上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

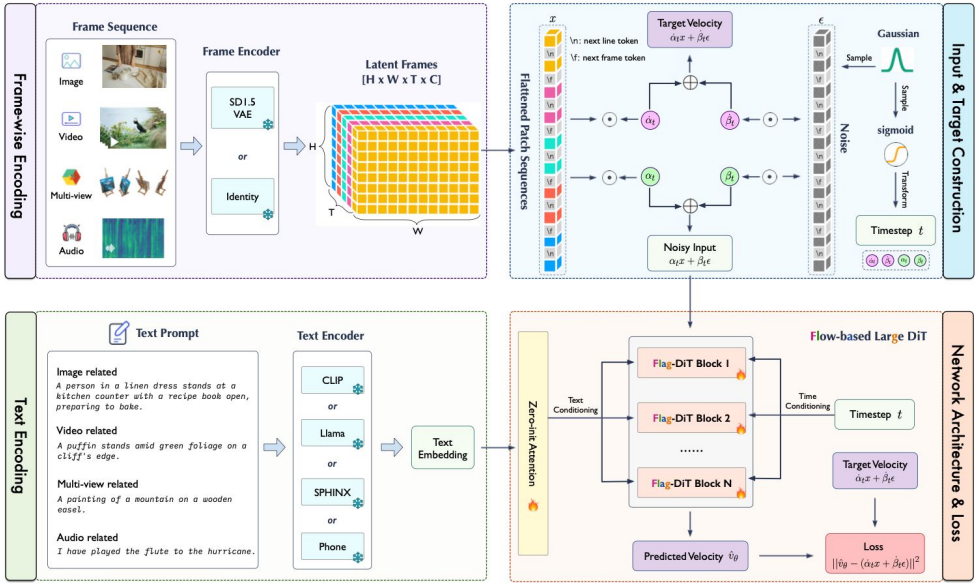
NEURAL INFORMATION PROCESSING SYSTEMS

Lumina-Next: Making Lumina-T2X Stronger and Faster with Next-DiT

Le Zhuo · Ruoyi Du · Han Xiao · Yangguang Li · Dongyang Liu · Rongjie Huang · Wenze Liu · Xiangyang Zhu · Fu-Yun Wang · Zhanyu Ma · Xu Luo · Zehan Wang · Kaipeng Zhang · Lirui Zhao · Si Liu · Xiangyu Yue · Wanli Ouyang · Yu Qiao · Hongsheng Li · Peng Gao

