TOKEN ALLOCATION FOR BETTER GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive image generation relies on visual tokenizers to convert images into discrete tokens. Recent variable-rate tokenizers dynamically allocate different token lengths to different images to optimize reconstruction. In this paper, we first show that token allocation optimized for a reconstruction objective, such as mean-square-error (MSE), might lead to worse generation. Next, we propose **GTA** (Generation-Aware Token Allocation), a variable-rate tokenization method optimized for generation. More specifically, we first introduce a latent-context formulation for variable-rate autoregressive modeling to enable generation with flexible token lengths. Further, we propose a selection criterion that assigns different token lengths to images for better generation. Empirically, GTA consistently improves downstream generation across different tokenizers. On GigaTok, GTA reduces gFID from 5.50 to 4.75 compared to its fixed-length counterpart.

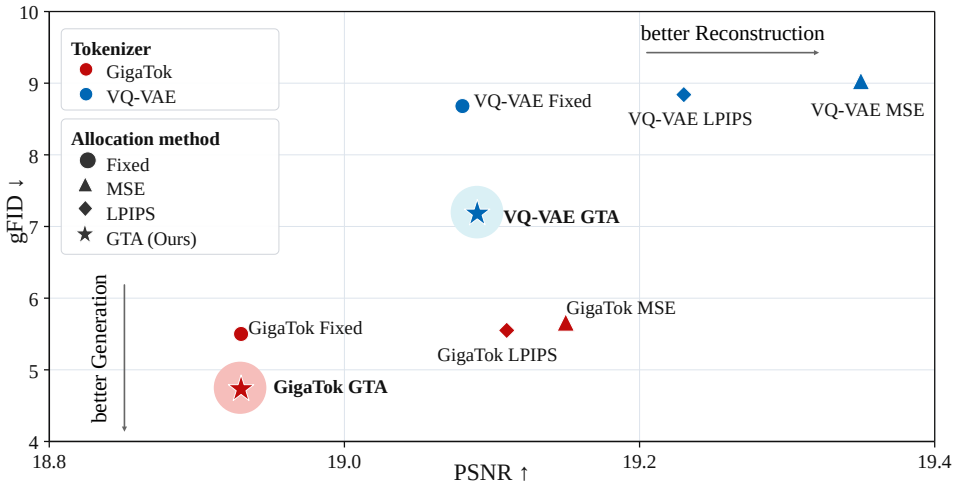


Figure 1: **Reconstruction–generation mismatch in variable-rate tokenization.** Under the same average token length $L = 192$, reconstruction-oriented allocations based on MSE (InfoTok-style (Ye et al., 2026)) and LPIPS (EVATok-style (Xiong et al., 2026)) improve their corresponding reconstruction objectives but do not translate into better downstream gFID. In contrast, GTA achieves substantially better generation quality on both GigaTok and VQ-VAE tokenizers.

1 INTRODUCTION

Discrete visual tokenizers are key to autoregressive visual content generation (Sun et al., 2024; Wang et al., 2024b). These models first convert high-dimensional images into discrete token sequences using a visual tokenizer (van den Oord et al., 2017; Esser et al., 2021), and then formulate image generation as next-token prediction. Most visual tokenizers represent every image with a fixed-length token sequence (Yu et al., 2024b; Xiong et al., 2025). While simple and convenient, fixed-length methods ignore the substantial variation in visual structure and reconstruction difficulty across images, potentially allocating unnecessary capacity to some images while limiting others.

On the other hand, recent visual tokenizers have explored variable-rate tokenization (Bachmann et al., 2025; Miwa et al., 2025), where different images are represented with different numbers

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

of visual tokens. In adaptive video tokenization, InfoTok allocates tokens according to information content, while EVATok uses learned routers to predict token assignments that balance reconstruction quality and token cost (Ye et al., 2026; Xiong et al., 2026). These works demonstrate improved reconstruction–token-cost trade-offs through adaptive allocation. This principle is also aligned with rate allocation methods (Li et al., 2014) and rate–distortion optimization in classical and learned compression (Shannon, 1959; Ballé et al., 2017).

In this paper, we observe a **reconstruction–generation mismatch** specifically in token allocation: under the same average token length, allocating more representation capacity to improve reconstruction can fail to improve, and even degrade downstream generation quality. This mismatch suggests that reconstruction alone is insufficient to guide rate allocation. For autoregressive generation, token length not only controls the reconstruction of images in the training dataset, but also determines the distribution of visual-token sequences used to train the downstream generator. This leads to our central question: How can we optimize token allocation for better generation?

