Soft Tail-dropping for Adaptive Visual Tokenization

Zeyuan Chen[1⋆], Kai Zhang[2], Zhuowen Tu[1], and Yuanjun Xiong[2]

[1] University of California San Diego, La Jolla, CA, USA
[2] Adobe, San Jose, CA, USA

Project Page: `https://zeyuan-chen.com/STAT/`

Abstract. We present **S**oft **T**ail-dropping **A**daptive **T**okenizer (STAT), a discrete tokenizer that learns adaptive visual representations. STAT adjusts the number of tokens allocated to each image according to its perceptual complexity. Specifically, it encodes an image into discrete tokens together with token-wise keep probabilities indicating whether each token is necessary for faithful reconstruction or can be safely dropped. Through this learned adaptivity, STAT achieves state-of-the-art reconstruction quality while using fewer tokens on average. When integrated with vanilla causal autoregressive (AR) modeling, STAT enables a content-aware generative model with adaptive-length sampling. The model achieves competitive or superior visual generation quality compared with other generative model families while exhibiting favorable scaling behavior that has been elusive in prior vanilla AR visual generation attempts.

Keywords: Adaptive visual tokenization · Autoregressive generation

1 Introduction

Images exhibit variation in visual complexity, ranging from simple structures with smooth homogeneous regions to complex layouts with rich high-frequency textures. Traditional visual compression methods, such as JPEG for images and modern video codecs including H.264 and HEVC, address this variability by allocating bits adaptively according to visual content. A similar principle of content-dependent representation also appears outside the vision domain. For instance, modern text tokenizers [43, 53] produce token sequences whose representation cardinality depends on the input content, serving as a key component behind the success of large language models [7,42,57,59]. In contrast, tokenizers in popular image generation models [4,41,56] typically anchor visual tokens on rigid 2D grids [16,27,46,61,66] or fixed-length 1D sequences [26,69], enforcing a constant compression ratio regardless of the underlying image content. Consequently, simple images may be over-represented while complex ones remain insufficiently encoded, which challenges generative models. Despite recent efforts [2, 39, 48] to

⋆ Work done during an internship at Adobe.

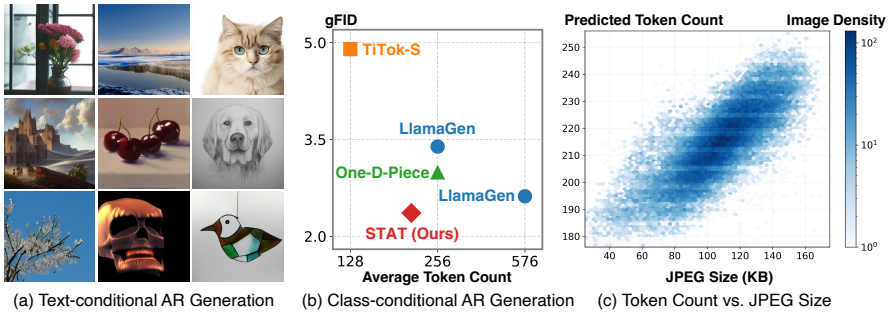

Fig. 1: STAT enables (a) high-quality text-conditional image generation and (b) achieves the best gFID in class-conditional ImageNet generation among visual tokenizers. STAT learns to allocate tokens adaptively based on image complexity, as validated in (c), where predicted token count strongly correlates with JPEG file size.

enable flexible representation cardinality in visual tokenizers, achieving adaptive alignment between token allocation and visual complexity within the tokenizer remains challenging. Moreover, the compatibility of content-adaptive visual tokenizers with generative models, particularly autoregressive (AR) models [5, 42] that naturally support variable length token sequences, remains underexplored.

In this work, we propose STAT, a visual tokenizer that employs a probabilistic token dropping mechanism for adaptive tokenization. Specifically, STAT encodes an image into 1D discrete tokens together with token-wise keep probabilities. In training, Bernoulli sampling is applied to each token to determine whether it is kept or dropped. To learn effective keep probability profiles, we introduce two priors. First, a content-adaptive prior encourages the sum of predicted probabilities to correlate with the perceptual complexity of the image. Second, a decreasing importance prior encourages probabilities to decrease from the head to the tail of the 1D token sequence, forming a soft tail-dropping pattern in which head tokens are more important as they are more likely to be retained for decoding. During inference, tokens whose keep probabilities exceed a fixed threshold are retained, producing visual representations with adaptive token counts determined by the probabilities. STAT achieves state-of-the-art reconstruction quality while using fewer tokens. Fig. 1 (c) further shows a strong correlation between predicted token counts and JPEG sizes.

More importantly, the adaptivity of STAT facilitates generative modeling. When integrated with a vanilla AR model that performs next-token prediction, STAT enables the model to achieve strong generation quality and favorable scaling behavior, comparable to or even surpassing diffusion-based models and other AR approaches that deviate from the simple next-token prediction paradigm.

To summarize, our contributions are:

– We propose STAT, a visual tokenizer that learns content-adaptive representations by adjusting token counts according to image perceptual complexity.

- We introduce a probabilistic token dropping mechanism with two priors that guide the learning of keep probability profiles, enabling STAT to allocate tokens in a content-adaptive manner.
- STAT achieves state-of-the-art reconstruction quality while using fewer tokens on average. When integrated with a vanilla autoregressive model, STAT enables adaptive content generation, reaching performance comparable to or surpassing other state-of-the-art generative models.

2 Related Work

2.1 Image Tokenization

Image tokenization plays an important role in modern visual generative models [4, 29, 49, 56], where images are mapped from the pixel space into a compressed high dimensional latent space, either continuous or discrete, to reduce the prohibitive computational cost of pixel space modeling and obtain a smoother distribution for generative learning. Variational Autoencoders (VAE) [27] are a representative model from continuous tokenizers that encode images into a continuous latent space. In contrast, Vector Quantized Variational Autoencoders (VQ-VAE) [61] introduce discrete visual tokenization. A VQ-VAE consists of an encoder, a quantizer with a codebook at the bottleneck, and a decoder. The quantizer maps continuous latent embeddings to codebook entries through nearest neighbor search, and the resulting discrete embeddings are decoded to reconstruct the image. Subsequent work has explored various improvements to VQ-VAE tokenizers, including architectural designs [6,29,46,66], training objectives [16, 45], and quantization techniques [14, 38, 67, 72].