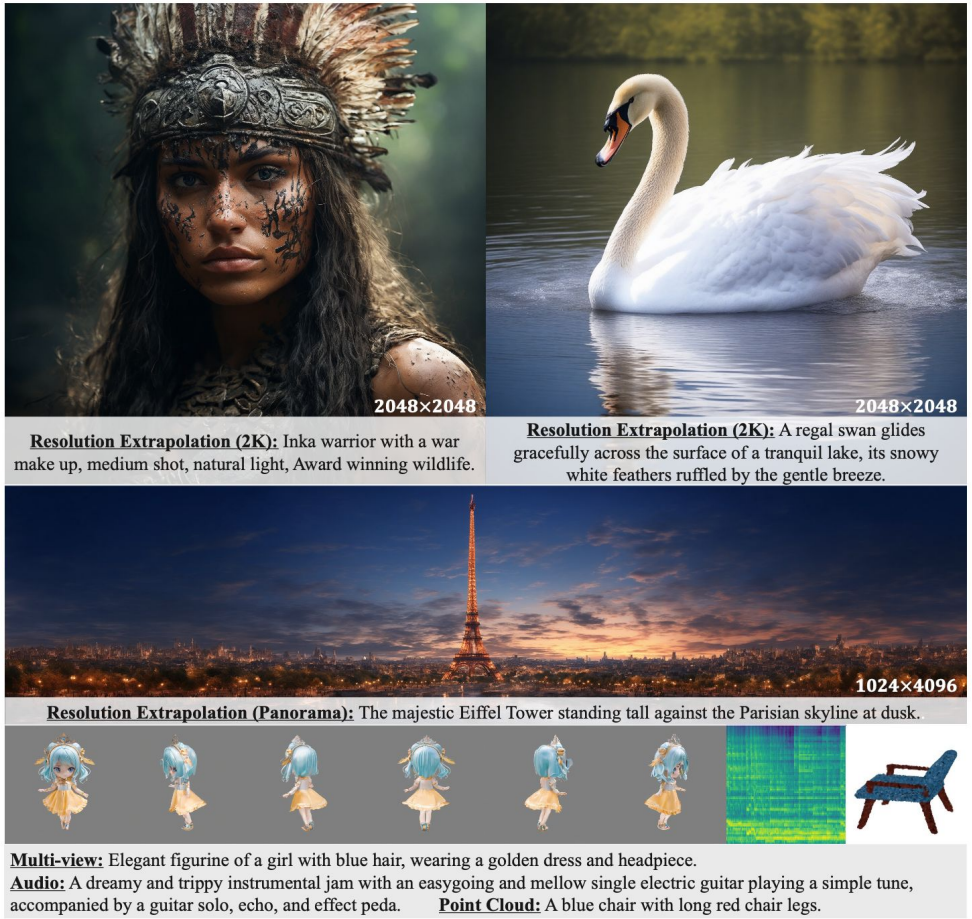
Overview

- Lumina-T2X proposed a versatile generative framework for multiple modalities
- Flow-based Large Diffusion Transformer (Flag-DiT)
  - Flow Matching Formulation
  - QK-Norm + RMSNorm
  - Zero-Init Attention
  - RoPE
  - .....
- 5B Flag-DiT with a 7B LLaMA
- Text-to-Image, Video, Multiview, and Audio Generation



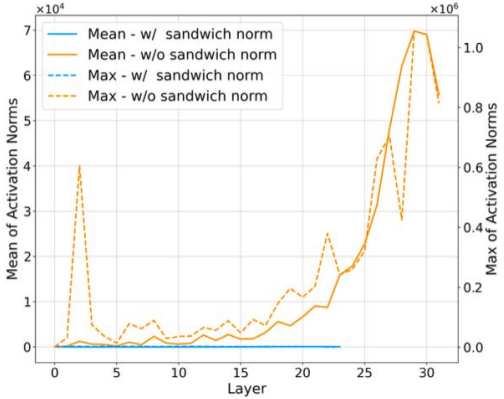
Overview

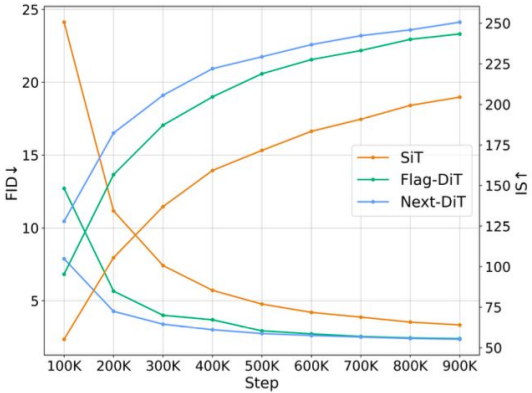
- Introduce Lumina-Next, making our model stronger and faster!
  - Improved Architecture
  - Stronger Extrapolation
  - Faster Sampling
  - More Languages
  - More Modalities

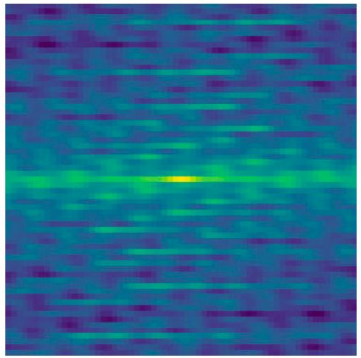


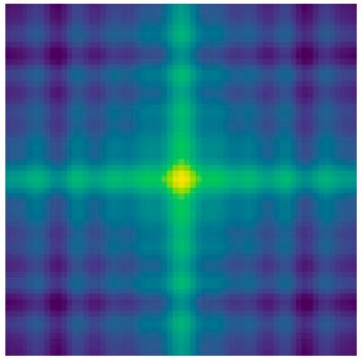
Improved Architecture - Next-DiT

- Cumbersome 1D RoPE + learnable tokens -> 3D RoPE
- Training&Inference instabilities -> Sandwich Normalization
- Huge memory bandwidth -> Grouped-Query Attention
- Faster convergence on ImageNet