To address this challenge, we introduce **generation-aware token allocation (GTA)**, a variable-rate tokenizer optimized for generation instead of reconstruction. Specifically, since image-dependent token length must be determined before visual content is fully generated, we introduce a latent context token (LCT), denoted by U, a single compressed token that is generated first to determine the token length in autoregressive generation. Further, we introduce a generation-aware criterion for token allocation, using the feature statistics sensitivity of token length. Empirically, GTA consistently improves downstream generation across both GigaTok and VQ-VAE tokenizers.

Our main contributions are summarized as follows:

- We identify **reconstruction–generation mismatch** in variable-rate visual tokenization: improving reconstruction through token allocation does not necessarily improve downstream generation, and can even degrade it under the same average token length.
- We introduce **generation-aware token allocation (GTA)**, a variable-rate tokenizer optimized for generation. GTA consists of a method that enables effective variable-length autoregressive generation and a criterion that allocates different token lengths to different images.
- Extensive experiments across token lengths and visual tokenizers show consistent generation improvements of GTA: gFID improves from 5.50 to 4.75 on GigaTok and from 8.68 to 7.20 on VQ-VAE.

2 Variable-Rate Tokenizer for Autoregressive Generation

2.1 Problem Formulation

Let $x \sim p_d$ denote an image drawn from the data distribution. We consider a length-controllable visual tokenizer $q_\phi(z \mid x, K)$, which represents image x as a discrete visual sequence z using K tokens. Given a target average training token length L, a token-allocation policy $r(K \mid x)$ determines the token length assigned to each image. We distinguish this training budget from $\bar{K}_{\text{gen}}$, the empirical average visual-token length of generated samples, excluding conditioning tokens. Together, the allocation policy and tokenizer induce the training distribution over visual-token sequences:

$$q_r(z) = \mathbb{E}_{x\sim p_d,\, K\sim r(\cdot|x)}[q_\phi(z \mid x, K)]. \tag{1}$$

Eq. 1 highlights a key role of token allocation: it determines not only how many tokens each sample receives, but also the discrete representation distribution presented to the downstream generator. For a fixed allocation r, we denote the final image distribution obtained by decoding samples from the learned autoregressive model as $p_g^r(x)$, the tokenizer decoder as $g(\cdot)$, the discrepancy between two distributions as $d(\cdot,\cdot)$, and the distortion between two images as $\Delta(\cdot,\cdot)$.

If the tokenizer is optimized **for reconstruction**, the optimization of token allocation is

$$\bar{r} = \arg\min_{r(K|x)} \mathbb{E}_{q_\phi(z|x,K)r(K|x)p_d(x)}[\Delta(x, g(z))], \text{ s.t. } \mathbb{E}_{r(K|x)p_d(x)}[K] = L. \tag{2}$$

Table 1: **Matched reconstruction can yield different generation quality.** Both pairs use the 50-epoch recipe. Within each pair, we match the corresponding reconstruction distortion while optimizing the allocation in opposite directions of the GTA score. The LPIPS-matched pair has LPIPS 0.2776, while the MSE-matched pair has MSE 0.01559.

			Reconstruction			Generation		
Constraint	**Train** L	**GTA Score** $\uparrow$	**LPIPS** $\downarrow$	**MSE** $\downarrow$	**PSNR** $\uparrow$	$\bar{K}_{\text{gen}}$	**gFID** $\downarrow$	**IS** $\uparrow$
LPIPS	192	4.88	.2776	.01559	19.10	192.39	**5.30**	**206.12**
	192	-4.99	.2776	.01561	19.09	189.43	6.64	188.78
MSE	192	4.88	.2776	.01559	19.10	191.25	**5.38**	**221.45**
	192	-5.01	.2787	.01559	19.08	190.99	6.57	186.10

On the other hand, if the tokenizer is optimized **for generation**, the optimization of token allocation is

$$r^\star = \arg\min_{r(K|x)} d(p_d, p_g^r), \text{ s.t. } \mathbb{E}_{r(K|x)p_d(x)}[K] = L. \tag{3}$$

2.2 Reconstruction–Generation Mismatch

Most existing variable-rate tokenizers follow the rate allocation convention in image and video compression (Li et al., 2014), *i.e.,* they use the reconstruction-optimized allocator $\bar{r}$. However, the reconstruction-optimized allocator might differ substantially from the generation-optimized allocator r^* and might not benefit generation.

We verify this difference with a controlled matched-RD experiment in Table 1. More specifically, we construct allocation pairs with the same average token length and closely matched reconstruction distortion. Despite their similar RD performance, they produce different downstream generation quality. This demonstrates a **reconstruction–generation mismatch**: reconstruction utility alone is insufficient to determine the generative utility of a token-length assignment.

