● PixArt-$\Sigma$ [5] | 0.6B | - | - | - | - | - | - | 0.52 |
| ● Lumina-Next [69] | 2.0B | - | - | - | - | - | - | 0.46 |
| ▶ SEED-X [14] | 17B | 0.97 | 0.58 | 0.26 | 0.80 | 0.19 | 0.14 | 0.49 |
| ▶ Chameleon [52] | 34B | - | - | - | - | - | - | 0.39 |
| ▶ LWM [33] | 7B | 0.93 | 0.41 | 0.46 | 0.79 | 0.09 | 0.15 | 0.47 |
| ★ **LiT** (512px) | 0.6B | 0.97 | 0.43 | 0.42 | 0.79 | 0.09 | 0.12 | 0.47 |
| ★ **LiT** (1024px) | 0.6B | 0.98 | 0.50 | 0.40 | 0.77 | 0.11 | 0.12 | 0.48 |

| Model | #Params. | Attention | Laptop | Latency (1K) | Latency (2K) |
|---|---|---|---|---|---|
| PixArt-$\Sigma$ [5] | 0.6B | KV Compress | ✗ | 4.38s | 32.16s |
| LiT | 0.6B | Linear | ✓ | 3.93s | 14.59s |

A photo of beautiful mountain with realistic sunset and blue lake, highly detailed, masterpiece

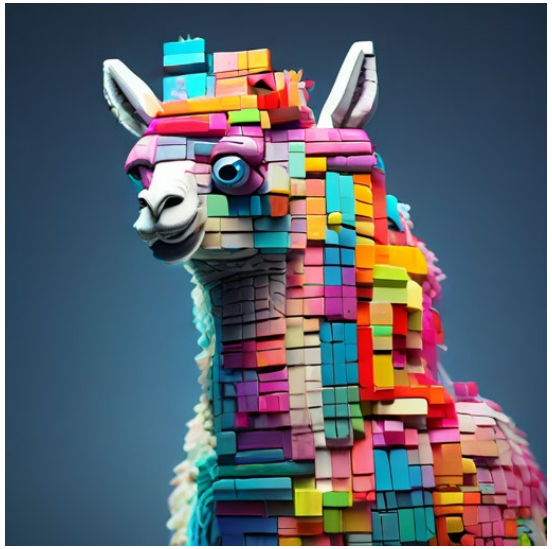
anthropomorphic profile of the white snow owl Crystal priestess , art deco painting, pretty and expressive eyes, ornate costume, mythical, ethereal, intricate, elaborate, hyperrealism, hyper detailed, 3D, 8K

a handsome 24 years old boy in the middle with sky color background wearing eye glasses, it's super detailed with anime style, it's a portrait with delicated eyes and nice looking face

Steampunk makeup, in the style of vray tracing, colorful impasto, uhd image, indonesian art, fine feather details with bright red and yellow and green and pink and orange colours, intricate patterns and details, dark cyan and amber makeup. Rich colourful plumes. Victorian style.

An illustration of a human heart made of translucent glass, standing on a pedestal amidst a stormy sea. Rays of sunlight pierce the clouds, illuminating the heart, revealing a tiny universe within.

A dog that has been meditating all the time



A alpaca made of colorful building blocks, cyberpunk

A blue jay standing on a large basket of rainbow macarons.

A car made out of vegetables.

A cute orange kitten sliding down an aqua slide. happy excited. 16mm lens in front. we see his excitement and scared in the eye. vibrant colors. water splashing on the lens

A realistic landscape shot of the Northern Lights dancing over a snowy mountain range in Iceland.

portrait photo of a girl, photograph, highly detailed face, depth of field

Frog, in forest, colorful, no watermark, no signature, in forest, 8k

Astronaut in a jungle, cold color palette, muted colors, detailed, 8k

dog

An extreme close-up of an gray-haired man with a beard in his 60s, he is deep in thought pondering the history of the universe as he sits at a cafe in Paris, his eyes focus on people offscreen as they walk as he sits mostly motionless, he is dressed in a wool coat suit coat with a button-down shirt , he wears a brown beret and glasses and has a very professorial appearance, and the end he offers a subtle closed-mouth smile as if he found the answer to the mystery of life, the lighting is very cinematic with the golden light and the Parisian streets and city in the background, depth of field, cinematic 35mm film.

Game-Art - An island with different geographical properties and multiple small cities floating in space

Pirate ship trapped in a cosmic maelstrom nebula, rendered in cosmic beach whirlpool engine, volumetric lighting, spectacular, ambient lights, light pollution, cinematic atmosphere, art nouveau style, illustration art artwork by SenseiJaye, intricate detail.

A boy and a girl fall in love

Editorial photoshoot of a old woman, high fashion 2000s fashion

Crocodile in a sweater

stars, water, brilliantly, gorgeous large scale scene, a little girl, in the style of dreamy realism, light gold and amber, blue and pink, brilliantly illuminated in the background.

beautiful lady, freckles, big smile, blue eyes, short ginger hair, dark makeup, wearing a floral blue vest top, soft light, dark grey background

3d digital art of an adorable ghost, glowing within, holding a heart shaped pumpkin, Halloween, super cute, spooky haunted house background

# Outline

- Overview
- Background
- Motivation
- Exploration Roadmap
- Evaluation
- Conclusion

# Conclusion

**Summary:**

We provide a suite of ready-to-use guidelines, answering how to convert a pre-trained DiT into an efficient linear DiT cost-effectively.

Paper: https://arxiv.org/pdf/2501.12976

On-device Demo: https://www.youtube.com/watch?v=X8aIrYYjFKU&t=1s