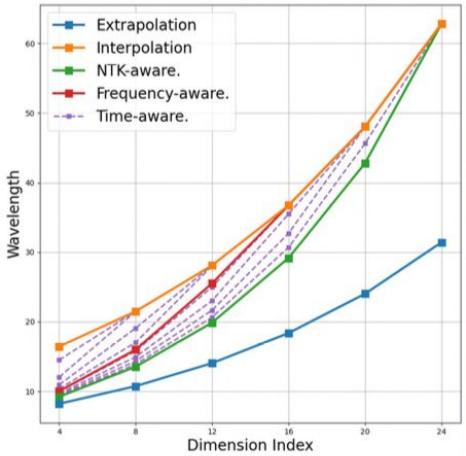






Stronger Resolution Extrapolation

- Naive methods: Position Interpolation (global), NTK-Aware Scaled RoPE (local)
- Frequency-Aware Scaled RoPE
- Time-Aware Scaled RoPE

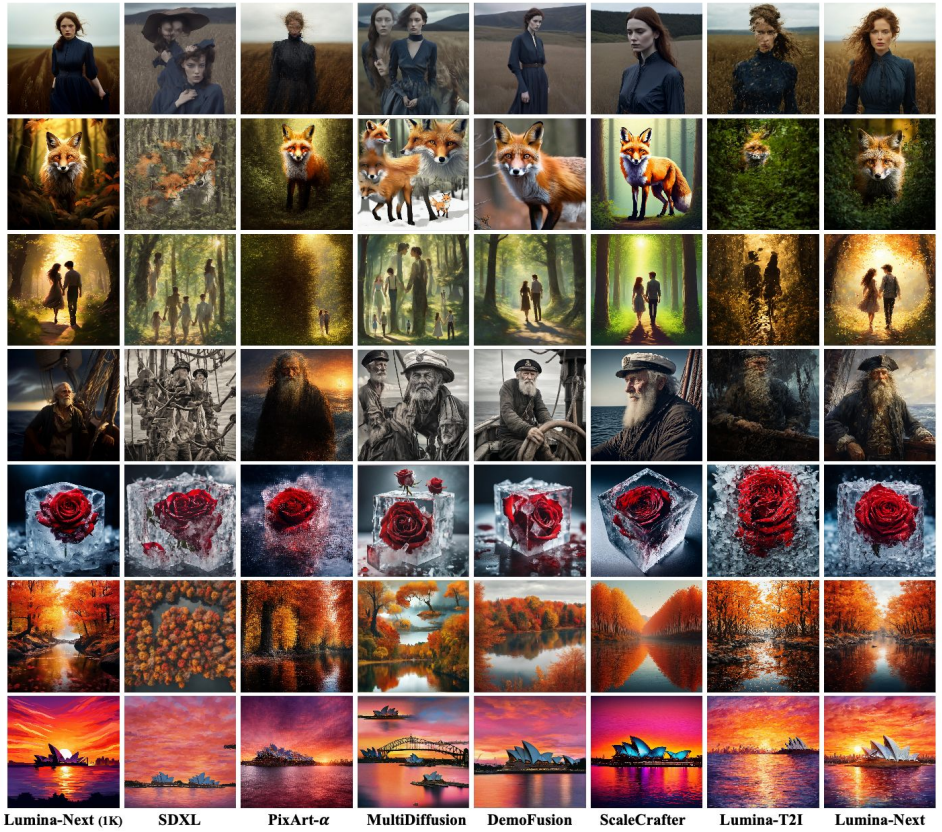


(a) Wavelength of the RoPE embeddings under different strategies

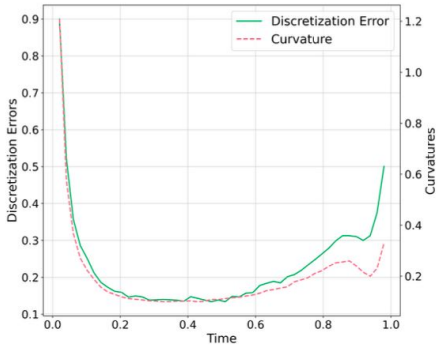
(b) Original Resolution (1K Inference) (c) Position Extrapolation (d) Position Interpolation (e) NTK-aware Scaled RoPE (f) Frequency-aware Scaled RoPE (g) Time-aware Scaled RoPE