3 Generation-Aware Token Allocation

The previous section defines the desired generation-oriented objective, but realizing it in practice introduces two challenges. First, training-time token allocation can depend on the observed image, whereas at generation time the token length must be determined before the visual content is available. Second, evaluating the true generative utility of a candidate allocation would require retraining the downstream AR model, making direct optimization infeasible at scale.

We reuse a single fixed-length AR checkpoint in two complementary roles. Its hidden representations provide the basis for learning image-level latent context, while its generated samples provide a distribution-level signal for generation-aware token allocation. We first describe latent-context modeling for variable-length generation, then introduce the criterion for selecting training-time token lengths. Figure 2 summarizes the resulting framework.

3.1 Latent Context for Variable-Length Generation

Latent context for generating token lengths. Training-time token allocation can assign an image-dependent token length because the complete image is observed. During autoregressive generation, however, the token length must be determined before the visual sequence is available. Given a generation condition c, directly modeling $p(K \mid c)$ provides no additional sample-level context before visual generation.

We address this asymmetry by introducing a discrete **Latent Context Token (LCT)** U, generated before the token length, to provide shared context for length selection and subsequent visual generation:

$$p(K \mid c) = \sum_u p(U = u \mid c)\, p(K \mid U = u, c). \tag{4}$$

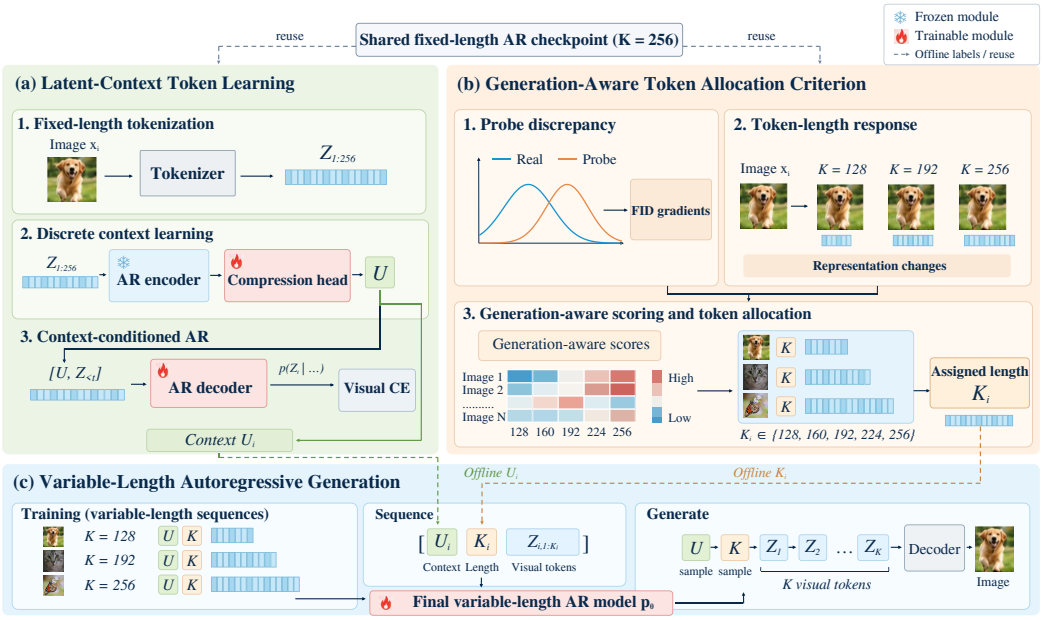


Figure 2: **Overview of the proposed framework.** A length-controllable tokenizer provides candidate representations at multiple token lengths. A fixed-length AR model provides both representations for learning latent context and a distributional probe for generation-aware token allocation. The final variable-length AR model generates latent context, length, and visual tokens autoregressively.

Learning latent context without token-length supervision. The latent context should capture image-level variation useful for autoregressive visual modeling without being tied to a particular token allocation. We therefore learn U entirely from fixed-length autoregressive modeling, without using GTA assignments or token-length labels. Given a complete fixed-length sequence, a posterior branch infers a discrete context U, which is then used to condition autoregressive prediction of the same visual sequence:

$$[U, Z_{1:256}].$$

Because every example uses the same visual-token length during this stage, learning the latent context is independent of the subsequent token allocation.

After latent-context learning, each training example is assigned a discrete context U_i. Given token lengths K_i selected by the criterion in Sec. 3.2, we combine them with U_i and train the final variable-length AR model on

$$[U_i, K_i, Z_{i,1:K_i}].$$

The resulting model factorizes as

$$p(U, K, Z \mid c) = p(U \mid c)\, p(K \mid U, c)\, p(Z \mid U, K, c). \tag{5}$$

At inference time, generation proceeds as

$$U \to K \to Z_{1:K},$$

so both the latent context and the image-dependent token length are generated autoregressively before the visual sequence is completed.