VQ tokenizers typically preserve a 2D spatial structure, where downsampled 2D feature maps at the bottleneck are discretized into tokens. To decouple the token count from the input dimension, 1D tokenizers such as TiTok [26, 69] concatenate 2D patch embeddings with a set of 1D latent tokens as the model input and distill information from the patch embeddings into these latent tokens, which are subsequently quantized. Flexible tokenizers [2,39,65] further learn ordered 1D token representations using nested dropout [48], enabling images to be represented with a variable number of tokens during training. These approaches still employs predetermined token counts at inference time. In contrast, STAT learns to reconstruct images with content adaptive token counts, representing an image with the number of tokens aligned with its visual complexity.

Among emerging works investigating adaptive tokenizers, CAT [54] is a continuous tokenizer and uses external complexity signals from large language models. ALIT [13] iteratively updates a 1D token sequence to identify the smallest number of tokens for reconstructing a given image, but requires multiple forward passes. KARL [12] learns loss-conditioned halting to approximate minimum image description in a single pass, but relies on an external loss heuristic and primarily focuses on learning compact representations rather than improving autoregressive generation. In contrast, STAT is single-pass and aligns token count prediction with image perceptual complexity via probabilities predicted within

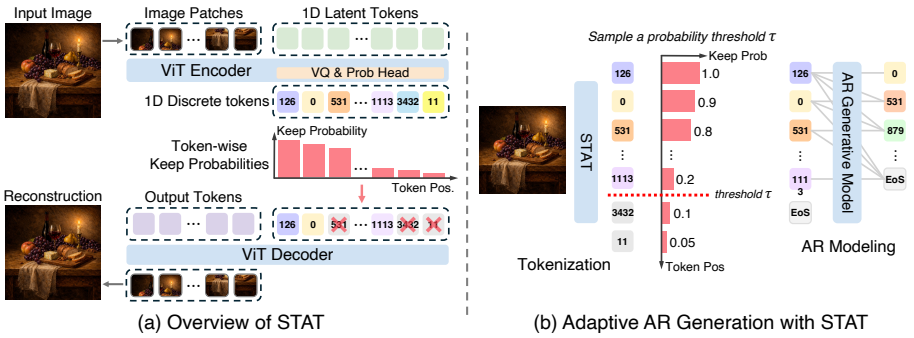

Fig. 2: (a) Overview of STAT. A ViT encoder, followed by a vector-quantization (VQ) layer and a probability prediction head, produces discrete tokens along with token-wise keep probabilities. Probabilistic token dropping generates a masked token sequence, which is then reconstructed using a ViT decoder. **(b) AR generation with STAT.** An End-of-Sequence (EoS) token is determined from the predicted probability profile via a sampled threshold, enabling adaptive-length autoregressive sampling.

the tokenizer. Moreover, STAT is designed for generative modeling and proves particularly suitable for vanilla autoregressive models, achieving compelling generation quality that scales favorably with the size of generative models.

2.2 Image Generation

Generative adversarial networks (GANs) [19] are a representative framework for early image generation. Subsequent works [3, 23–25, 52] have been proposed to advance GAN-based models in terms of both fidelity and diversity. More recently, diffusion models [15, 22, 31, 44, 49, 50, 55, 64, 70] have become a dominant paradigm for image generation due to their high generation quality and stable training dynamics.

With the development of VQ-VAE and other discrete tokenizers, discrete generative models have also gained increasing attention in recent years. These models can generally be divided into two categories. The first category consists of BERT-style models [4, 30, 67], which typically adopt a Transformer encoder architecture with bidirectional attention and generate images by progressively unmasking tokens in a random order. The second category includes GPT-style autoregressive generative models [16, 29, 56, 66], which are based on decoder-only architectures with causal attention. These models are easy to integrate with other modalities such as language. LlamaGen [56] is a representative work in this category and shows that with powerful image tokenizers and proper training recipes, autoregressive models can achieve performance competitive with diffusion models, although they still face scaling bottlenecks in model capacity.

Recently, numerous efforts have explored new autoregressive style pipelines for visual generation. VAR [58] reformulates autoregressive learning for images as next scale prediction to enable a coarse to fine generation process. MAR [32]

introduces a diffusion loss to BERT style generative models, allowing them to operate with continuous tokenizers. RandAR [40] and xAR [47] extend next token prediction with additional variations to strengthen decoder-only autoregressive visual generation. While these approaches achieve strong generation performance, integrating them with large language models requires additional architectural modifications [8, 20, 73].

3 Method

STAT adopts a 1D tokenizer architecture that decouples tokens from 2D spatial patches, providing greater flexibility. During training, it is first optimized for prefix reconstruction by decoding images from prefixes of token sequences with varying lengths. It then learns content adaptivity by predicting token-wise keep probabilities with two priors. Once trained, STAT can be seamlessly integrated with autoregressive (AR) models to enable content-adaptive visual generation.

3.1 Architecture

We build STAT on a 1D Transformer-based [11] tokenizer that operates on latent token sequences decoupled from fixed 2D grids, laying the foundation for variable-length decoding and adaptive token allocation.

Given an image $x \in \mathbb{R}^{B\times C\times H\times W}$, a patch embedding layer maps it to $\mathbf{P} \in \mathbb{R}^{B\times N\times D}$, where $N = H/f \cdot W/f$ for downsampling factor f. We concatenate $\mathbf{P}$ with L learnable latent tokens $\mathbf{L} \in \mathbb{R}^{B\times L\times D}$ and encode the sequence:

$$[z_p, z_l] = \mathrm{Enc}([\mathbf{P}, \mathbf{L}]). \tag{1}$$

Here z_p and z_l denote output patch tokens and latent tokens. We retain only z_l, which is mapped by a vector quantizer to the nearest codebook entries, producing discrete latents z_q. The decoder produces the reconstruction result $\hat{x}$ from z_q together with a set of learnable output tokens $\mathbf{O}$:

$$\hat{x} = \mathrm{Dec}([z_q, \mathbf{O}]). \tag{2}$$

This architecture distills any input image into a 1D token sequence, making STAT naturally compatible with both variable-length prefix reconstruction (Sec. 3.2.1) and content-adaptive token allocation (Sec. 3.2.2).