Stronger Resolution Extrapolation

- The improved architecture and Time-Aware Scaled RoPE enables efficient and flexible resolution extrapolation with any aspect-ratio



Faster Sampling

- The overall time complexity of sampling can be written as N × T
  - Number of function evaluations N
  - Inference time of a single function evaluation T
- Fewer sampling steps
  - The time discretization errors of flow models are different from diffusion models
  - Propose new time schedules integrated with higher-order ODE solvers
  - 

$$\textsc{Rational} \quad t' = \frac{t}{\sigma - t + \sigma t}$$

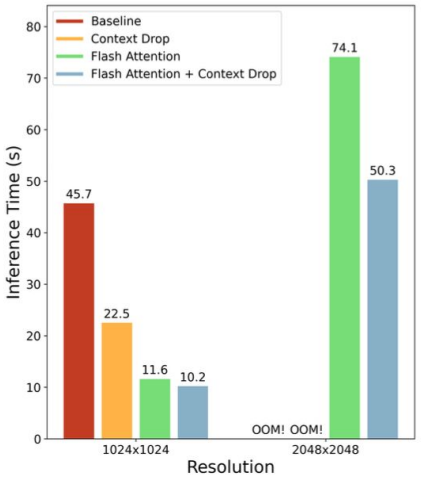
$$\textsc{Sigmoid} \quad t' = \begin{cases} \frac{1}{1+\exp(-\alpha(t-\mu))} & \text{if } t < \mu \\ 1 - \frac{1}{1+\exp(\beta(t-\mu))} & \text{if } t \geq \mu \end{cases}$$

Faster Sampling

- Fewer sampling steps
  - The time discretization errors of flow models are different from diffusion models
  - Propose new time schedules integrated with higher-order ODE solvers

Faster Sampling

- Faster network evaluation
  - Context Drop reduces spatial redundancy of key and value tokens by average pooling
  - Add time-awareness tailored for denoising process



Stronger Language Understanding

- Instead of using CLIP or T5 as text encoder, Lumina-Next uses decoder-based LLM, like Gemma
- LLMs exhibit much strong language understanding capabilities, which enables zero-shot multilingual generation

Stronger Language Understanding

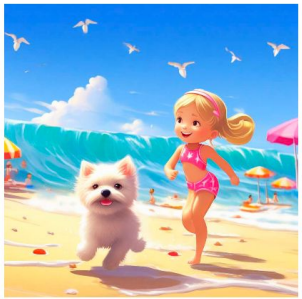
- Understanding Emojis

Stronger Language Understanding

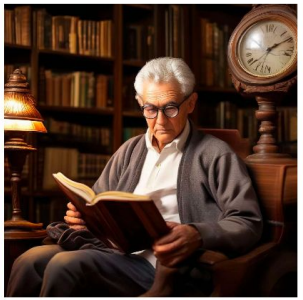
- Understanding Chinese poems

Stronger Language Understanding

- Understanding long and detailed prompts



A young girl plays joyfully with a small, fluffy white dog on a sandy beach. She's wearing a pink swimsuit with white polka dots and a matching headband. Her blonde hair is tied into pigtails. The girl and the dog run towards the gently crashing waves, leaving footprints behind them. In the distance, there are colorful beach umbrellas and people lounging on beach towels. The sky is clear and blue, with seagulls soaring above.



