"Principal Components" Enable A New Language of Images

by ***Xin Wen****, *Bingchen Zhao**, *Ismail Elezi*, *Jiankang Deng*, and *Xiaojuan Qi*

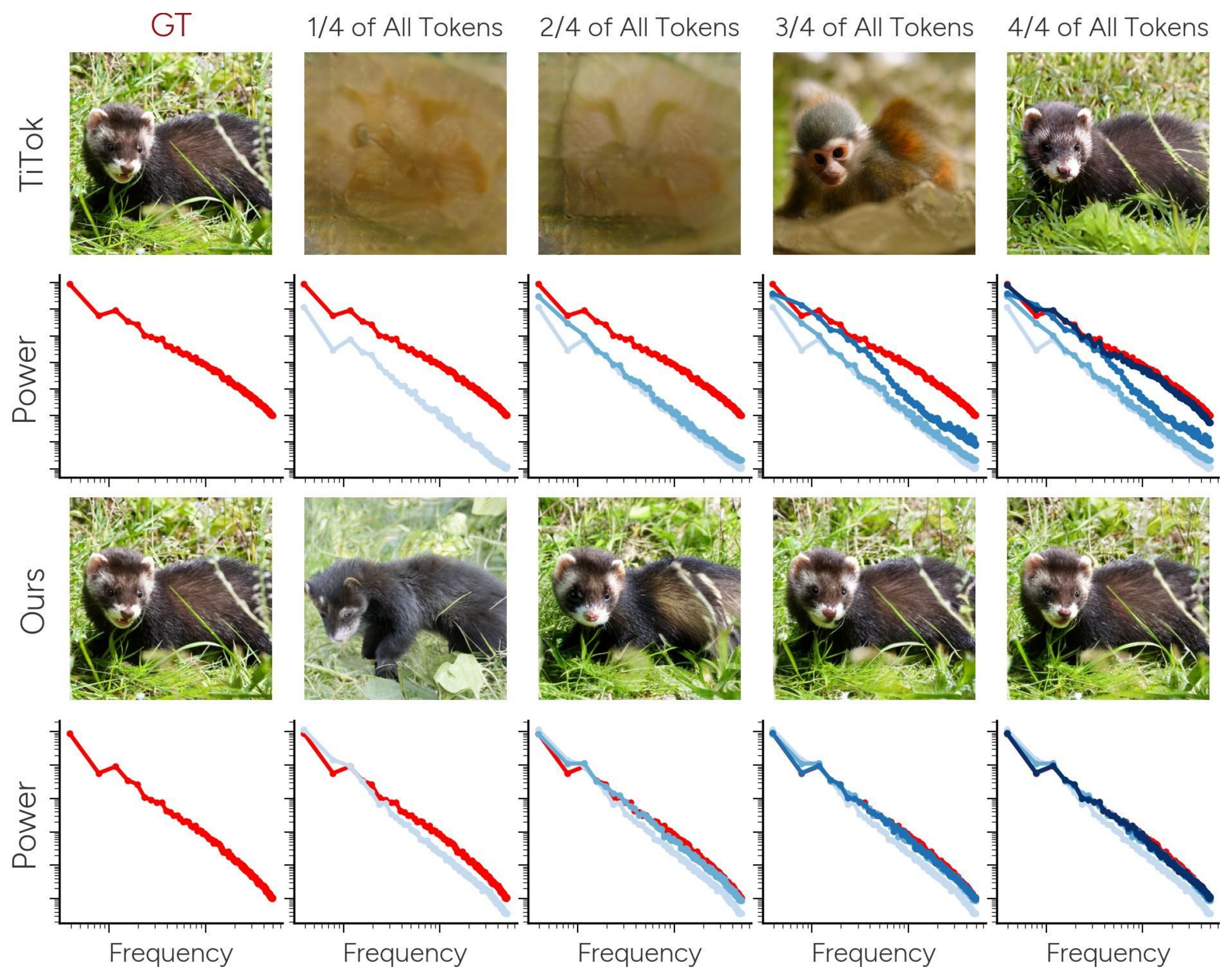
ICCV OCT 19-23, 2025 HONOLULU HAWAII

香 港 大 學 THE UNIVERSITY OF HONG KONG
THE UNIVERSITY of EDINBURGH
HUAWEI
IMPERIAL

What makes this tokenizer different?

