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

Supplementary Material

G. Extended Related Works

Image tokenizers. Tokenizers are central to modern image generation. Diffusion/flow models are typically built on continuous latents (*i.e.*, VAEs) and achieve state-of-the-art results, whereas AR models commonly rely on discrete latents (*e.g.*, VQ-VAE variants). Early AR systems use *2D* tokenizers such as MAGViT [17], Open-MAGViT2 [24], and LlamaGen [37]. A known limitation is that the token count grows with resolution—*e.g.*, with stride 16, a 1024×1024 image yields 4096 tokens—making scaling costly. DA-AR reduces tokens via stronger downsampling (e.g., $32\times$ compression $\Rightarrow$ 1024 tokens at 1024^2), and CAT introduces LLM-guided *dynamic* compression for 2D VAEs; however, these encodings remain only partially flexible and still couple token length to spatial resolution. Recent work proposes *1D* tokenizers with higher compression. TiTok [52] introduced the paradigm; TA-TiTok [16] extends training (with/without text conditioning). Subsequent approaches refine objectives, quantization, and training curricula [4, 5, 32]. Despite strong compression, many early 1D designs assume fixed training grids and output a *fixed* latent length, limiting generalization to unseen resolutions/aspect ratios. One-D-Piece and FlexTok [1, 26] enable *dynamic* latent lengths (e.g., 1–256 tokens) but primarily at fixed resolutions, so cross-resolution generalization remains challenging. Instella-T2I [44] and Layton [48] scale 1D tokenizers to 1024^2 using as few as 128–256 tokens, yet the encodings are still tied to specific canvases and do not natively handle arbitrary aspect ratios. A-Token [22] moves toward arbitrary resolutions, but requires careful high-resolution training and remains sensitive to data/compute budgets. In contrast, **VibeToken** scales 1D tokenization to *resolution-agnostic* encodings: it decouples latent length from spatial size, supports dynamic token budgets, and generalizes across resolutions and aspect ratios without retraining per resolution. This flexibility enables the AR generator to operate with a fixed, user-controllable compute budget while retaining strong fidelity.

Dynamic Resolution-adaptation for Transformers. Vision Transformers (ViT) [8] are typically trained on fixed patch grids, making *length generalization* difficult due to absolute positional encodings and quadratic attention costs. A line of work improves flexibility primarily for *discriminative* tasks: ResFormer [40], DeiT [41], ViTAR [12], and NaViT [6] explore position encodings, packed batching, and attention variants to better handle variable input sizes, with ViTAR/NaViT showing notably stronger scaling than their baselines. FlexiViT [2] enables *compute-adaptive* inference via dynamic patch sizes. Beyond absolute PEs, recent methods leverage Axial-RoPE and learned RoPE variants to improve cross-resolution extrapolation [13]. On the *generative* side, Diffusion Transformers (DiTs) [29] have been scaled to native resolutions with FiT-v1/v2 [23, 42] and NiT [43] through RoPE-based encodings and training strategies inspired by NaViT. However, autoregressive (AR) image generation remains largely fixed-resolution: token counts grow with resolutions and next-token training is brittle to length shifts. We bridge this gap by pairing a *resolution-agnostic* 1D tokenizer, **VibeToken**, with a LlamaGen-style AR head, **VibeToken-Gen**. The tokenizer decouples token *length* from spatial resolution, while the generator conditions on the target canvas; together, they enable arbitrary resolutions and aspect ratios under a constant, user-controllable token budget.

H. Experimental Setup

This section details the training and inference setups for **VibeToken** (tokenizer) and **VibeToken-Gen** (AR generator).

H.1. VibeToken

Table F summarizes the hyperparameters for **VibeToken** and its ablations. We follow TiTok/TA-TiTok conventions, adopting TA-TiTok's single-stage training (no text conditioning).

Ablations. We train a small encoder/decoder VAE for 200,000 iterations on $4\times$H100 with per-GPU batch size 32. The latter half of training uses an adversarial loss as well. Ablation models are trained at two resolutions, 256×256 and 512×512.