An elderly man sits comfortably in a cozy library, reading a thick, leather-bound book. He's wearing a grey cardigan over a white shirt, with round spectacles perched on his nose. His white hair is neatly combed back. Behind him, wooden bookshelves are filled with rows of books, some old and some new. A vintage lamp on a side table casts a warm light, illuminating his serene face. A classic grandfather clock ticks softly in the background.

A serene, coastal fishing village nestled along a rocky shoreline. Quaint, colorful houses with terracotta roofs line the narrow, cobblestone streets. Fishing boats bob gently in the harbor, their nets drying in the sea breeze. Seagulls circle overhead as the sun casts a golden glow over the calm, turquoise waters.

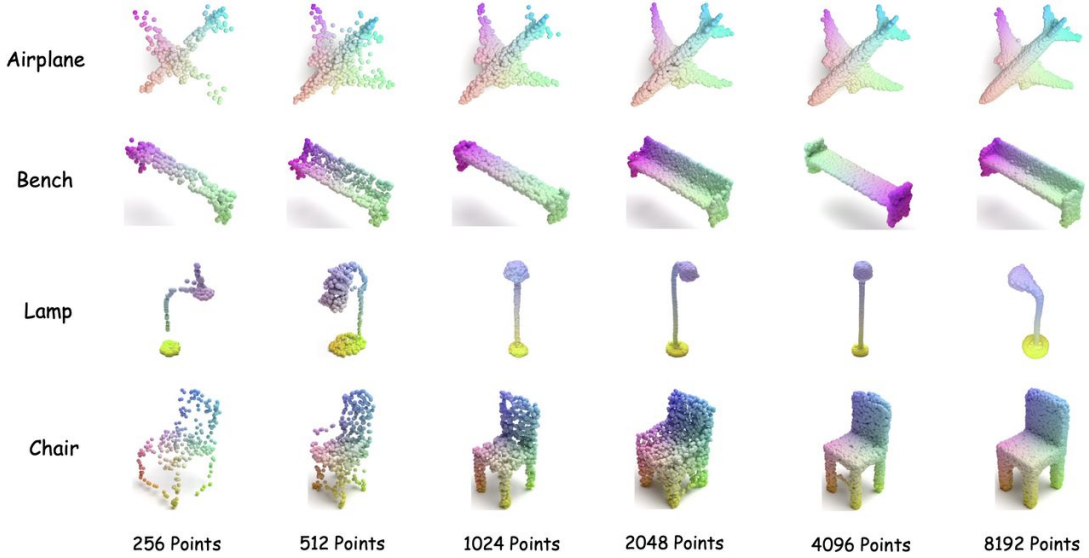
A grand, medieval castle perched atop a rocky cliff, overlooking a vast, undulating landscape of forests and meadows. The castle's imposing stone towers and turrets rise majestically against a backdrop of a brilliant blue sky. A drawbridge spans a deep, mist-filled moat, and banners flap in the gentle breeze, displaying regal coats of arms.