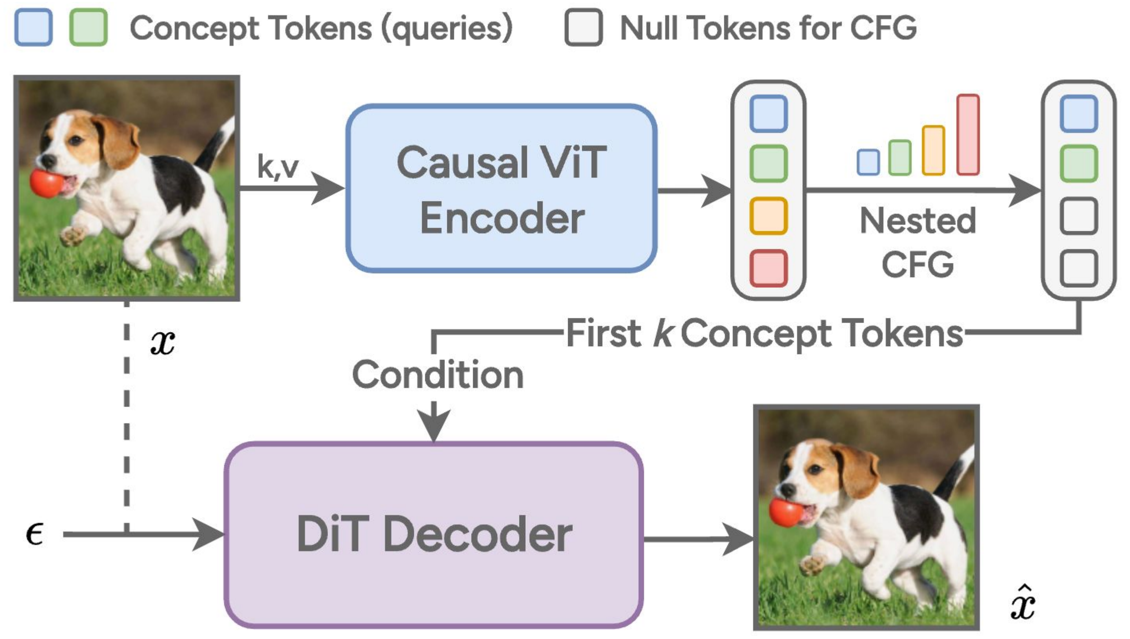
- **1D Causal Tokenization** – An ordered structure where token importance follows a hierarchical pattern.
- **PCA-Like Structure** – Earlier tokens contain most significant information, while later tokens refine details.
- **Semantic-Spectrum Decoupling** – ensuring tokens capture high-level rather than low-level artifacts.

1. Semantic-Spectrum Decoupling



TiTok: both semantic details and spectral power increase simultaneously.

Ours:
1) Semantic clarity w/ few tokens.
2) Consistent spectrum across token counts.
3) Coarse-to-fine hierarchy, mirroring global precedence effect in human vision.