Architecture. LCT is implemented by reusing the same fixed-length AR checkpoint with $K = 256$ that also serves as the GTA probe, without introducing a separate visual backbone. A frozen copy processes the complete fixed-length visual sequence and maps its hidden representation through a 16-way discrete bottleneck to obtain U. A second copy, initialized from the same checkpoint, is trained for context-conditioned visual autoregressive modeling. After this fixed-length latent-context learning stage, the posterior branch is used only to assign U_i to the training data and

is discarded at inference. LCT-training details are documented in Appendix C, with final AR training and sampling in Appendix D.

3.2 Generation-Aware Token Allocation Criterion

With the variable-length generation model in place, we now specify how to select token lengths for training images under a shared average budget. The allocation criterion uses the fixed-length AR checkpoint as a distributional probe and does not require latent-context assignments.

Generation discrepancy from an AR probe. We use a fixed-length AR model as a probe of the current generation distribution. Specifically, we use $\mathcal{D}_{\mathrm{FID}}$, the Fréchet distance in a fixed pretrained feature space, to instantiate the discrepancy d in Sec. 2.1. Let (μ_p, Σ_p) denote the feature statistics of samples generated by the probe and (μ_r, Σ_r) those of real data. At the probe distribution, we compute

$$g_\mu = \nabla_{\mu_p} \mathcal{D}_{\mathrm{FID}}, \qquad G_\Sigma = \nabla_{\Sigma_p} \mathcal{D}_{\mathrm{FID}}. \tag{6}$$

These gradients specify directions in feature-statistics space that locally change the discrepancy between the probe generation distribution and real data.

Token-length-induced representation changes. A length-controllable tokenizer provides the complementary signal: how the representation of an individual training example changes with token length. Let f_i^K denote the feature of image i reconstructed using K visual tokens. Relative to a reference token length K_0, replacing the reference feature $f_i^{K_0}$ with f_i^K in the empirical dataset induces changes $\Delta\mu_i^K$ and $\Delta\Sigma_i^K$ in the corresponding feature statistics.

Because the probe gradient and the token-length-induced response are expressed in the same statistics space, we use their alignment as a local surrogate for the generative utility of an image–length assignment. A first-order expansion gives the predicted discrepancy change $g_\mu^\top \Delta\mu_i^K + \langle G_\Sigma, \Delta\Sigma_i^K \rangle_F$. We define the GTA score as

$$s(i, K) = -\left[g_\mu^\top \Delta\mu_i^K + \left\langle G_\Sigma, \Delta\Sigma_i^K \right\rangle_F \right]. \tag{7}$$

A larger score indicates that the representation change induced by assigning token length K to image i is more aligned with locally reducing the discrepancy exposed by the AR probe.

Length-constrained token allocation. Given N training examples and target average token length L, we select token lengths by optimizing

$$\max_{\{K_i\}_{i=1}^N} \sum_{i=1}^N s(i, K_i) \qquad \text{s.t.} \quad \sum_{i=1}^N K_i = NL, \quad K_i \in \mathcal{K}. \tag{8}$$

We use $\mathcal{K} = \{128, 160, 192, 224, 256\}$, a fixed-length AR probe with $K = 256$, and reference token length $K_0 = 192$. In practice, we express covariance changes through raw second moments and enforce the target length using a global price search followed by discrete length repair. The finite-sample score and exact-length allocation procedure are provided in Appendix B.

4 Experiments

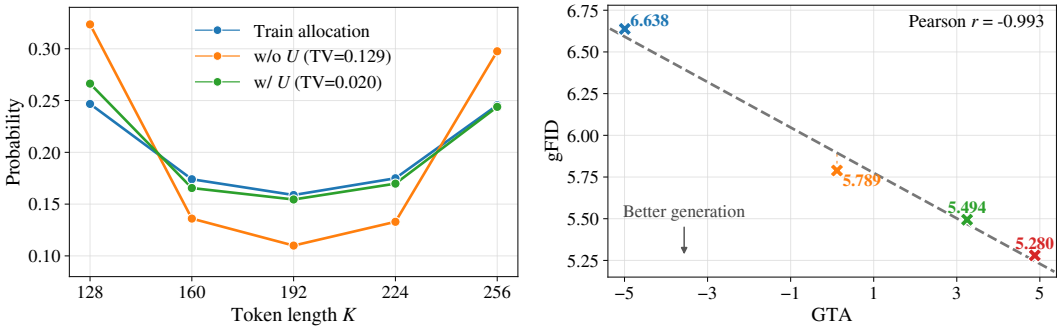
4.1 Experimental Setup

Model setting. We study token allocation while keeping the visual tokenizer fixed within each tokenizer family. We consider length-controllable variants of GigaTok Xiong et al. (2025) and VQ-VAE van den Oord et al. (2017) with candidate lengths $\{128, 160, 192, 224, 256\}$ and use $L = 192$ as the default average token budget. All allocation methods use LlamaGen-B as the downstream AR model.