3.2 Adaptive Visual Tokenizer

Learning adaptivity in STAT proceeds in two stages. In the first stage, we train STAT for variable-length prefix reconstruction, where the model reconstructs images from prefixes of varying lengths while the token count is externally randomized and independent of image content. This prepares the model for the second stage, where STAT learns to allocate tokens adaptively according to perceptual complexity by predicting token-wise keep probabilities.

3.2.1 Variable-length Prefix Reconstruction To prepare for learning adaptivity, we first train the tokenizer to reconstruct images from prefixes of its produced 1D token sequences with variable lengths.

Following prior work [2,39], STAT is trained with hard tail-dropping. In each iteration, a keep length $K \sim \mathcal{U}(L_{\min}, L_{\max})$ is sampled from a uniform distribution, and only the first K quantized tokens are retained for reconstruction:

$$\hat{x} = \mathrm{Dec}([z_q \odot m,\, \mathbf{O}]), \quad m_i = \mathbf{1}[i < K]. \tag{3}$$

Here $m \in [0,1]^L$ is a token selection mask indicating which latent tokens are retained for decoding. This random tail truncation enables STAT to reconstruct images from prefixes of varying lengths, but the number of tokens used for reconstruction is externally randomized and independent of image content.

3.2.2 Content-Adaptive Token Allocation To make token counts adaptive to visual complexity, STAT learns to predict a token-wise keep probability profile for each image and applies probabilistic token dropping accordingly.
Token-wise Keep Probabilities. We design the encoder to output latent tokens $z_l \in \mathbb{R}^{B\times L\times D}$ together with their keep probabilities:

$$p_{j,i} = \sigma\big(g_\theta(z_l[j,i])\big), \tag{4}$$

where j indexes the batch dimension and i indexes the token position. g_θ is a position-aware multilayer perceptron (MLP), and σ denotes the sigmoid function. Each token i for image x_j is stochastically retained via independent Bernoulli sampling $m_{j,i} \sim \mathrm{Bernoulli}(p_{j,i})$. Reconstruction is performed from the masked discrete latent tokens $z_{q,j} \odot m_j$, where $m_j = [m_{j,1}, m_{j,2}, \ldots, m_{j,L}]$ denotes the token selection mask for image x_j. Since Bernoulli sampling is non-differentiable, we adopt a straight-through estimator (STE) to enable gradient flow in training.

We can compute the expected number of tokens used for image x_j by:

$$T_j = \sum_{i=0}^{L-1} p_{j,i}. \tag{5}$$

Thus, the decoding token count in training is decided by the predicted keep probabilities instead of an externally randomized value as in the first stage. We next describe how probabilities are supervised for adaptive token allocation.
Content-adaptive prior. To allocate tokens adaptively across images, we impose a prior that encourages the expected token count to increase with perceptual complexity. For each image x_j, we compute a perceptual reconstruction error $L_{\mathrm{perc},j}$ (e.g., LPIPS) as a proxy for complexity, and correlate this quantity with the expected token count T_j. Across all samples in a batch, we optimize:

$$\mathcal{L}_{\mathrm{content}} = \big(1 - \mathrm{corr}(L_{\mathrm{perc}}, T)\big)^2, \tag{6}$$

which encourages a strong positive correlation between the complexity proxy L_{perc} and the allocated token count T. $\mathcal{L}_{\mathrm{content}}$ acts as a content-adaptive prior

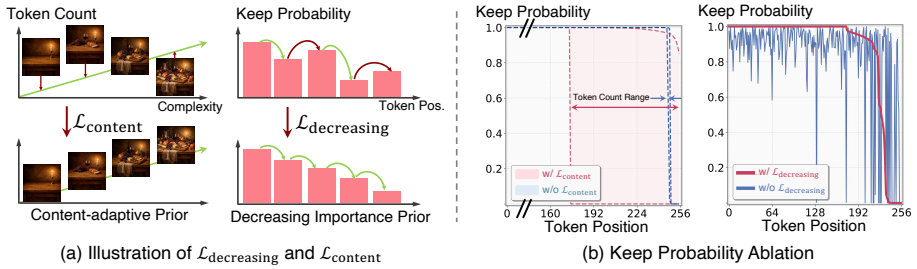

Fig. 3: Illustration and ablation of the priors. (a) presents the illustration of the content-aware prior and the decreasing importance prior. (b) shows ablation studies on predicted token probabilities with/without these two priors.

on token counts, nudging the tokenizer to assign more tokens to harder images and fewer to simpler ones. By contrasting all samples in one batch, this loss encourages the model to properly allocate tokens without the need to normalize the complexity proxy precisely. In practice, we compute the correlation using the Pearson correlation coefficient. An illustration of the prior can be found in Fig. 3.

Decreasing Importance Prior. The token-wise keep probability reflects token importance, as it determines the likelihood that a token is retained for decoding. Tokens with higher probabilities therefore contribute more to reconstruction and are considered more important. To shape this importance profile, we introduce the decreasing importance prior encouraging a monotonically decreasing pattern by penalizing upward steps (see Fig. 3):

$$\mathcal{L}_{\text{decrease}} = \sum_{i=1}^{L-1} \max\left(0,\, p_{j,i} - p_{j,i-1}\right). \tag{7}$$

This prior encourages earlier tokens to carry higher importance than later ones. Such a decreasing importance profile aligns with the causal structure of AR models, where earlier tokens serve as context for subsequent predictions, thereby improving compatibility between STAT and AR generative models. Moreover, it induces a dropping behavior resembling the tail-dropping pattern used in the first training stage, which motivates the name "soft tail-dropping".

