UNIVERSITY OF MINNESOTA
Driven to Discover℠

NEURAL INFORMATION PROCESSING SYSTEMS

IMAGINE
imagine.umn.edu

SPOTLIGHT

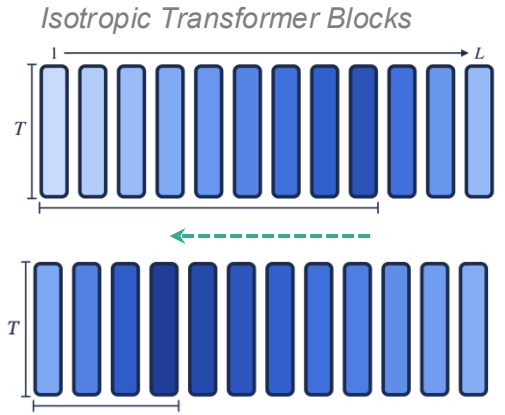
UDT: Reconciling U-Nets and Diffusion Transformers with Data-Adaptive Token Reduction

Junno Yun, Yaşar Utku Alçalar, and Mehmet Akçakaya
University of Minnesota, Minneapolis

Motivation

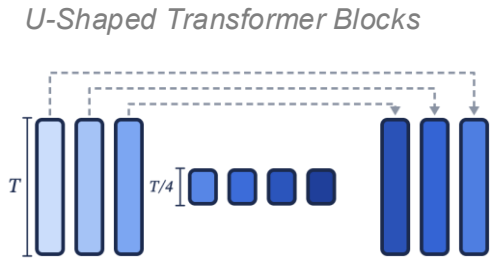
Standard Diffusion Transformers





- **Consist of sequential isotropic transformer blocks**
- **Learn progressively stronger semantic representations from early to later layers**
  → Creates an imbalanced encoder-decoder structure, limiting later-layer denoising capacity
- **Training regularization**, such as representation alignment (REPA), encourages earlier formation of strong representations
  → Effectively shifting the encoder forward and freeing later layers for denoising, *but still requires additional regularization designs or external encoders*

Existing U-shaped Diffusion Transformer



- **Explicit Encoder-Decoder Architectures :** Spatial down-/up-sampling and skip connections
- **Multi-scale hierarchical modeling through downsampling**, together with **skip connections** to mitigate information loss, **improves representation learning and accelerates convergence**.
- ***Howerver,*** *Fixed 2×2 local downsampling* ignores cross-image token similarity and distorts token features representation, weakening token-wise interactions in self-attention.

Proposed Method : U-Net Diffusion Transformer (UDT)

"Can we introduce an **encoder-decoder architecture** into diffusion transformers ***without*** disrupting their *representational characteristics and self-attention dynamics*?"

1. **Transformer-friendly Spatial Down-/Up- Sampling : Token Merge (ToMe[1])**

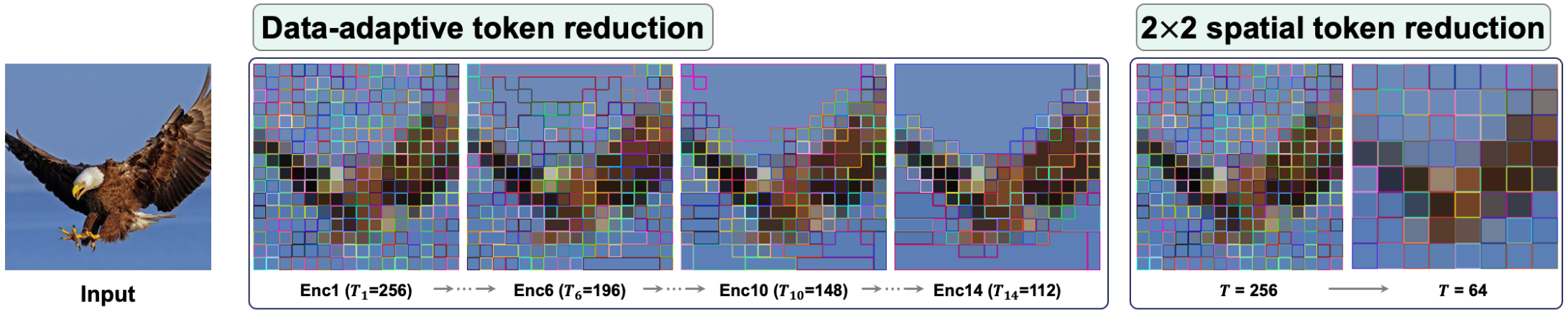
2. **Encoder-Decoder in Diffusion Transformer : U-Net Architecture**