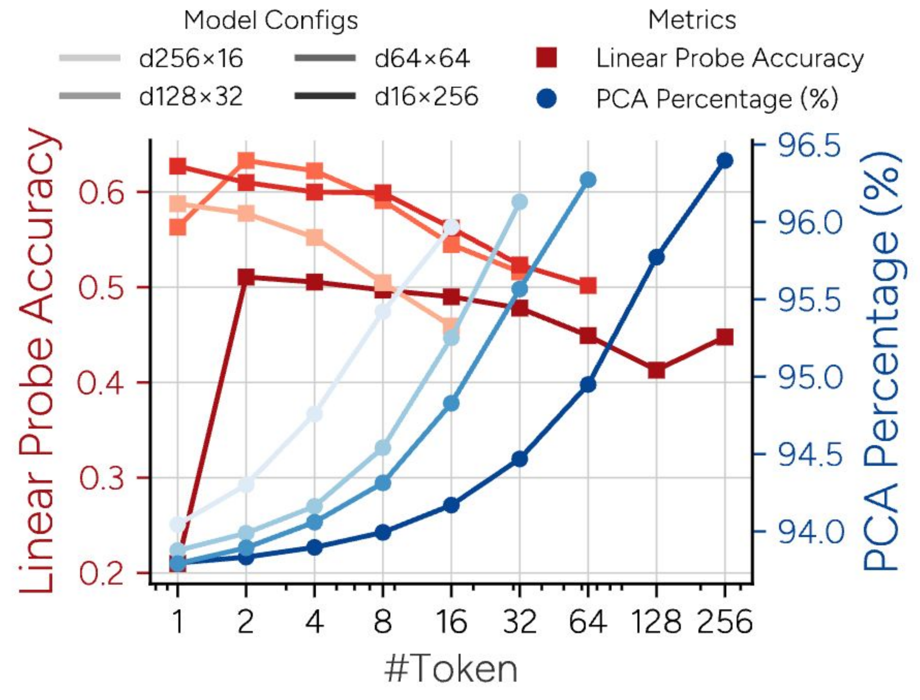
2. Semanticist

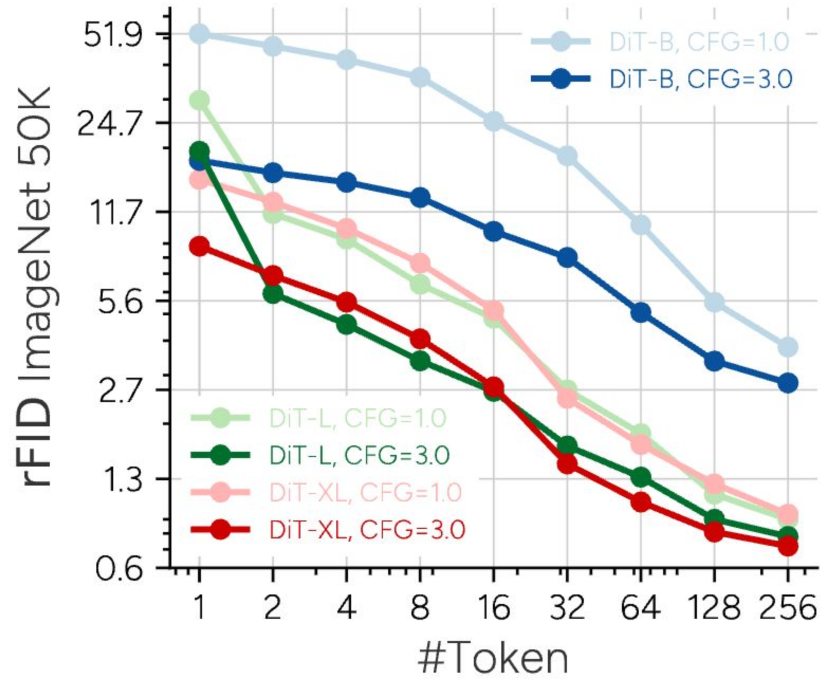
1) Causal ViT Encoder
2) Nested Dropout for PCA-like Structure
3) Diffusion-Based DiT Decoder

3. Reconstruction with A Variable Number of Tokens

Method	#Token	Dim.	VQ	rFID↓	PSNR↑	SSIM↑	Gen. Model	Type	#Token	#Step	gFID↓	IS↑
MaskBit [55]	256	12	✓	1.61	–	–	MaskBit	Mask.	256	256	1.52	328.6
RCG (cond.) [27]	1	256	✗	–	–	–	MAGE-L	Mask.	1	20	3.49	215.5
MAR [28]	256	16	✗	1.22	–	–	MAR-L	Mask.	256	64	1.78	296.0
TiTok-S-128 [60]	128	16	✓	1.71	–	–	MaskGIT-L	Mask.	128	64	1.97	281.8
TiTok-L-32 [60]	32	8	✓	2.21	–	–	MaskGIT-L	Mask.	32	8	2.77	194.0
VQGAN [13]	256	16	✓	7.94	–	–	Tam. Trans.	AR	256	256	5.20	280.3
ViT-VQGAN [57]	1024	32	✓	1.28	–	–	VIM-L	AR	1024	1024	4.17	175.1
RQ-VAE [26]	256	256	✓	3.20	–	–	RQ-Trans.	AR	256	64	3.80	323.7
VAR [49]	680	32	✓	0.90	–	–	VAR-$d16$	VAR	680	10	3.30	274.4
ImageFolder [29]	286	32	✓	0.80	–	–	VAR-$d16$	VAR	286	10	2.60	295.0
LlamaGen [48]	256	8	✓	2.19	20.79	0.675	LlamaGen-L	AR	256	256	3.80	248.3
CRT [39]	256	8	✓	2.36	–	–	LlamaGen-L	AR	256	256	2.75	265.2
Causal MAR [28]	256	16	✗	1.22	–	–	MAR-L	AR	256	256	4.07	232.4
SEMANTICIST w/ DiT-L	256	16	✗	0.78	21.61	0.626	ϵLlamaGen-L	AR	32	32	2.57	260.9
SEMANTICIST w/ DiT-XL	256	16	✗	0.72	21.43	0.613	ϵLlamaGen-L	AR	32	32	2.57	254.0

Competitive understanding, reconstruction, and AR generation.
Top: quantitative results on ImageNet reconstruction and autoregressive generation.
Left: linear probing; Right: reconstruction FID.

