VibeToken

Scaling 1D Image Tokenizers and Autoregressive Models for Dynamic Resolution Generations

64 tokens. **Any resolution.** Constant 179 GFLOPs.

Maitreya Patel[1,2,*], Jingtao Li[2], Weiming Zhuang[2], Yezhou Yang[1], Lingjuan Lv[2]

[1] *Arizona State University* [2] *SonyAI*

CVPR JUNE 3-7, 2026 DENVER COLORADO

* Maitreya is now at Adobe

ImageNet-1k samples · 256² – 1024² · variable aspect ratios

63.4×

more efficient than LlamaGen at 1024²

179 G FLOPs vs. 11.3 T FLOPs

THE PROBLEM

AR image models can't follow diffusion to high resolution

Diffusion

Generalizes to arbitrary resolutions natively

✓ FiT, NiT scale to any (H, W)

✓ Constant token count per pass

✓ Production-grade today

Autoregressive

Locked to the resolution it was trained on

✗ Token count grows with pixels

✗ Attention cost is $O(T^2)$

✗ Trained at fixed 256^2 / 512^2

LlamaGen at 1024×1024:

4,096 tokens → ≈11 TFLOPs per forward pass

THE IDEA

Push all the resolution-handling into the tokenizer

> *"Can we encode any-resolution images into a **fixed, small** number of tokens — and decode them back?"*

If yes:

1 **Train AR at fixed length**

L = 64–256 tokens

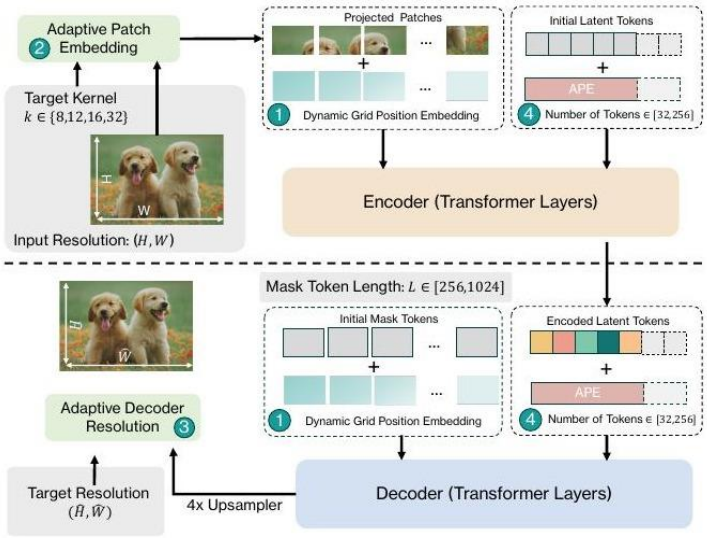
2 **Tokenizer shoulders scaling**

any (H, W) → 64 tokens

3 **Inference cost is constant**

independent of resolution

VibeToken — a resolution-agnostic 1D tokenizer



Encoder → 32–256 latent tokens → decoder at any (H, W)

1 **Dynamic grid positional embedding**

Bilinearly resize a learned 32×32 grid to any patch lattice.

2 **Adaptive patch embedding**

Single base kernel resized on-the-fly to k ∈ {8, 12, 16, 32}.

3 **Adaptive decoder resolution**

Decode 4× upsampled, then a learned conv lands on (H, W).

4 **Dynamic-length tokenization**

Sample L ∈ [32, 256] every step — both encoder & decoder train on it.

Bending the encoder to any input shape

① Dynamic grid positional embedding

Learn a 32×32 grid of position codes; bilinearly resize to match the current patch lattice (T_H, T_W).

Preserves spatial inductive bias across any resolution & aspect ratio — and is cheaper than learnable RoPE.

② Adaptive patch embedding (FlexiViT-style)

One learned base kernel W_kmax; for any k ∈ {8, 12, 16, 32} we derive W_k on the fly by bilinear weight-resizing.

No per-k parameters. Consistent features across patch sizes.

→ The two pieces together cut FLOPs ~3× while improving rFID at high res.