[1]Bolya, Daniel, et al. ICLR 2023

Proposed Method : U-Net Diffusion Transformer (UDT)

Token Merge for Spatial Reduction in DiT

- **Downsampling:** Tokens are progressively merged based on attention-key similarity across layers
- **Upsampling:** Recorded merge indices enable approximate recovery of the original token layout

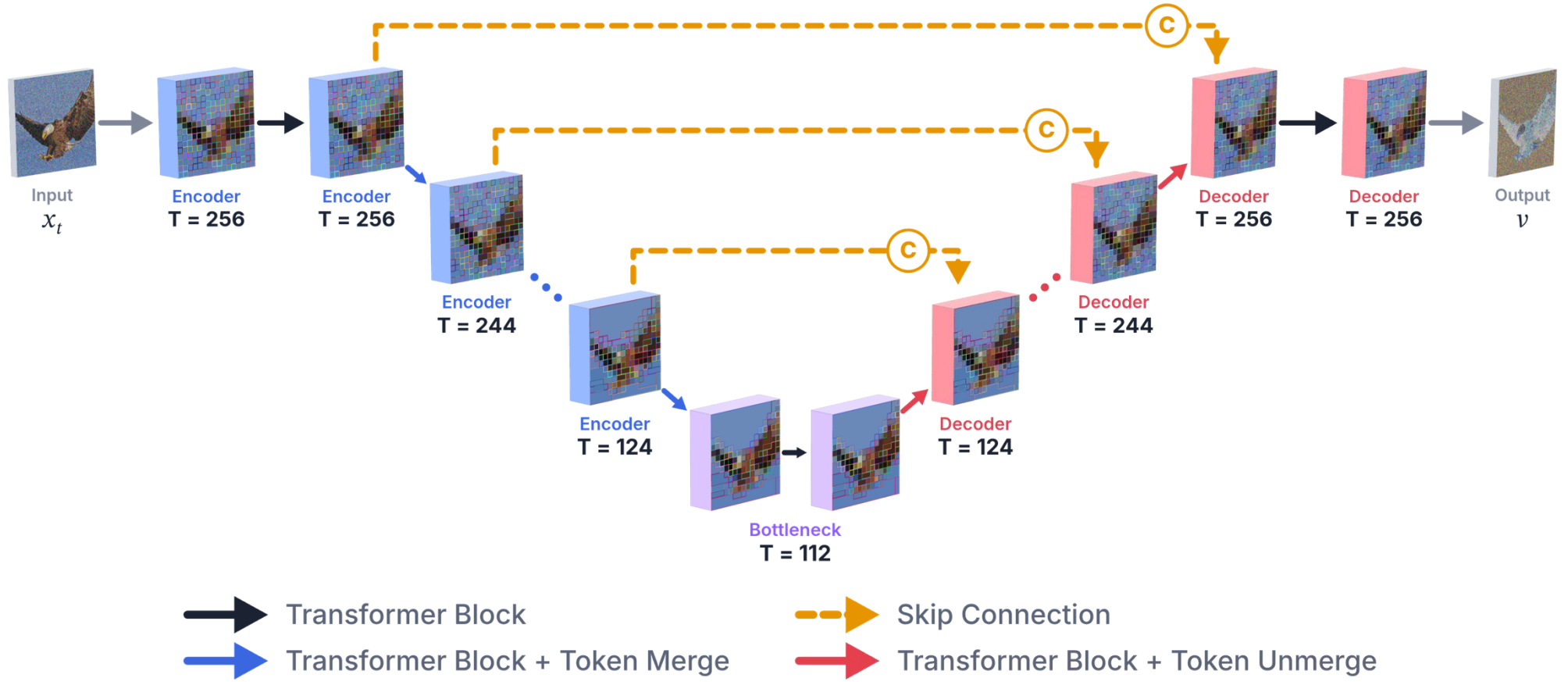


- Gradual data-adaptive token reduction
- Parameter-free
- Preserving the DiT token representation dimension