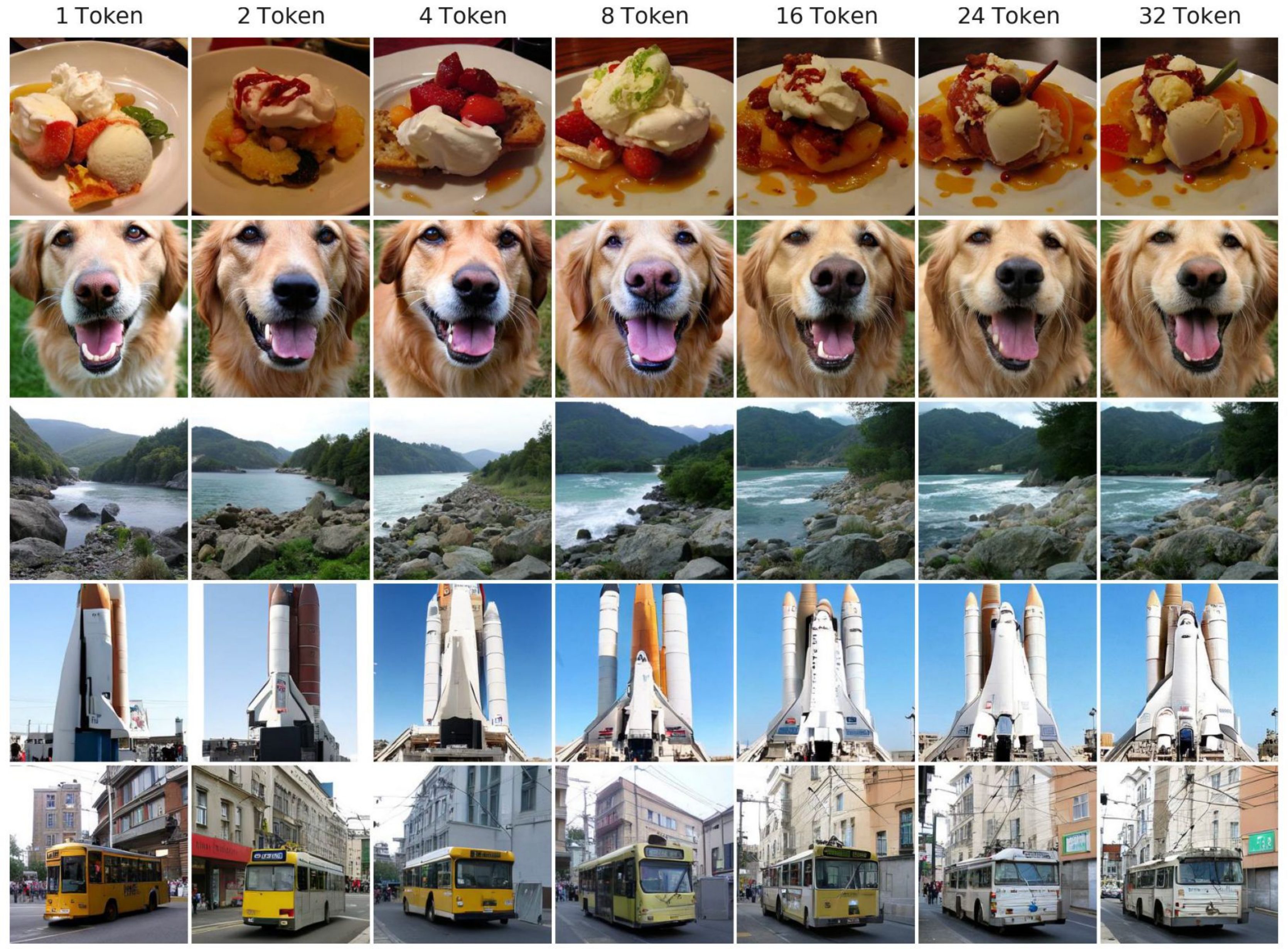




4. AR Generation with A Variable Number of Tokens (w/ LlamaGen)

Code & Models

Project Page



5. Discussion

Potential Applications

- Unified understanding & generation, image compression, data analysis, etc.

Limitations/Future Work

- **Inference Speed:** Diffusion decoding is slower than direct pixel regression.
- **Alternative Architectures:** Flow-matching or consistency models could improve efficiency.
- **Random Ordering:** Compatible to random-ordered tokens could be useful.