Token Budget Regularization. The above priors shape the relative allocation of tokens across images but do not constrain the global token budget. For each image x_j, we compute the average keep probability across its tokens:

$$\bar{p}_j = \frac{1}{L} \sum_{i=0}^{L-1} p_{j,i}. \tag{8}$$

We regularize this quantity toward a prior sparsity level $p^* = 0.5$ using a KL divergence between Bernoulli distributions:

$$\mathcal{L}_{\text{sparse}} = \text{KL}(\text{Bern}(p^*) \parallel \text{Bern}(\bar{p}_j)). \tag{9}$$

This term encourages the average keep probability to approach p^*, thereby controlling the global token budget. In practice, p^* serves as a soft target: reconstruction favors keeping more tokens, while this regularization promotes token efficiency, and their balance determines the final token usage.

Overall Objective. The adaptive tokenizer is trained with the composite loss:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{GAN}} + \mathcal{L}_{\text{VQ}} + \lambda_{\text{content}}\mathcal{L}_{\text{content}} \\ +\lambda_{\text{decrease}}\mathcal{L}_{\text{decrease}} + \lambda_{\text{sparse}}\mathcal{L}_{\text{sparse}}.\end{aligned} \tag{10}$$

This yields STAT, whose representations adapt to visual complexity and exhibit a decreasing importance profile compatible with AR modeling. Please see the supplementary material for hyperparameters and training details.

Inference. At inference time for reconstruction, tokens with keep probabilities $p > 0.5$ are retained for decoding, while the remaining tokens are dropped.

3.3 Visual Generation with Adaptive Tokenizer

The primary purpose of visual tokenizers is to provide an effective representation space for generative models. To evaluate this capability, we integrate STAT with a vanilla AR model and demonstrate that it enables adaptive visual generation.

Tokenization. Given an image x, STAT produces a discrete token sequence $q = (q_0, \ldots, q_{L-1})$ together with a keep probability profile $p = (p_0, \ldots, p_{L-1})$, which is nearly monotonic via the prior $\mathcal{L}_{\text{decrease}}$. We leverage this probability profile to enable adaptive generation by introducing a special end-of-sequence (EoS) token. The EoS position is defined as:

$$t_{\text{eos}} = \min\{i \mid p_i < \tau\}, \tag{11}$$

where τ is a probability threshold randomly sampled at each iteration to allow variable EoS positions. We replace the token at position t_{eos} with the EoS token and discard all subsequent tokens, as illustrated in Fig. 2.

Autoregressive Modeling. STAT converts an image into a discrete sequence terminated by an end-of-sequence (EoS) token, $q = (q_0, \ldots, q_{t_{\text{eos}}-1}, \text{EoS})$, where t_{eos} denotes the EoS position. The AR model is trained to model the distribution over this sequence as:

$$p_\phi(q) = \prod_{t=0}^{t_{\text{eos}}} p_\phi(q_t \mid q_{<t}), \tag{12}$$

where $q_{t_{\text{eos}}} = \text{EoS}$, and $p_\phi(\cdot)$ denotes the probability distribution parameterized by the AR model with learnable parameters ϕ. Please refer to the supplementary material for details in tokenization and autoregressive model training.

4 Experiments

In experiments, we first study the reconstruction performance of STAT. We then evaluate STAT for image generation by integrating it with a vanilla autoregressive (AR) model. Next, we conduct ablation studies to analyze the impact of our

Fig. 4: Keep-probability curves predicted by STAT and the resulting token counts for images of varying complexity. Tokens with $p > 0.5$ are retained. More complex images receive higher token counts, which strongly correlate with their JPEG file sizes.

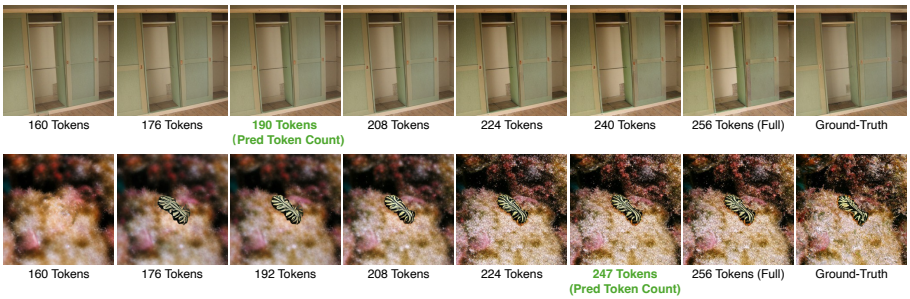

Fig. 5: Adaptive image reconstruction by STAT. Reconstructions with increasing token budgets are shown from left to right. We observe that the predicted token count (green) roughly aligns with the minimum tokens required for faithful reconstruction.

design components. Finally, we investigate the scalability of STAT along three axes: task, resolution, and input domain.

Training setup. By default, STAT is trained on the ImageNet [9] training set at a resolution of 256×256. The patch-embedding downsampling factor is set to $f = 16$, and the vector quantizer uses a codebook of size 4096 with a code dimension of 12. The latent token sequence length is 256. Full training details and hyperparameter settings are provided in the supplementary material.

Metrics. We report reconstruction FID (rFID) [21] and PSNR for image reconstruction. We evaluate generation FID (gFID) [21], Inception Score (IS) [51], and Precision/Recall [28] for image generation. For video reconstruction, we measure reconstruction FVD [60] (rFVD), PSNR, and LPIPS [71] to assess the quality.

4.1 Image Reconstruction

In Fig. 4, we show the keep-probability profiles learned by STAT, which allocate more tokens to visually complex images. Fig. 5 shows reconstructions with increasing token budgets, where the predicted token count corresponds to the onset of a quality plateau and additional tokens yield marginal improvements.