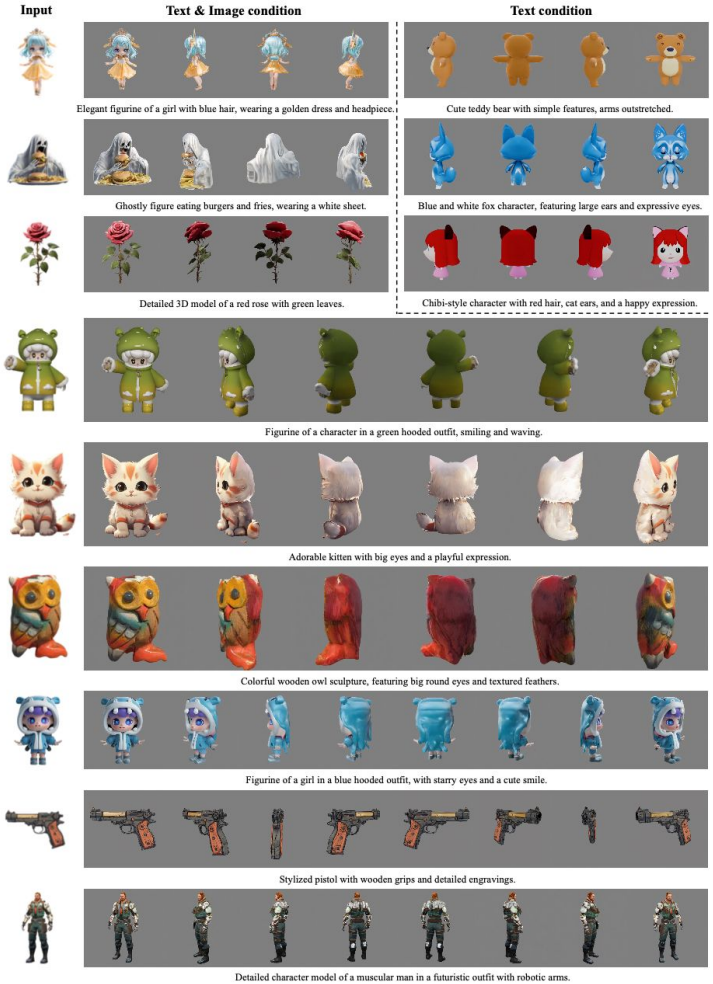
在广袤无垠的沙漠中，夜幕降临时，天际被无数璀璨的星星点缀得如同一幅银色的华丽锦绸。月亮将温柔的月光洒在金黄色的沙丘上，投射出柔和的光影。顽强的树挺立在沙丘之间，静静地伫立在这寂静的夜色中。一个孤独的旅人缓缓行走，仿佛在追寻着星辰的指引。

一间布置典雅的复古书房，墙上挂满了古老的书画、陈旧的地图和稀有的收藏品。红木书架上摆放着各类珍贵的典籍，书脊泛着岁月的光泽。一张大书桌镇坐在房间中央，上面放着一本打开的古书，旁边是一支装饰精美的鹅毛笔和一个古董墨水瓶。昏黄的台灯发出柔和的光芒，映照在桌上的黄铜地球仪上。窗外的晚霞透过厚重的窗帘洒进屋内，给整个房间增添了一抹温暖的色彩。

More Modalities

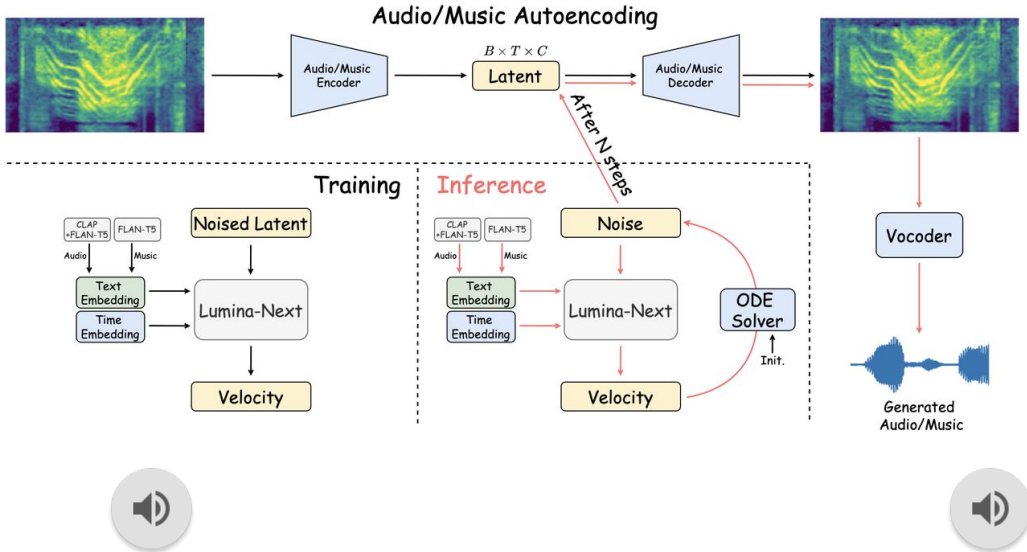
- Multiview and point cloud





More Modalities

- Audio and music generation



A honking horn from an oncoming train

A large bell rings out multiple times

More Modalities

- Audio and music generation

This instrumental progressive rock song features a complex, electric guitar solo with impressive tapping techniques.