Baselines. We compare GTA with three baselines. For the **InfoTok-style** baseline, adaptive token rates are motivated by the ELBO-based information criterion in InfoTok (Ye et al., 2026). The method further shows that reconstruction error alone is sufficient for token-length selection in practice. Following its MSE-based implementation, we use MSE reconstruction error as the allocation

Table 2: **System-level comparison of variable-length autoregressive generation.** All methods use the same average training token length $L = 192$. Best results are shown in **bold** and second-best results are underlined.

		Reconstruction			Generation				
Tokenizer	Method	LPIPS↓	PSNR↑	rFID↓	$\bar{K}_{\text{gen}}$↓	gFID↓	sFID↓	IS↑	$\text{FD}_{\text{DINOv2}}$ ↓
GigaTok	Fixed	0.2742	18.93	1.79	192.00	5.50	6.39	212.03	243.47
	INFOTOK-style	0.2737	**19.15**	1.98	**185.56**	5.66	6.22	200.65	259.08
	EVATOK-style	**0.2671**	19.11	1.82	187.12	5.55	6.09	203.85	260.70
	GTA (Ours)	0.2806	18.93	**1.52**	185.96	**4.75**	**5.98**	**225.15**	**243.12**
VQ-VAE	Fixed	0.3057	19.08	**6.61**	192.00	8.68	11.77	174.65	409.16
	INFOTOK-style	0.3056	**19.35**	8.44	189.51	9.03	11.66	177.93	406.61
	EVATOK-style	**0.2968**	19.23	6.65	**184.02**	8.84	11.37	177.55	406.24
	GTA (Ours)	0.3105	19.09	6.98	195.65	**7.20**	**9.19**	**180.50**	**389.97**



(a) Token-length distribution modeling

(b) GTA score vs. downstream gFID

Figure 3: **Diagnostics of latent-context modeling and generation-aware token allocation. Left:** Generated token-length distributions with and without LCT.LCT substantially reduces the distribution mismatch, showing improved token-length modeling. **Right:** GTA score versus downstream gFID across representative token-allocation strategies.

signal. For the **EVATok-style** baseline, we use LPIPS improvement across token lengths, following its quality–cost formulation (Xiong et al., 2026). All adaptive methods use the same candidate lengths and average token budget.

Evaluation and protocol. We evaluate reconstruction with LPIPS (Zhang et al., 2018), PSNR, and rFID, and generation with gFID, sFID (Nash et al., 2021), $\text{FD}_{\text{DINOv2}}$ (Stein et al., 2023), IS (Salimans et al., 2016), and the average generated visual-token length. Both rFID and gFID use Fréchet Inception Distance (Heusel et al., 2017).

Unless otherwise specified, AR models are trained for 300 epochs; controlled studies use a 50-epoch recipe where noted. We sweep CFG from 1.5 to 3.0 in increments of 0.25 and select the setting with the lowest gFID. All other generation metrics are reported at the same CFG. Compared variants otherwise share the same optimization, sampling, and evaluation protocol. Additional details are provided in the Appendix.

4.2 EXPERIMENTAL RESULTS

Overall performance. Across both tokenizer families, the GTA system reduces gFID by approximately 14%–17% over fixed-length tokenization and achieves the best sFID, $\text{FD}_{\text{DINOv2}}$, and IS (Table 2). InfoTok-style and EvaTok-style allocation improve their respective reconstruction objectives but yield worse gFID than the fixed-length baseline, showing that better reconstruction does not necessarily translate into better generation.

Robustness across token lengths. Across average visual-token lengths $L \in \{160, 192, 224\}$, reconstruction-oriented strategies achieve the best values of their respective distortion metrics, whereas the GTA system consistently achieves the best gFID and sFID (Table 3). The reconstruction–generation mismatch therefore persists across the tested token lengths.

Table 3: **Effect of token length on token allocation.** Results on GigaTok under the 50-epoch training recipe across average token lengths $L \in \{160, 192, 224\}$. Each cell reports results in this order. Best results at each length are shown in **bold**.