Final tokenizer. The scaled **VibeToken** uses multi-vector quantization (MVQ) with 8 codebooks and 256 latent dimensions per token, factorized into 8 sub-codes of 32 dimensions each. The maximum latent length is variable in $[32, 256]$; each token comprises 8 sub-tokens *without* increasing the AR sequence length. Concretely, following the RQ-Transformer style prediction-head, the AR embedding/output layers combine the 8 sub-codes within the channel dimension after UniTok, so time length L is unchanged.

Config	Ablations	VibeToken
Model Type	VAE	MVQ
Token size	16	256
# of codebooks	—	8
Vocab size	—	32768
Encoder/Decoder	SS	SL/LL
Discriminator start	100,000	300,000
Quantizer weight	1	1
Discriminator weight	0.1	1
Perceptual weight	1.1	1.1
Reconstruction weight	1	1
Commitment cost	—	0.25
KL weight	1×10^{-6}	—
# of tokens	256	32–256
Optimizer	AdamW	
Learning rate	0.0001	
GAN LR	0.0001	
β_1	0.9	
β_2	0.999	
Weight decay	0.0001	
Scheduler	Cosine	
Warmup steps	10,000	
Dataset	ImageNet1k	ImageNet1k
Batch size / GPU	32	8
Gradient accumulation	1	2
GPUs	4×H100s	8×H100s
Precision	bf16	bf16
Training steps	200,000	600,000
Variable resolutions	`{"256x256": 0.5, "512x512": 0.5}`	`{"256x256": 0.3, "512x512": 0.3, "384x256": 0.1, "256x384": 0.1, "512x384": 0.1, "384x512": 0.1}`

Table F. **VibeToken** pretraining setup.

Training schedule. We train the **SL** and **LL** variants for 600,000 iterations on a single node with 8×H100 GPUs, batch size 8 per GPU, and gradient accumulation 2 (effective batch 128). Training images are sampled over six resolutions/aspect ratios between 256×256 and 512×512 with probabilities listed in Table F. Despite training below 512^2, **VibeToken** generalizes to higher resolutions (1024^2) without sacrificing reconstruction performance.

H.2. VibeToken-Gen

Table G lists pretraining settings for **VibeToken-Gen** **B** (≈ 90M) and **XXL** (≈ 1.4B). Aside from model size, the setups are identical where possible (epochs and GPU count scale with size).

Tokenizer and tokens. We fix the tokenizer to **VibeToken**-`MVQ-LL`. Tokens of length $L \in \{64, 128, 256\}$ are extracted over eight resolutions according to Table G. For simplicity, the encoder patch size is 16×16.

Model and training. Following LlamaGen, we train on ImageNet-1k with ten-crop augmentation. We apply QK LayerNorm for stability. To predict MVQ sub-codes efficiently, we attach a lightweight 4-layer residual transformer head after UniTok that predicts the 8 sub-codes per token; this reduces FLOPs by keeping the temporal length L fixed while factorizing predictions across sub-codes. We observed instability with `bfloat16`; all AR training is therefore conducted in `fp32`.

Config	GPT-B	GPT-XXL
Tokenizer & Data		
Tokenizer	VibeToken-MVQ-LL	
Tokens	64 / 128 / 256	
Resolutions	`{"256x256": 0.3, "512x512": 0.3, "384x256": 0.07, "256x384": 0.07, "512x384": 0.07, "384x512": 0.07, "256x512": 0.06, "512x256": 0.06}`	
Encoder patch size	16x16 (fixed for simplicity)	
Dataset	ImageNet1k	
Model Architecture		
Prediction head layers	4	
# of codebooks	8	
Vocab size	32768	
Sequence length	64 / 128 / 256	
Class-dropout prob.	0.1	
Additional norm layers	QK	
Dropout prob.	0.1	
Training Setup		
Epochs	300	150
Optimizer	AdamW	
Learning rate	0.0001	
β_1	0.9	
β_2	0.95	
Weight decay	0.05	
Precision	None	
Schedule	Linear	
GPUs	4×H100–8×H100	8×H100

Table G. Hyperparameters for *GPT-B* and *GPT-XXL* configurations of **VibeToken-Gen**.