Position embedding & patch ablation

Small VAE setting, 200k iters

Variant	GFLOPs	rFID 256²	rFID 1024²
RoPE	299	2.93	280.7
Learnable RoPE	445	1.54	93.5
Dynamic Grid	299	1.71	131.2
+ Adaptive Patch	**90**	**1.37**	**5.4**

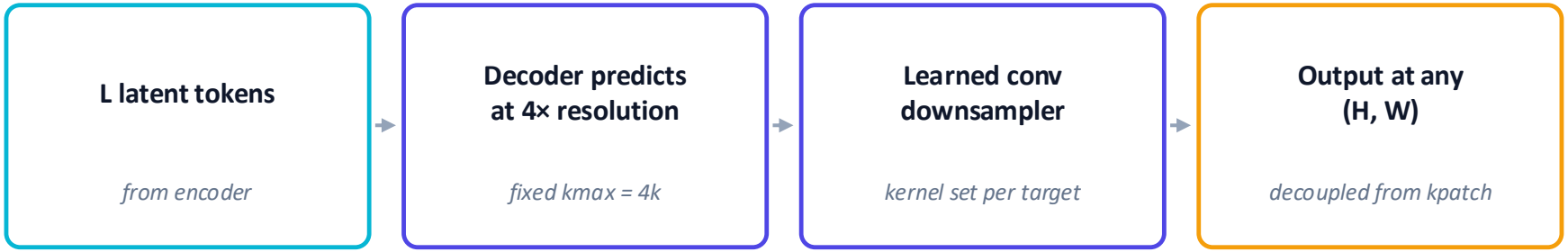
Adaptive patch alone takes 1024² rFID from 131 → 5.4.

Full-scale VibeToken-LL: Grid uses 33% fewer FLOPs than Learnable RoPE at equal or better rFID.

DECODER DESIGN

Decoding to any output shape — and getting SR for free

③ Adaptive decoder resolution



FREE LUNCH

Native 4× super-resolution

Encode a low-res image, swap the downsampler kernel, decode at 4× the size. No separate upsampler. No extra training.

Standard recipe

Train a tokenizer + train a separate SR diffusion model (SDXL / Flux upsampler) on top.

VibeToken ships it built-in.

One model, any token budget

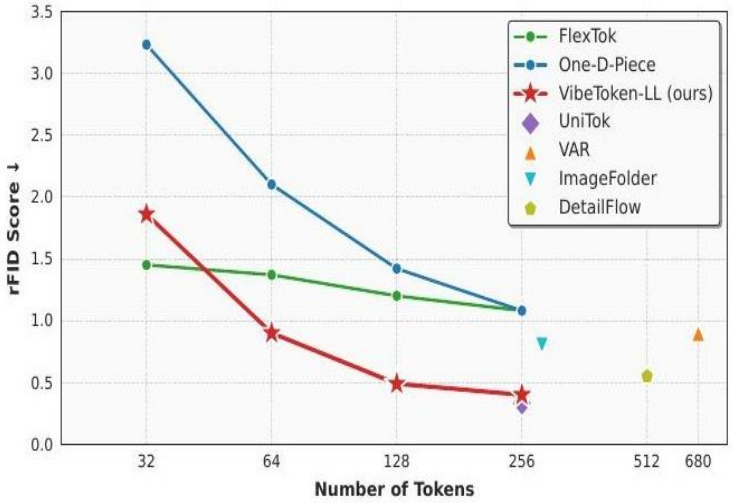
④ Train both encoder & decoder on variable L

Each step: sample L ~ Uniform[32, 256]. Encoder produces exactly L tokens; decoder consumes exactly L — no padding.

Why not tail-token drop?

Methods like One-D-Piece train at fixed max-length and drop tokens at inference — leaving a quality gap on short sequences.

Training natively on every length closes that gap.



rFID vs. latent length L at 256²

VibeToken (red) dominates at every L.

TRAINING STRATEGY

Mix everything during training — resolution, aspect, length

Resolution mix

- 256^2 – 512^2 training only
- Per-sample independent input vs. target resolution
- Generalizes up to 1024^2 unseen

Aspect ratio mix

- {1:1, 1:2, 2:1, 2:3, 3:2}
- 60% square, 40% non-square
- Native support without retraining

Length & quantization

- L ~ Uniform[32, 256] each step
- MVQ with 8 codebooks (vocab 32k)
- Single-stage like TA-TiTok

Total training budget — ImageNet-1k

600 K	8 × H100	300 / 150
tokenizer iterations	*single node*	*AR epochs (B / XXL)*

RECONSTRUCTION · TABLE 2

VibeToken — the only 1D tokenizer with native resolution generalization

Tokenizer	Type	Tokens	rFID 256^2	rFID 512^2	rFID 1024^2	Arbitrary
LlamaGen-Tok	2D · 16× compr.	256 → 4096	2.19	0.70	2.01	2.48
Open-MAGVIT-v2	2D · 16× compr.	256 → 4096	1.17	0.50	1.32	1.52
IBQ	2D · 16× compr.	256 → 4096	0.97	0.40	1.26	1.42
UniTok	1D fixed-res	256	0.33	—	—	—
FlexTok / One-D-Piece	1D dyn.-len.	1–256	1.08	—	—	—
VibeToken-LL (ours)	**1D · dyn-len + dyn-res**	**32–256**	**0.40**	**0.51**	**2.40**	**3.60**