Quantitative evaluation is conducted on ImageNet [9] and COCO [34]. We compare STAT with representative 2D and 1D tokenizers. In Tab. 1, STAT achieves the best rFID on ImageNet while using fewer average tokens than other tokenizers, demonstrating the effectiveness of adaptive token allocation. Moreover, STAT allows flexible adjustment of the token budget at inference, either by

Table 1: ImageNet image reconstruction at 256×256 resolution. We compare STAT with representative 2D [4, 16, 36, 56] and 1D tokenizers [2, 39, 69]. STAT achieves the best rFID while using fewer tokens then other tokenizers. We further evaluate different inference-time thresholding (Thr 0.01) and manually increasing the token budget (+X Tokens), which improves rFID to 0.88 with an average of 240 tokens. CB denotes the codebook size.

Type	Tokenizer	#Tokens	CB	rFID↓	PSNR↑
2D	Taming VQ-GAN [16]	256	16384	4.98	19.40
	MaskGIT VQ-GAN [4]	256	1024	2.28	-
	Open-MAGVIT2 [36]	256	262144	1.17	22.64
	LlamaGen [56]	256	4096	3.02	19.99
	LlamaGen [56]	256	16384	2.19	20.79
1D	TiTok-L [69]	32	4096	2.21	15.96
	TiTok-B [69]	64	4096	1.70	17.13
	TiTok-S [69]	128	4096	1.71	17.80
	FlexTok [2]	256	64000	1.08	17.70
	One-D-Piece-L [39]	256	4096	1.08	19.04
1D	STAT	220	4096	1.15	20.22
	STAT(Thr 0.01)	230	4096	0.99	20.25
	STAT(Thr 0.01 +10)	240	4096	0.88	20.02

Table 2: COCO image reconstruction at 256×256 resolution. STAT achieves the lowest rFID with competitive PSNR, demonstrating strong generalization.

Tokenizer	#Tokens	CB	rFID↓	PSNR↑
Taming VQ-GAN [16]	256	16384	19.29	19.57
LlamaGen [56]	256	16384	8.11	20.42
TiTok-S [69]	128	4096	9.22	17.27
One-D-Piece-L [39]	256	4096	8.04	17.94
STAT	221	4096	6.93	19.80

Table 3: ImageNet class-conditional generation across tokenizers. We train an autoregressive image generative model (LlamaGen-XL, 775M) under the same settings. The model trained with STAT significantly outperforms other tokenizers.

Tokenizer	#Tokens	CB	gFID↓	IS↑	Pre.↑	Rec.↑
LlamaGen [56]	256	16384	3.39	227.1	0.81	0.54
LlamaGen [56]	576	16384	2.62	244.1	0.80	0.57
TiTok-S [69]	128	4096	4.90	191.7	0.77	0.56
One-D-Piece-L [39]	256	4096	2.99	235.1	0.81	0.59
STAT	223	4096	2.36	244.0	0.78	0.62

lowering the threshold for keeping tokens or manually adding tokens, which further improves rFID with a slight decrease in PSNR, indicating a trade-off where additional tokens encourage the tokenizer to prioritize perceptual realism over pixel-level fidelity. We further evaluate generalization on COCO, where STAT still achieves the best rFID as shown in Tab. 2).

4.2 Image Generation

We investigate whether the strong reconstruction capacity of STAT translates to improved generation. We integrate STAT into a vanilla AR image generation framework using the same backbone as LlamaGen [56]. An End-of-Sequence (EoS) token is learned to adaptively determine the effective token length during generation. As summarized in Tab. 4, the vanilla AR model equipped with STAT achieves generation quality comparable to state-of-the-art AR and diffusion-based models, despite its minimalist design. This demonstrates that STAT is well suited for AR generative modeling, improving both performance and efficiency without requiring elaborate architectural heuristics. We further observe strong scaling behavior when increasing the model size from 775M to 3B parameters. While LlamaGen-STAT consistently improves with model scaling, LlamaGen with 256 tokens largely converges at 1.4B and shows minimal gains when scaled to 3B. These results confirm the effectiveness of STAT and the learned adaptive AR generation scheme. Qualitative examples of adaptive generation are in Fig. 6, where the model adjusts EoS positions according to the generated content.

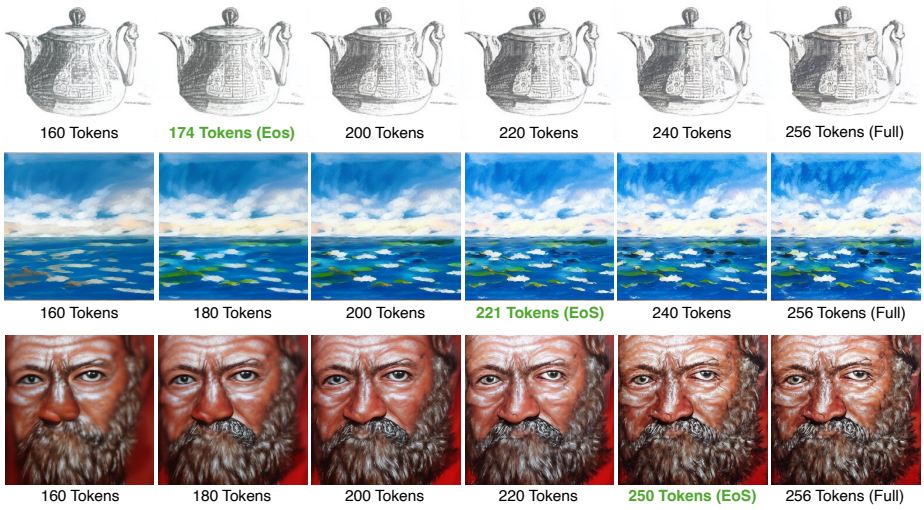

Fig. 6: Adaptive-length AR image generation with STAT. Images generated with increasing token budgets are shown from left to right. Right images share prefix tokens with those to the left, revealing how generation evolves as more tokens are produced. The model predicts the EoS token to adaptively terminate generation (green).

In Tab. 3, we compare different visual tokenizers within the same AR generation framework, including the improved VQ tokenizer from LlamaGen, TiTok-S, and One-D-Piece-L. STAT delivers the best generation performance, surpassing all baseline tokenizers by a substantial margin.