Method	LPIPS ↓	PSNR ↑	rFID ↓	gFID ↓	sFID ↓
Fixed	.3065 / .2742 / .2467	18.16 / 18.93 / 19.70	2.69 / 1.79 / 1.35	6.99 / 6.51 / 5.79	7.69 / 6.71 / 5.87
InfoTok-style	.3069 / .2738 / .2454	**18.32 / 19.16 / 19.92**	2.99 / 1.98 / 1.40	6.88 / 6.33 / 5.87	7.42 / 6.36 / 5.76
EVATok-style	**.3004 / .2671 / .2408**	18.28 / 19.11 / 19.88	2.73 / 1.82 / 1.35	6.53 / 6.08 / 5.83	6.95 / 6.18 / 5.72
GTA (Ours)	.3116 / .2806 / .2519	18.13 / 18.93 / 19.72	**2.08 / 1.52 / 1.19**	**5.74 / 5.28 / 5.17**	**6.33 / 6.07 / 5.49**

4.3 Ablation Studies

4.3.1 Joint Effects of Token Allocation and Latent Context

We first disentangle token allocation from latent context by independently shuffling the assignments K_i and U_i across training images while preserving their marginal distributions.

Generation is substantially more sensitive to disrupting the token-length assignments than to disrupting the latent contexts (Table 4). Even with the correct U_i, shuffling K_i largely removes the generation benefit, showing that matching representation capacity to the appropriate samples is the primary factor. Latent context provides a complementary gain, but cannot compensate for an incorrect image–length assignment. We next isolate the allocation criterion and then study latent-context modeling in more detail.

Table 4: **Joint ablation of allocation and latent context.** With marginals preserved, shuffling only token lengths causes a larger gFID degradation than shuffling only contexts. Checkmarks denote the original sample-specific assignments.

Latent context	Token length	gFID ↓	sFID ↓	IS ↑
✓	✓	**5.28**	**6.07**	**203.67**
Shuffled	✓	5.45	6.37	201.87
✓	Shuffled	5.82	6.12	195.23
Shuffled	Shuffled	5.85	6.38	198.53

4.3.2 Generation-Aware Token Allocation Criterion

Effect of the allocation criterion. We isolate token allocation from latent-context modeling by comparing different allocation strategies without LCT.

The GTA criterion achieves a clear generation advantage even in the absence of LCT (Table 5), showing that the improvement already arises from the allocation rule itself. The strong negative association between GTA score and downstream gFID in Fig. 3 further supports the score as a useful surrogate for generation-oriented allocation. Importantly, the same allocation also improves $\mathrm{FD}_{\mathrm{DINOv2}}$, despite being constructed in the ADM Inception feature space, providing initial evidence that the gain is not restricted to the metric used to define the score.

Table 5: **Ablation of token allocation.** The GTA criterion achieves the best gFID and $\mathrm{FD}_{\mathrm{DINOv2}}$, indicating that its benefit is not confined to the Inception feature space used to construct the GTA score.

Allocation	$\bar{K}_{\mathrm{gen}}$ ↓	gFID ↓	sFID ↓	IS ↑	$\mathrm{FD}_{\mathrm{DINOv2}}$ ↓
Fixed	192.00	5.50	6.39	**212.03**	243.47
InfoTok-style	185.57	5.66	6.22	200.65	259.08
EVATok-style	188.03	5.55	**6.08**	203.85	260.70
GTA-criterion	**185.50**	**5.15**	6.14	209.06	**237.95**

Sensitivity to the scoring feature space. We next test the criterion itself by changing the feature representation used to construct the GTA score. Specifically, we replace the default Inception features with DINOv2 features while keeping the remaining allocation pipeline unchanged.

DINO-based GTA continues to improve generation under both gFID and $\mathrm{FD}_{\mathrm{DINOv2}}$ (Table 6). Together with the cross-metric improvement of the default Inception-based criterion, this result shows that generation-aware allocation is not tied to a particular feature representation and reduces the concern that GTA merely optimizes the feature space used to define its score.

Table 6: **Scoring feature space.** DINO-GTA replaces the default Inception features in GTA with DINOv2 features.

Allocation	gFID ↓	$\mathrm{FD}_{\mathrm{DINOv2}}$ ↓
Fixed	5.50	243.47
MSE	5.66	259.08
LPIPS	5.55	260.70
DINO-GTA	5.32	240.11
GTA	**5.15**	**237.95**

Table 7: **LCT across allocation strategies.** Comparison of models with and without LCT under different token-allocation strategies.

	w/o LCT			w/ LCT		
Allocation	gFID ↓	sFID ↓	IS ↑	gFID ↓	sFID ↓	IS ↑
Fixed	5.50	6.39	**212.03**	5.26	6.38	**229.28**
MSE	5.66	6.22	200.65	5.22	**5.83**	222.67
LPIPS	5.55	**6.08**	203.85	5.33	5.88	224.46
GTA	**5.15**	6.14	209.06	**4.75**	5.98	225.15

4.3.3 Latent-Context Modeling

LCT across allocation strategies. We further examine how latent-context modeling interacts with different token-allocation strategies. LCT improves generation across all allocations, but does not remove the importance of the allocation rule: the best performance is achieved when LCT is combined with GTA (Table 7). Together with the shuffling ablation in Table 4, this result shows that effective variable-length generation relies on both generation-aware token allocation and sample-specific latent context. The strongest performance is obtained when the assigned token length and latent context are both matched to the visual content.