Inference and evaluation. Unless stated otherwise, quantitative results use classifier-free guidance (CFG) fixed across runs with sampling parameters: temperature= 1.0, top-k=0, top-p=1.0. Qualitative samples use CFG= 4.0, temperature= 0.9, top-k=500, top-p=1.0. The decoder patch size is fixed to 16×16 for resolutions ≤ 512×512 and 32×32 for higher resolutions.

I. Ablations

I.1. VibeToken

Image Super-Resolution. Table H evaluates the *native* super-resolution capability of **VibeToken** against diffusion/flow upsamplers (e.g., SDXL/Flux upsamplers). Existing tokenizers generally *lack* native SR and require a separately trained upsampler; **VibeToken** does not. We use 10k FFHQ images at 1024×1024, bicubically downsample to 256×256 and 512×512 for 4× and 2× SR, respectively, and report PSNR/SSIM/LPIPS. Diffusion upsamplers tend to achieve higher PSNR (favoring smoothness), while **VibeToken** yields stronger SSIM and lower LPIPS, indicating better structure and perceptual fidelity. Thus, **VibeToken**

Model	Type	NFEs	Scale	PSNR↑	SSIM↑	LPIPS↓
SDXL-Upscaler	Diffusion	20		**27.97**	0.710	0.317
Flux-Upscaler	Diffusion	20	2×	25.17	0.609	0.377
VibeToken-LL	VQ-VAE	1		24.98	**0.838**	**0.261**
SDXL-Upscaler	Diffusion	20		**29.10**	0.721	0.361
Flux-Upscaler	Diffusion	20	4×	25.88	0.637	0.357
VibeToken-LL	VQ-VAE	1		24.11	**0.805**	**0.310**
VibeToken-LL	VQ-VAE	1	1×	24.91	0.839	0.263

Table H. **Super-resolution ablation.** Ground truth high resolution is fixed to 1024×1024 and bicubic downsampling is used to get low resolution counterparts.

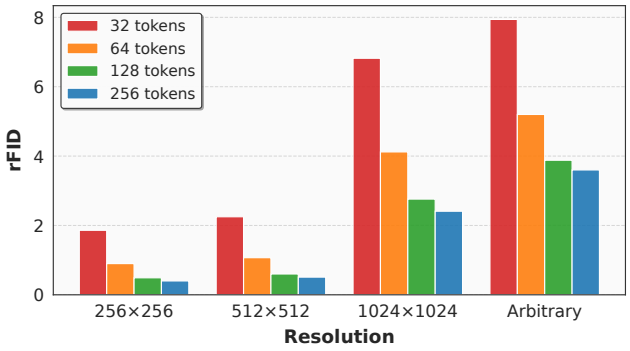

Figure E. **Token-length Analysis.** rFID performance of VibeToken-LL with different token lengths across the resolutions. Notably, reference images are different for each resolutions, hence, rFID scale and performance is not comparable between each resolutions.

delivers competitive SR *without* high-resolution pretraining or an extra upsampler. Qualitative examples are shown in Figure N.

Token Length *vs.* Resolutions. Figure E shows that VibeToken maintains strong reconstruction quality across resolutions while supporting dynamic-length encoding. In practice, **128 tokens** offer an excellent quality–compute balance for reconstruction; **256 tokens** provide small additional gains at higher cost. Notably, for *generation*, **64 tokens** are most effective—later tokens primarily capture high-frequency details that benefit reconstruction more than synthesis. This separation suggests using short sequences for AR generation and longer sequences when reconstruction fidelity is paramount.

Generalization beyond ImageNet. Table J compares the VibeToken-LL performance on MSCOCO for two resolutions along with IBQ and other 1D tokenizers. It can be observed that despite VibeToken only trained on ImageNet (single entity focused) it generalizes to more complex images and achieves SoTA performance on 256×256 resolution and gets competitive performance to 2D tokenizer.