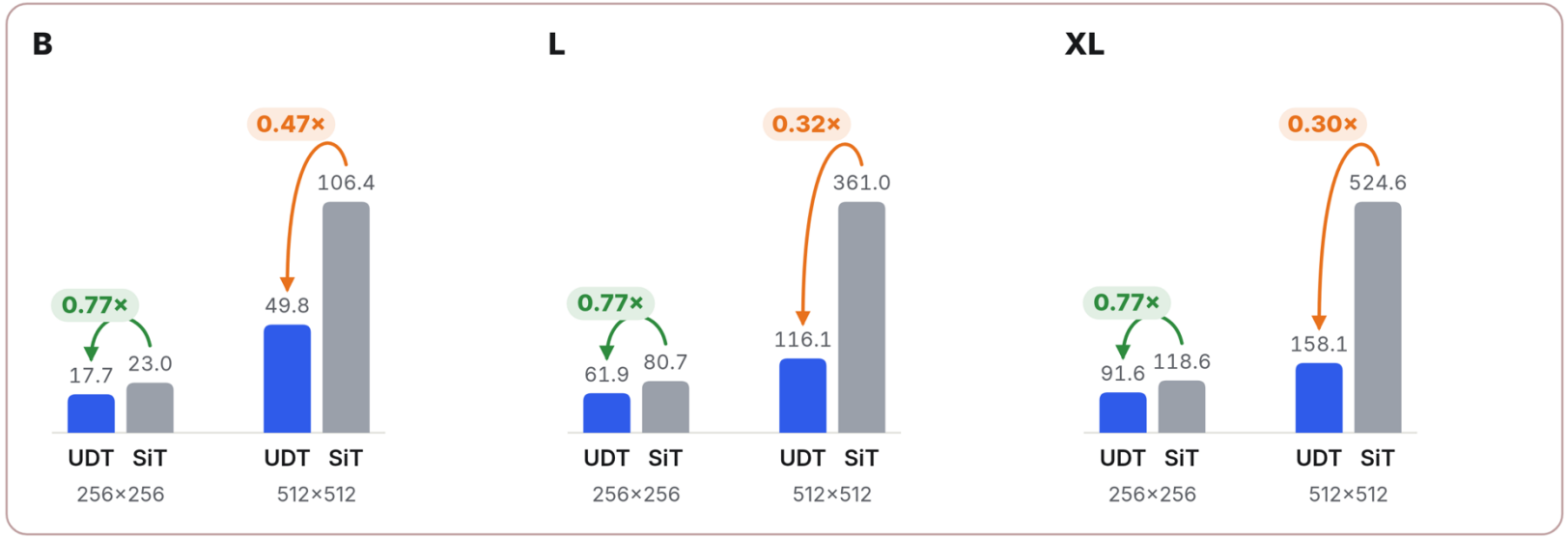
- Fixed 2×2 local downsampling
- Increased parameters count
- Token dimension transformation via 2×2 grouping and projection

Proposed Method : U-Net Diffusion Transformer (UDT)

U-Net Architecture



Computational Benefits of UDT



- At 256×256, **UDT needs only ~0.77× the GFLOPs of SiT** at every model size (B, L, XL)

- **Compute stays low as resolution grows**
  - Going from 256×256 to 512×512 raises the token count 4×
  - SiT at 512×512 needs ~4.5× the GFLOPs while **UDT needs only 1.3–2.2×**
  - At 512×512, **UDT needs only 0.30 – 0.47× the GFLOPs of SiT**

Faster Convergence, Enhanced Generative Performance

without CFG

ImageNet 256×256

FID **7.0** at 60 epochs (**UDT** *without* REPA)

60 — 80 (SiT + REPA) — 1400

FID **7.6** **at 40 epochs (UDT with REPA)**

40 — 1400

25~40× faster

at 1400 epochs (SiT) FID **7.9**

0 — 1400

with CFG

ImageNet 256×256

SD-VAE

	Epoch	FID↓	IS↑
SiT-XL/2	1400	2.06	270.3
SiT-XL/2 + REPA	800	1.42	305.7

FID **1.42** & IS 308.4 at 350 epochs (**UDT** *without* REPA)