Effect of latent context. With the GTA criterion fixed, we compare models without latent context, with shuffled context, and with sample-specific context.

LCT improves gFID and sFID and reduces K-TV, bringing the generated token-length distribution closer to the target allocation (Table 8; Fig. 3, left). The average generated length remains nearly unchanged, suggesting that the improvement does not simply come from generating more tokens. Sample-specific context also yields lower visual-token NLL than shuffled U, indicating that its correspondence with visual content helps autoregressive modeling.

Table 8: **Ablation of latent-context modeling.** With the GTA criterion fixed under the 50-epoch setting at $L = 192$, we compare models without latent context, with shuffled context, and with the learned sample-specific context.

LCT	gFID ↓	sFID ↓	IS ↑	$\bar{K}_{\mathrm{gen}}$	K-TV ↓	Visual NLL ↓
w/o U	5.798	6.513	**204.86**	190.241	0.129	8.353
Shuffled U	5.458	6.366	201.87	190.586	0.025	8.354
w/ U	**5.280**	**6.066**	203.67	190.699	**0.020**	**8.347**

5 Related Work

5.1 Visual Tokenization for Autoregressive Generation

Discrete visual tokenizers map images into compact code sequences for autoregressive generation (van den Oord et al., 2017; Esser et al., 2021; Sun et al., 2024). Prior work has explored different representation structures and quantization schemes, including residual quantization, finite scalar quantization, and binary spherical quantization (Lee et al., 2022; Mentzer et al., 2024; Zhao et al., 2025). RQ-VAE represents each spatial feature with a stack of codes through successive residual

quantization. FSQ instead quantizes individual latent dimensions to fixed scalar levels, forming an implicit codebook. MAGVIT-v2 and OmniTokenizer extend tokenization across images and videos, while TiTok develops compact one-dimensional visual tokens (Yu et al., 2024a; Wang et al., 2024a; Yu et al., 2024b).

Recent studies increasingly consider how visual representations interact with downstream generation. GigaTok and ViTok show that better reconstruction does not necessarily lead to comparable gains in generation (Xiong et al., 2025; Hansen-Estruch et al., 2025). More broadly, REPA-E jointly adapts the visual autoencoder and generative model, while tokenizer post-training adapts the tokenizer using latents from a trained generator (Leng et al., 2025; Qiu et al., 2025). These studies highlight the importance of both representation fidelity and the learnability of visual tokens for downstream generative models.

5.2 Adaptive Tokenization and Rate Allocation

Flexible tokenizers allow different inputs to use different numbers of visual tokens. FlexTok and One-D-Piece provide variable-length image representations with controllable reconstruction quality (Bachmann et al., 2025; Miwa et al., 2025). FlexTok combines nested dropout with a rectified-flow decoder to reconstruct images from variable-length token prefixes. Its token sequence progresses from high-level semantics to finer details. One-D-Piece uses Tail Token Drop during training to support reconstruction at different token budgets. ElasticTok adapts token counts to visual complexity, while LaVIT uses dynamic discrete visual sequences for multimodal modeling (Yan et al., 2025; Jin et al., 2024). CAT also adjusts compression ratios according to input complexity (Shen et al., 2025).

InfoTok studies adaptive tokenization from an information-theoretic compression perspective, while EVATok learns quality–cost-oriented routing for variable-length video tokenization (Ye et al., 2026; Xiong et al., 2026). Despite these differences, existing allocation criteria are largely defined from a representation-side or rate–distortion perspective. They focus on the information or reconstruction quality preserved by additional tokens, but do not directly account for how token assignments affect the downstream generative model.

6 Discussion and Limitations

Our results suggest that token allocation is not purely a compression decision: different image–length assignments can induce different training distributions for the downstream generator even at similar reconstruction quality. Adaptive tokenization should therefore consider downstream generative utility in addition to rate–distortion criteria.

GTA introduces additional offline computation to evaluate candidate token lengths using a fixed AR probe. This cost is incurred only before final AR training and avoids retraining the downstream generator for every candidate allocation. Future work could further reduce this overhead by predicting allocation scores from intermediate representations.