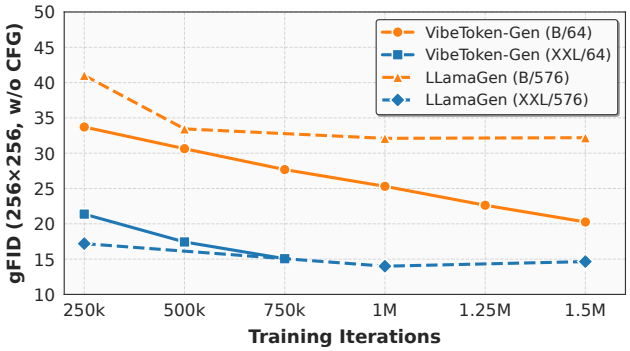

Figure F. **Convergence Analysis.** gFID performance of VibeToken-Gen and baseline LlamaGen models as training progresses. The results are shown on 256×256 without classifier-free guidance.

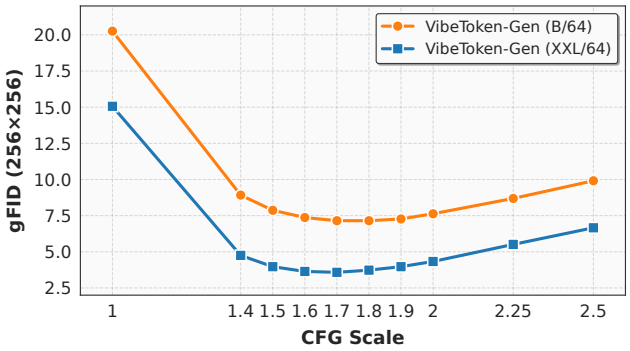

Figure G. **CFG Analysis.** gFID performance of VibeToken-Gen-B/XXL with respect to classifier-free guidance scale on 256×256.

Extended Reconstruction Results. Table I provides the detailed results on ImageNet across reconstruction metrics such as PSNR, SSIM, LPIPS and rFID. VibeToken consistently outperforms the 1D tokenizers and performs competitively to 2D tokenizers. Importantly, 2D tokenizers have 1024 tokens for 512×512 resolutions while VibeToken only requires 256 tokens and can support as low as 64 tokens.

I.2. VibeToken-Gen

Convergence. Figure F compares training dynamics of VibeToken-Gen and LlamaGen at Base and XXL variants, evaluated on 256^2 *without* CFG to expose true likelihood learning. VibeToken-Gen continues improving with training and surpasses LlamaGen, whereas LlamaGen saturates early and shows limited gains thereafter. This indicates better scaling of VibeToken-Gen, attributable to shorter sequences and resolution-agnostic 1D tokens. Due to compute limits, we train VibeToken-Gen-`XXL/64` to 750k iterations (≈150 epochs); based on the `B/64` trend, we expect further improvements with longer training.

Model	Type	ImageNet 256×256				ImageNet 512×512			
		PSNR↑	SSIM↑	LPIPS↓	rFID↓	PSNR↑	SSIM↑	LPIPS↓	rFID↓
MAGVIT-v2	2D	22.62	0.5870	0.2050	1.17	24.67	0.6510	0.2090	0.50
IBQ	2D	22.95	0.5960	0.1980	0.97	**25.19**	0.6620	0.2010	**0.40**
LLaMAGen	2D	21.44	0.5320	0.2240	2.10	23.76	0.6140	0.2180	0.65
TiTok-B	1D	-	-	-	1.70	-	-	-	1.52
VAR	1D	-	-	-	0.90	-	-	-	-
ImageFolder	1D	-	-	-	0.80	-	-	-	-
UniTok	1D	-	-	-	0.33	-	-	-	-
One-D-Piece-L	1D	19.04	-	-	1.08	-	-	-	-
DetailFlow	1D	-	-	-	0.55	-	-	-	-
Instella	1D	-	-	-	-	22.25	0.7040	-	1.32
VibeToken-SL	1D	24.47	0.8069	0.1137	0.43	22.86	0.7541	0.1998	0.55
VibeToken-LL	1D	**25.04**	**0.8194**	**0.1048**	**0.40**	23.37	**0.7649**	**0.1867**	0.51

Table I. ImageNet reconstruction results across the diverse metrics at 256×256 and 512×512. Best is shown in bold and second best is underlined. VibeToken consistently outperforms 1D tokenizers and performs competitively to 2D tokenizer while having higher compression rates.