FID **1.38** **& IS 306.3 at 320 epochs (UDT with REPA)**

VA-VAE

FID **1.35** **& IS 322.0 at 500 epochs (UDT)**

ImageNet 512×512

FID **1.58** **& IS 327.7 at 200 epochs (UDT *without* REPA)**

	Epoch	GFLOPs	FID↓	IS↑
SiT-XL/2	600	524.6	2.62	252.2
SiT-XL/2 + REPA	200	524.6	2.10	274.6
UDT-XL/2	**100**	**158.1**	**1.63**	**320.5**
UDT-XL/2	**200**	**158.1**	**1.58**	**327.7**

Drop-in Replacement for Isotropic DiT Variants

ϵ-prediction

Model	Epoch	FID↓
DiT-B/2	80	43.5
UDT-DiT-B/2	80	**30.6**
DiT-L/2	80	23.3
DiT-L/2 + REPA	80	15.6
UDT-DiT-L/2	80	**13.5**
DiT-XL/2	80	19.5
DiT-XL/2 + REPA	80	12.3
UDT-DiT-XL/2	80	**11.2**

Pixel-level DiT

Model	Epoch	FID↓	FID*↓
JiT-B/16†	50	82.6	13.6
JiT-B/16†	200	66.1	4.7
UDT-JiT-B/16+	100	64.6	6.3
UDT-JiT-B/16+	200	**58.9**	**4.6**
JiT-L/16†	50	62.2	6.9
JiT-L/16†	200	48.0	3.0
UDT-JiT-L/16+	50	47.9	5.8
UDT-JiT-L/16+	200	**37.7**	**2.8**

VAE Modification

Model	Tokenizer	Epoch	FID*↓	IS*↑
LightningDiT-XL/1†	VA-VAE	64	2.11	252.3
UDT-XL/1+	VA-VAE	64	**1.99**	**265.8**
LightningDiT-XL/1†	VA-VAE	800	1.35	295.3
UDT-XL/1+	VA-VAE	**500**	**1.35**	**322.0**

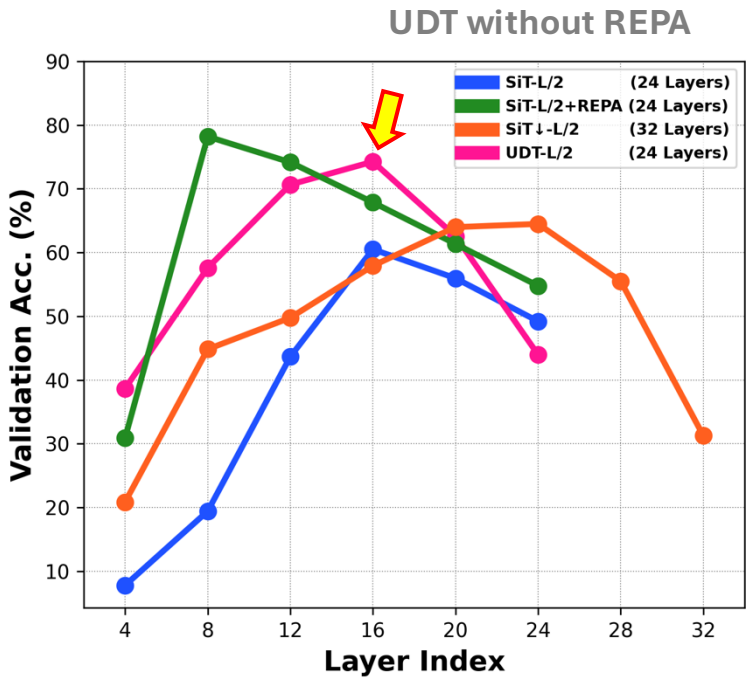
T2I Model

Model	Iter.	FID*↓
MMDiT	150K	5.5
MMUDT	150K	**4.7**

- **Drop UDT into any isotropic DiT for faster training and better generation**

Improving Representations

Linear Probing



(a) Classification on ImageNet

Method	Layer (Total)	mIoU	pixAcc
SiT-L/2	16 (24)	38.30	84.81%
SiT↓-L/2	24 (32)	41.02	85.92%
UDT-L/2	16 (24)	**55.00**	**89.78%**

(b) Semantic Segmentation on PASCAL

Links & Contact Information

Please refer to the paper and code for additional results and details.

Paper

Project Page

IMAGINE Lab's homepage

Email : {yun00049, alcal029, akcakaya}@umn.edu

Thank you !