More broadly, our study treats the visual tokenizer as a controlled component to isolate token allocation, although tokenizer design itself can affect downstream generation (Yu et al., 2024b; Xiong et al., 2025). Jointly optimizing the tokenizer, token allocation, and autoregressive generator, as well as extending GTA to finer or continuous token-length spaces, remain promising directions for future work.

7 Conclusion

This paper introduced Generation-Aware Token Allocation to address the reconstruction–generation mismatch in variable-length visual tokenization. GTA combines a generation-aware token allocation criterion with LCT modeling to support sample-specific variable-length autoregressive generation. Across GigaTok and VQ-VAE, GTA consistently improves downstream generation over reconstruction-oriented and fixed-length alternatives. Our results suggest that adaptive visual tokenization should allocate representation capacity according to downstream generative utility, rather than reconstruction quality alone.

REFERENCES

Roman Bachmann, Jesse Allardice, David Mizrahi, Enrico Fini, Oğuzhan Fatih Kar, Elmira Amirloo, Alaaeldin El-Nouby, Amir Zamir, and Afshin Dehghan. FlexTok: Resampling images into 1D token sequences of flexible length. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2241–2292. PMLR, 2025.

Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. In *International Conference on Learning Representations*, 2017.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021.

Philippe Hansen-Estruch, David Yan, Ching-Yao Chuang, Orr Zohar, Jialiang Wang, Tingbo Hou, Tao Xu, Sriram Vishwanath, Peter Vajda, and Xinlei Chen. Learnings from scaling visual tokenizers for reconstruction and generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22023–22043. PMLR, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chengru Song, Dai Meng, Di Zhang, Wenwu Ou, Kun Gai, and Yadong Mu. Unified language-vision pretraining in LLM with dynamic discrete visual tokenization. In *International Conference on Learning Representations*, 2024.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11523–11532, 2022.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. REPA-E: Unlocking VAE for end-to-end tuning of latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18262–18272, 2025.

Bin Li, Houqiang Li, Li Li, and Jinlei Zhang. λ domain rate control algorithm for high efficiency video coding. *IEEE Transactions on Image Processing*, 23(9):3841–3854, 2014. doi: 10.1109/TIP.2014.2336550.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *International Conference on Learning Representations*, 2024.

Keita Miwa, Kento Sasaki, Hidehisa Arai, Tsubasa Takahashi, and Yu Yamaguchi. One-D-Piece: Image tokenizer meets quality-controllable compression. In *ICML Workshop on Tokenization (TokShop)*, 2025.

Charlie Nash, Jacob Menick, Sander Dieleman, and Peter Battaglia. Generating images with sparse representations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7958–7968. PMLR, 2021.

Kai Qiu, Xiang Li, Hao Chen, Jason Kuen, Xiaohao Xu, Jiuxiang Gu, Yinyi Luo, Bhiksha Raj, Zhe Lin, and Marios Savvides. Image tokenizer needs post-training. *arXiv preprint arXiv:2509.12474*, 2025.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Claude E. Shannon. Coding theorems for a discrete source with a fidelity criterion. In *IRE National Convention Record*, volume 7, Part 4, pp. 142–163, 1959.

Junhong Shen, Kushal Tirumala, Michihiro Yasunaga, Ishan Misra, Luke Zettlemoyer, Lili Yu, and Chunting Zhou. CAT: Content-adaptive image tokenization. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6306–6315, 2017.

Junke Wang, Yi Jiang, Zehuan Yuan, Binyue Peng, Zuxuan Wu, and Yu-Gang Jiang. OmniTokenizer: A joint image-video tokenizer for visual generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024b.

Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. GigaTok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18770–18780, 2025.

Tianwei Xiong, Jun Hao Liew, Zilong Huang, Zhijie Lin, Jiashi Feng, and Xihui Liu. EVATok: Adaptive length video tokenization for efficient visual autoregressive generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23249–23259, 2026.

Wilson Yan, Volodymyr Mnih, Aleksandra Faust, Matei Zaharia, Pieter Abbeel, and Hao Liu. ElasticTok: Adaptive tokenization for image and video. In *International Conference on Learning Representations*, 2025.

Haotian Ye, Qiyuan He, Jiaqi Han, Puheng Li, Jiaojiao Fan, Zekun Hao, Fitsum Reda, Yogesh Balaji, Huayu Chen, Sheng Liu, Angela Yao, James Y. Zou, Stefano Ermon, Haoxiang Wang, and Ming-Yu Liu. InfoTok: Adaptive discrete video tokenizer via information-theoretic compression. In *International Conference on Learning Representations*, 2026.

Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion – tokenizer is key to visual generation. In *International Conference on Learning Representations*, 2024a.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization. In *International Conference on Learning Representations*, 2025.