Model	256×256			512×512		
	FID↓	PSNR↑	SSIM↑	FID↓	PSNR↑	SSIM↑
IBQ (2D)	6.14	22.56	0.603	**4.12**	**24.50**	0.656
TiTok-SL/256	6.84	21.60	0.726	–	–	–
One-D-Piece-LL/256	8.02	18.41	0.612	–	–	–
VibeToken-LL/256 (ours)	**3.59**	**24.71**	**0.828**	4.46	22.68	**0.757**

Table J. Comparison of VibeToken-LL on MSCOCO validation set across 256 × 256 and 512 × 512 resolutions. Best results are in bold, and second-best are underlined.

Model vs. Sec/Img	256×256	1024×1024
Tokenizer enc./dec. only		
LlamaGen-Tok	**0.005**	0.082
VibeToken-LL	0.017	**0.017**
End-to-end incl. tok. dec. + generation		
LlamaGen / XXL	**0.20**	32.79
VibeToken-Gen / XXL	0.46	**0.46**

Table K. Inference speed (seconds per image) comparison across resolutions.

Impact of classifier-free guidance. Figure G shows smooth behavior across CFG scales. After convergence, both `B` and `XXL` models achieve their best gFID around 1.7–1.8. Accordingly, we use CFG= 1.75 and CFG= 2.0 when reporting VibeToken-Gen results.

Efficiency Analysis. Table K shows the comparison of VibeToken and VibeToken-Gen for reconstruction and generation tasks with respect to LlamaGen. It can be observed that LlamaGen (theoretically) requires 32 seconds to generate 1024×1024 resolution images while VibeToken-Gen can do sub-one second generations.

J. Qualitative Results

Figures H, I, J, K, L, and M illustrate samples from six ImageNet1k classes generated by VibeToken-Gen-`(XXL/64)` at random resolutions. VibeToken-Gen consistently produces high-resolution, variable-aspect-ratio images using only 64 tokens. Some crops or truncations may appear; we attribute these artifacts to the randomized cropping used during AR training.

K. Limitations and Future Works

Our experiments focus on ImageNet-1k and class-conditional generation; extending to text-to-image, open-vocabulary settings, and video is an important next step. As a resolution *generalist* model, VibeToken-Gen can trail single-resolution specialists at exactly 256^2/512^2 under the same budget. Further scaling, longer training, and stronger data augmentation may reduce this gap. Methodologically, exploring other AR variants (*e.g.* randomized orderings, scale-wise training), alternative quantization schemes, and larger pretraining corpora could improve quality. Finally, applying resolution-agnostic tokenization to unified mul-

timodal modeling (*e.g.* image–text–video) is a promising direction for significantly efficient production-grade generative systems.

Figure H. **Class 985.** Qualitative examples of generated images using **VibeToken-Gen** (XXL/64) on arbitrary resolutions.

Figure I. **Class 980.** Qualitative examples of generated images using **VibeToken-Gen** (XXL/64) on arbitrary resolutions.

Figure J. **Class 387.** Qualitative examples of generated images using **VibeToken-Gen** (XXL/64) on arbitrary resolutions.

Figure K. **Class 250.** Qualitative examples of generated images using **VibeToken-Gen** (XXL/64) on arbitrary resolutions.

Figure L. **Class 88.** Qualitative examples of generated images using **VibeToken-Gen** (XXL/64) on arbitrary resolutions.

Figure M. **Class 33.** Qualitative examples of generated images using **VibeToken-Gen** (XXL/64) on arbitrary resolutions.

Figure N. **Super-resolution Task.** Qualitative examples of native 4× super-resolution ability of **VibeToken** with respect to simple bilinear resize operation.