4.3 Ablation Study

We conduct ablation studies on the key design components of STAT to understand the sources of its strong performance in adaptive generation.

Priors. We analyze the impact of the content-adaptive prior and the decreasing importance prior during training. As shown in Fig. 3, removing the content-adaptive prior causes the tokenizer to converge to a trivial solution in which nearly all images use the same token count. The decreasing importance prior encourages a soft tail-dropping pattern with smoothly decreasing probabilities, aligning with the causal ordering of token importance in AR generation.

Adaptive vs. Fixed Token Allocation. To assess whether the adaptivity learned by STAT benefits generation, we compare STAT with a controlled tokenizer variant that uses a fixed token count for all images under the same token budget. As shown in Tab. 5, the fixed-token variant achieves similar reconstruction quality due to its simpler objective of learning a fixed-length representation. However, adaptive tokenization leads to substantially better generation performance, reducing gFID from 2.73 to 2.36. This suggests that allowing the model to adjust token length according to the generated content improves token efficiency and enables higher-quality generation.

Table 4: ImageNet class-conditional image generation. We compare autoregressive models built on STAT with a broad range of generative approaches, including GANs [3, 23, 52], diffusion models [10, 37, 41, 49], and AR models with bidirectional [4, 32, 67, 69] or causal attention [16, 29, 35, 36, 40, 56, 58, 68]. All results are evaluated on ImageNet-1k at 256×256 resolution. Replacing the original tokenizer in LlamaGen with STAT improves the 3B model to a gFID of 1.75, outperforming the LlamaGen baseline by 0.43 while using less than half the number of tokens.

Type	Model	#Para.	gFID↓	IS ↑	Precision ↑	Recall ↑	#Tokens
GAN	BigGAN [3]	112M	6.95	224.5	0.89	0.38	1
	GigaGAN [23]	569M	3.45	225.5	0.84	0.61	1
	StyleGAN-XL [52]	166M	2.30	265.1	0.78	0.53	1
Diffusion	ADM [10]	554M	4.59	186.7	0.82	0.53	250
	LDM-4 [49]	400M	3.60	247.7	–	–	250
	DiT-XL [41]	675M	2.27	278.2	0.83	0.57	250
	SiT-XL [37]	675M	2.06	270.3	0.82	0.59	250
Bi-directional AR	MaskGIT-re [4]	227M	4.02	355.6	–	–	256
	MAGVIT-v2 [67]	307M	1.78	319.4	–	–	256
	MAR-L [32]	479M	1.98	290.3	–	–	64
	MAR-H [32]	943M	1.55	303.7	0.81	0.62	256
	TiTok-S-128 [69]	287M	1.97	281.8	–	–	128
Advanced Causal AR	VAR [58]	600M	2.57	302.6	0.83	0.56	680
	VAR [58]	2.0B	1.92	350.2	0.82	0.59	680
	SAR-XL [35]	893M	2.76	273.8	0.84	0.55	256
	RAR-XXL [68]	955M	1.50	306.9	0.80	0.62	256
	RAR-XL [68]	1.5B	1.48	326.0	0.80	0.63	256
	RandAR-XL [40]	775M	2.25	317.8	0.80	0.60	256
	RandAR-XXL [40]	1.4B	2.15	322.0	0.79	0.62	256
Vanilla Causal AR	VQGAN [16]	1.4B	5.20	280.3	–	–	256
	RQTran.-re [29]	3.8B	3.80	323.7	–	–	256
	Open-MAGVIT-v2-XL [36]	1.5B	2.33	271.8	0.84	0.54	256
	LlamaGen-XL-256 [56]	775M	3.39	227.1	0.81	0.54	256
	LlamaGen-XXL-256 [56]	1.4B	3.09	253.6	0.82	0.53	576
	LlamaGen-3B-256 [56]	3.1B	3.06	279.7	0.84	0.54	576
	LlamaGen-XL-576 [56]	775M	2.62	244.1	0.80	0.57	576
	LlamaGen-XXL-576 [56]	1.4B	2.34	253.9	0.80	0.59	576
	LlamaGen-3B-576 [56]	3.1B	2.18	263.3	0.81	0.58	576
Vanilla Causal AR	LlamaGen-STAT-XL	775M	2.36	244.0	0.78	0.62	223
	LlamaGen-STAT-XXL	1.4B	1.91	290.2	0.79	0.63	227
	LlamaGen-STAT-3B	3.1B	1.75	300.6	0.80	0.64	229

Soft vs. Hard Tail-Dropping. We next investigate whether the mechanism used by STAT to achieve adaptivity contributes to improved generation performance. During training, STAT performs token-wise dropping according to the predicted keep probabilities. To better understand the effect of this design, we implement a hard tail-dropping variant. This variant remains adaptive, with the token count determined by the sum of predicted keep probabilities across token positions, but tokens are deterministically truncated from the tail during training. As shown in Tab. 5, the hard tail-dropping variant achieves similar reconstruction quality, since tail truncation results in a simpler distribution of masked tokens for decoding. However, for image generation, our Soft Tail-Dropping significantly outperforms this variant (gFID 2.36 vs. 2.67). This indicates that the stochastic perturbation introduced by probabilistic token-wise dropping im-

Table 5: Ablation study of STAT. We compare various dropping settings in training and prove that the soft tail-dropping in STAT achieves the best generation result with strong reconstruction quality.

	Reconstruction			Generation				
Dropping Method	#Token	rFID↓	PSNR↑	#Token	gFID↓	IS↑	Pre.↑	Rec.↑
Fixed Token Allocation	220	1.15	20.35	220	2.73	234.2	0.77	0.62
Hard Tail-Dropping	222	1.15	20.30	223	2.67	231.2	0.76	0.63
Fixed EoS Threshold	220	1.15	20.22	221	2.49	240.7	0.79	0.61
Soft Tail-Dropping	220	1.15	20.22	223	2.36	244.0	0.78	0.62

Table 6: Text-conditional generation on GenEval. STAT improves over the LlamaGen [56] tokenizer across all metrics.