16×–64×

fewer tokens than 2D tokenizers at 1024^2

Only 1D model

with arbitrary resolution + aspect support

0.40 rFID

matches IBQ at 256^2 with 32–256 tokens

A LlamaGen-style AR on top — resolution goes free

Decouples compute from resolution

Inference FLOPs depend only on L and AR depth — not on (H, W).

Trained at 256² – 512² · generalizes to 1024².

179 GFLOPs

per forward pass — at every resolution

LlamaGen scales 711 G → 11.3 T as (H,W) grows

ImageNet-1k samples · 256^2 – 1024^2 · variable aspect ratios

GENERATION · TABLE 3

One AR model — class-conditional gFID across resolutions

Higher resolutions (512^2 – 1024^2) — gFID, lower is better

Model	Type	512^2	512×768	1024×768	1024^2	Average
EDM2-L	Diff.	1.92	5.62	39.54	64.32	23.23
FiTv2-XL	Diff.	2.93	20.61	176.66	259.11	124.32
NiT-XL	Diff.	1.80	4.45	5.14	5.87	6.05
VibeToken-Gen (XXL)	**AR**	**3.69**	**5.32**	**3.84**	**3.54**	**5.53**

Beats NiT at 1024^2

3.54 vs 5.87 gFID

First AR model

to be resolution-generalist

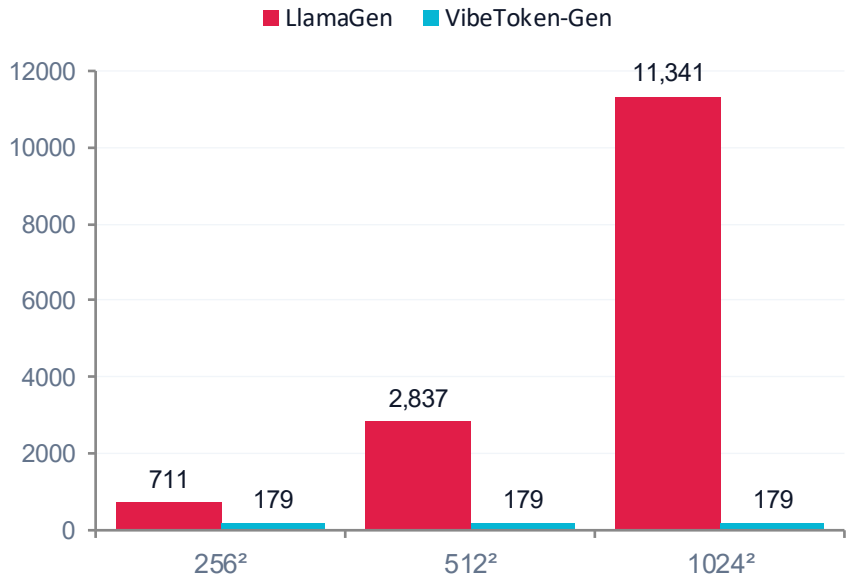
FiT-v2 collapses

off-training resolutions (>259 gFID)

EFFICIENCY

Constant cost — and seconds, not minutes, at 1024²

Inference compute (GFLOPs)



End-to-end wall-clock (s / image, XXL)

Model	256²	1024²
LlamaGen (XXL)	0.20	32.79
VibeToken-Gen (XXL)	**0.46**	**0.46**
NiT-XL (diffusion)	—	1.08

63.4× fewer FLOPs at 1024²

71× faster than LlamaGen at 1024²

2.35× faster than NiT diffusion

TAKEAWAYS

Resolution-agnostic tokens unlock AR for production

- **First 1D tokenizer** to natively support arbitrary resolutions and aspect ratios.
- **AR closes the flexibility gap with diffusion** 3.54 vs 5.87 gFID at 1024^2, 2.35× faster than NiT.
- **Inference cost decoupled from pixels** 179 GFLOPs at every resolution — L controls the budget.
- **Native 4× super-resolution** no separate upsampler, one forward pass.
- **Next:** text-to-image, video, unified multimodal.

Project page

github.com/SonyResearch/VibeToken