The upbeat instrumental song features punchy digital drums, lively piano harmony, and funky bass line, accompanied by perky synthesised violins and various background sounds superimposed by music, creating a happy and lively atmosphere with a fast tempo and a sound of beeping.

More Modalities

- Beyond generation, we can easily adapt Lumina-Next to any resolution recognition

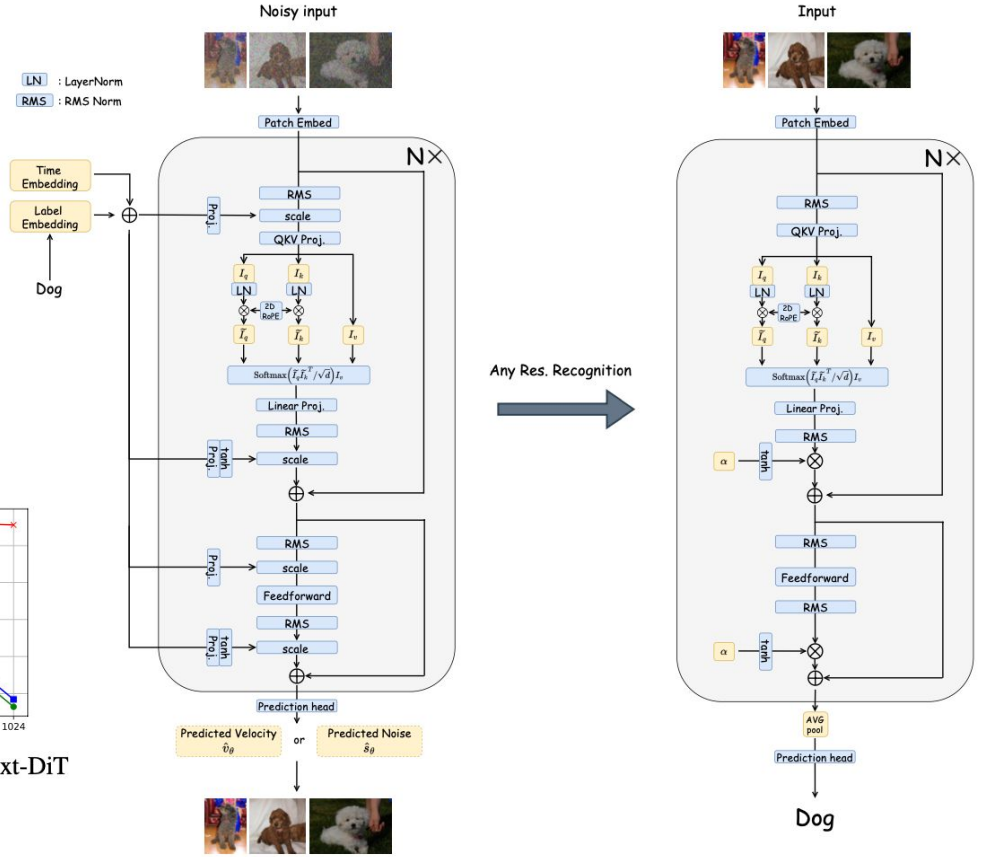


Table 1: Comparison of Next-DiT with DeiT [73] on ImageNet classification.

Model	Params	Setting	Resolution	Top-1 Acc(%)
DeiT-base [73]	86M	300E	224 × 224	81.8
Next-DiT	86M	100E	224 × 224	81.6
Next-DiT	86M	300E	224 × 224	82.3
DeiT-base [73]	86M	300E	Flexible	67.2
Next-DiT	86M	300E+30E	Flexible	84.2

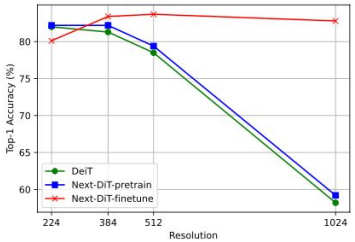


Figure 19: Performance of Next-DiT across different resolutions.

Thanks for Listening!

Code & Model