Tokenizer	S. Obj.	T. Obj.	Count.	Colors	Position	C. Attri.	Overall
LlamaGen [56]	0.91	0.18	0.19	0.57	0.04	0.01	0.32
STAT	0.99	0.36	0.40	0.81	0.10	0.06	0.45

Table 7: ImageNet image reconstruction at 512×512 resolution. We evaluate the scalability of STAT at a higher resolution and compare it with representative image tokenizers [4, 56, 69]. Despite using significantly fewer tokens, STAT achieves competitive reconstruction quality. CB denotes the codebook size.

Tokenizer	#Tokens	CB	rFID↓	PSNR↑
MaskGIT VQ-GAN [4]	1024	1024	1.97	-
TiTok-L-64 [69]	64	4096	1.77	-
TiTok-B-128 [69]	128	4096	1.52	-
LlamaGen [56]	1024	16384	0.70	23.03
STAT(Thr 0.5)	471	4096	0.87	20.98
STAT(Thr 0.01)	482	4096	0.82	20.98
STAT(Thr 0.01 +20)	501	4096	0.75	20.88

proves the robustness of our tokenizer to imperfect token sequences produced during autoregressive generation.

Variable vs. Fixed EoS Threshold. Adaptive visual generation enabled by STAT introduces an EoS token whose position is determined by a sampled threshold τ. As a result, the same image may have different EoS positions across training iterations due to different sampled τ values. To evaluate whether this variability benefits generation, we compare it with a variant that uses the same tokenizer and a fixed threshold $\tau = 0.5$ for EoS during generation training. As shown in Tab. 5, fixing the EoS threshold leads to worse generation quality (gFID increases from 2.36 to 2.49). This result suggests that varying the EoS position provide an additional form of data augmentation for generation.

4.4 Scalability of STAT

Beyond evaluation on ImageNet and COCO at 256×256, we further study the scalability of STAT along three axes: task, resolution, and input domain.

Task: Text-conditional Image Generation. We evaluate text-conditional image generation by comparing STAT with the VQ tokenizer from LlamaGen [56] on the GenEval [18] benchmark. STAT consistently outperforms the LlamaGen tokenizer across all metrics. Example generations are shown in Fig. 7.

Resolution: Reconstruction at 512×512. We present the results in Tab. 7. STAT achieves competitive reconstruction quality compared with LlamaGen [56], despite using less than half the number of tokens on average.

Input Domain: Video Reconstruction. We extend STAT to the video domain by applying it to 16-frame video clips. As shown in Tab. 8, STAT-Video achieves state-of-the-art reconstruction quality, obtaining the best PSNR and LPIPS among all compared video tokenizers while maintaining competitive rFVD. Moreover, its adaptive tokenization allows STAT-Video to further improve reconstruction quality by increasing the token budget similar to image reconstruction.

Fig. 7: Text-conditional Image Generation. Text prompts are below the images.

Table 8: UCF-101 video reconstruction at 128×128 resolution. We compare STAT-Video with prior video tokenizers [1,17,33,62,63,65]. All methods are evaluated on 16-frame clips following [33]. STAT-Video achieves the best PSNR and LPIPS with competitive rFVD. "(+X Tokens)" denotes manually adding X more token budget during inference, and † indicates models trained for twice as many epochs.

Tokenizer	#Tokens	Codebook Size	rFVD ↓	PSNR ↑	LPIPS ↓
TATS [17]	1024	16384	157	24.22	0.206
OmniTokenizer [63]	1280	8192	62	27.76	0.112
ElasticTok [65]	1022	64000	230	-	-
Cosmos-Tokenizer-DV [1]	1280	64000	140	-	-
LARP [62]	1024	8192	24	-	-
LARP† [62]	1024	8192	20	28.71	0.075
AdapTok [33]	1024	8192	36	25.72	0.114
STAT-Video	997	8192	29	28.77	0.064
STAT-Video (+100 Tokens)	1097	8192	26	29.08	0.062
STAT-Video (+200 Tokens)	1197	8192	23	29.27	0.060
STAT-Video (+300 Tokens)	1297	8192	22	29.40	0.059

5 Conclusion

We present STAT, a content-adaptive 1D visual tokenizer with probabilistic token dropping that aligns token length with image perceptual complexity. STAT matches or surpasses state-of-the-art tokenizers in both reconstruction and generation while using fewer tokens. Our results suggest that such an adaptive tokenizer provides a strong inductive bias for high-dimensional generative modeling and may help enable unified multimodal autoregressive systems.

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
2. Bachmann, R., Allardice, J., Mizrahi, D., Fini, E., Kar, O.F., Amirloo, E., El-Nouby, A., Zamir, A., Dehghan, A.: Flextok: Resampling images into 1d token sequences of flexible length. In: Forty-second International Conference on Machine Learning (2025)
3. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: ICLR (2019)
4. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: CVPR (2022)
5. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: ICML (2020)
6. Chen, Z., Xu, H., Song, G., Xie, Y., Zhang, C., Chen, X., Wang, C., Chang, D., Luo, L.: X-dancer: Expressive music to human dance video generation. In: ICCV (2025)
7. Chen, Z., Zhang, X., Xu, H., Xie, J., Tu, Z.: Cvp: Central-peripheral vision-inspired multimodal model for spatial reasoning. In: WACV (2026)
8. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
10. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS (2021)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
12. Duggal, S., Byun, S., Freeman, W.T., Torralba, A., Isola, P.: Single-pass adaptive image tokenization for minimum program search. In: NeurIPS (2025)
13. Duggal, S., Isola, P., Torralba, A., Freeman, W.T.: Adaptive length image tokenization via recurrent allocation. In: First Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models (2024)
14. El-Nouby, A., Muckley, M.J., Ullrich, K., Laptev, I., Verbeek, J., Jégou, H.: Image compression with product quantized masked image modeling. TMLR (2023)
15. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
16. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR (2021)
17. Ge, S., Hayes, T., Yang, H., Yin, X., Pang, G., Jacobs, D., Huang, J.B., Parikh, D.: Long video generation with time-agnostic vqgan and time-sensitive transformer. In: ECCV (2022)
18. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. NeurIPS (2023)
19. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. NeurIPS (2014)

20. Han, J., Liu, J., Jiang, Y., Yan, B., Zhang, Y., Yuan, Z., Peng, B., Liu, X.: Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. arXiv preprint arXiv:2505.14683 (2024)
21. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS (2017)
22. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. arXiv preprint arxiv:2006.11239 (2020)
23. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: CVPR (2023)
24. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
25. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: CVPR (2020)
26. Kim, D., He, J., Yu, Q., Yang, C., Shen, X., Kwak, S., Chen, L.C.: Democratizing text-to-image masked generative models with compact text-aware one-dimensional tokens. In: ICCV (2025)
27. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
28. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. NeurIPS (2019)
29. Lee, D., Kim, C., Kim, S., Cho, M., Han, W.S.: Autoregressive image generation using residual quantization. In: CVPR (2022)
30. Lezama, J., Chang, H., Jiang, L., Essa, I.: Improved masked image generation with token-critic. In: ECCV (2022)
31. Li, B., Wang, C.Y., Xu, H., Zhang, X., Armand, E., Srivastava, D., Xiaojun, S., Chen, Z., Xie, J., Tu, Z.: Overlaybench: A benchmark for layout-to-image generation with dense overlaps. NeurIPS (2026)
32. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. In: NeurIPS (2024)
33. Li, Y., Tian, C., Xia, R., Liao, N., Guo, W., Yan, J., Li, H., Dai, J., Li, H., Yang, X.: Learning adaptive and temporally causal video tokenization in a 1d latent space. arXiv preprint arXiv:2505.17011 (2025)
34. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
35. Liu, W., Zhuo, L., Xin, Y., Xia, S., Gao, P., Yue, X.: Customize your visual autoregressive recipe with set autoregressive modeling. arXiv preprint arXiv:2410.10511 (2024)
36. Luo, Z., Shi, F., Ge, Y., Yang, Y., Wang, L., Shan, Y.: Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. arXiv preprint arXiv:2409.04410 (2024)
37. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: SIT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: ECCV (2024)
38. Mentzer, F., Minnen, D., Agustsson, E., Tschannen, M.: Finite scalar quantization: Vq-vae made simple. In: ICLR (2024)
39. Miwa, K., Sasaki, K., Arai, H., Takahashi, T., Yamaguchi, Y.: One-d-piece: Image tokenizer meets quality-controllable compression. arXiv preprint arXiv:2501.10064 (2025)

40. Pang, Z., Zhang, T., Luan, F., Man, Y., Tan, H., Zhang, K., Freeman, W.T., Wang, Y.X.: Randar: Decoder-only autoregressive visual generation in random orders. In: CVPR (2025)
41. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
42. Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training (2018), https://api.semanticscholar.org/CorpusID:49313245
43. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog (2019)
44. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
45. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML (2021)
46. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. NeurIPS (2019)
47. Ren, S., Yu, Q., He, J., Shen, X., Yuille, A., Chen, L.C.: Beyond next-token: Next-x prediction for autoregressive visual generation. In: ICCV (2025)
48. Rippel, O., Gelbart, M., Adams, R.: Learning ordered representations with nested dropout. In: ICML (2014)
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
50. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS (2022)
51. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. NeurIPS (2016)
52. Sauer, A., Schwarz, K., Geiger, A.: Stylegan-xl: Scaling stylegan to large diverse datasets. In: SIGGRAPH (2022)
53. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: ACL (2016)
54. Shen, J., Tirumala, K., Yasunaga, M., Misra, I., Zettlemoyer, L., Yu, L., Zhou, C.: Cat: Content-adaptive image tokenization. In: NeurIPS (2025)
55. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
56. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024)
57. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
58. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. In: NeurIPS (2024)
59. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
60. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
61. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. NeurIPS (2017)

62. Wang, H., Suri, S., Ren, Y., Chen, H., Shrivastava, A.: Larp: Tokenizing videos with a learned autoregressive generative prior. arXiv preprint arXiv:2410.21264 (2024)
63. Wang, J., Jiang, Y., Yuan, Z., Peng, B., Wu, Z., Jiang, Y.G.: Omnitokenizer: A joint image-video tokenizer for visual generation. NeurIPS (2024)
64. Wang, Y., Xu, H., Zhang, X., Chen, Z., Sha, Z., Wang, Z., Tu, Z.: Omnicontrolnet: Dual-stage integration for conditional image generation. In: CVPRW (2024)
65. Yan, W., Mnih, V., Faust, A., Zaharia, M., Abbeel, P., Liu, H.: Elastictok: Adaptive tokenization for image and video. ICLR (2025)
66. Yu, J., Li, X., Koh, J.Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., Wu, Y.: Vector-quantized image modeling with improved vqgan. arXiv preprint arXiv:2110.04627 (2021)
67. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Birodkar, V., Gupta, A., Gu, X., Hauptmann, A.G., Gong, B., Yang, M.H., Essa, I., Ross, D.A., Jiang, L.: Language model beats diffusion–tokenizer is key to visual generation. In: ICLR (2024)
68. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Randomized autoregressive visual generation. arXiv preprint arXiv:2411.00776 (2024)
69. Yu, Q., Weber, M., Deng, X., Shen, X., Cremers, D., Chen, L.C.: An image is worth 32 tokens for reconstruction and generation. NeurIPS (2024)
70. Zeng, G., Zhang, X., Wang, Z., Xu, H., Chen, Z., Li, B., Tu, Z.: Yolo-count: Differentiable object counting for text-to-image generation. In: ICCV (2025)
71. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
72. Zhao, Y., Xiong, Y., Krahenbuhl, P.: Image and video tokenization with binary spherical quantization. In: ICLR (2025)
73. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model (2024), `https://api.semanticscholar.org/CorpusID:271909855`