Mutual Enhancement Between Global Tokens and Patch Tokens: From Theory to Practice

Anonymous ACL submission

Abstract

Accurate and effective discrete image tokenization is crucial for long image sequence processing. However, current methods rigidly compress all content at a fixed rate, ignoring the variable information density of images and leading to either redundancy or information loss. Inspired by information entropy, we propose **TaTok**, a **T**heoretically grounded **a**daptive image **Tok**enization framework. We rigorously identify two key drawbacks in existing methods: information insufficiency when reconstructing images with patch tokens alone, and information redundancy among patch tokens. To address these, we introduce global tokens that model mutual information across patch tokens, and a **D**ynamic **T**oken **F**iltering (**DTF**) algorithm based on cumulative conditional entropy to eliminate redundancy. Experiments confirm TaTok's state-of-the-art performance, delivering a 1.3× gFID improvement and 8.7× inference speedup. By allocating tokens according to information richness, TaTok enables more compressed yet accurate image tokenization, offering valuable insights for future research.

1 Introduction

Accurate and effective discrete image tokenization is crucial for processing long image sequences (Van Den Oord et al., 2017; Razavi et al., 2019a; Yu et al., 2021a). However, the inherent complexity and variable information density of images pose significant bottlenecks for current tokenization methods—these methods rigidly compress all content at a fixed rate, leading to either redundancy or information loss (Lee et al., 2022).

Discrete visual tokenizers typically comprise an encoder, a quantizer, and a decoder (Esser et al., 2021). The encoder splits an image into patch tokens and embeds them into a vector sequence (Vaswani, 2017; Raffel et al., 2020), which the quantizer discretizes (Radford et al., 2019; Brown, 2020) for the decoder to reconstruct the image. Since reconstruction is trivial with unrestricted sequence length (e.g., retaining all RGB pixels) (Zhang et al., 2022; OpenAI, 2022), a high-performance tokenizer essentially acts as a compressor that preserves core information in compact tokens. Developing efficient, scalable, and theoretically sound tokenizers is thus pivotal for advancing unified multimodal models (Google, 2023).

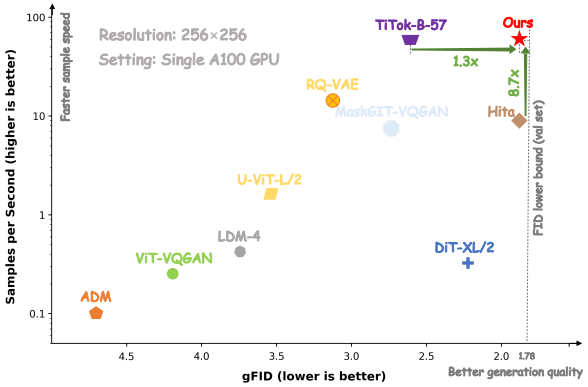

Figure 1: The comprehensive Speed-Quality Comparison of TaTok and Existing Approaches for ImageNet 256×256 Image Generation. The sampling speed is measured with an A100 GPU.

In real-world scenarios, images carry varying amounts of information due to differences in distribution, texture, and other characteristics (Kaplan et al., 2020; Henighan et al., 2020), suggesting that the optimal token count should adapt to image-specific features. However, most existing tokenizers apply a fixed compression rate to all images, regardless of architectural or quantization advances. This one-size-fits-all strategy yields redundant tokens for simple images and insufficient tokens for complex ones (causing key information loss), making downstream understanding or generation tasks inefficient or even intractable (Yu et al., 2021b, 2022). This is one of the core bottleneck limiting current tokenizers.

To deeply dissect this bottleneck, we conduct theoretical analysis via information entropy theory (Gray, 2011), formally proving that existing discrete visual tokenization methods have two fundamental defects: (1) **information insufficiency** (patch tokens alone cannot fully capture image global information) and (2) **information redundancy** (prevalent among patch tokens). Existing studies attempt to tackle these limitations but lack a principled integrated solution (Yu et al., 2024; Zheng et al., 2025a): hierarchical tokenization methods (e.g., ViT-VQGAN (Yu et al., 2021a), Swin Transformer (Liu et al., 2021)) capture global context via patch aggregation but raise computational complexity and fail to offset global information loss (Theorem 2.1 proves the irreducible reconstruction loss lower bound of patch-only schemes), while dynamic token selection approaches (e.g., sparse tokenizers (Nash et al., 2021)) reduce redundancy via heuristic rules (e.g., feature magnitude thresholding) without theoretical guidance on sufficient information for reconstruction, often trading quality for compression. Critically, these methods treat global information modeling and redundancy elimination as separate tasks, failing to synergistically address both defects (Chang et al., 2022; Zheng et al., 2025a).

Against this backdrop, inspired by information entropy (Gray, 2011) and rate-distortion (Berger, 2003) theory, we propose **TaTok**, a **T**heoretically grounded **a**daptive image tokenization framework. TaTok addresses the dual defects via two complementary mechanisms: *global information completion* to compensate for the inherent insufficiency of patch tokens, and *adaptive redundancy pruning* to dynamically match the number of tokens with the image's information density.

To tackle information insufficiency, we theoretically derive the necessity of introducing a learnable global token that explicitly models the holistic semantic and structural information of images (Lemma 3.1). Unlike the auxiliary global features in previous works (Yu et al., 2024; Zheng et al., 2025a), our global token is optimized to minimize conditional entropy, thereby directly enhancing the mutual information between the augmented token sequence and the original image. As formally proven in Theorem 3.2, this augmentation not only enriches the information carried by the token sequence but also reduces the theoretical infimum of reconstruction loss, laying a rigorous foundation for high-fidelity image recovery with a limited number of tokens.

For redundancy elimination, we propose a **D**ynamic **T**oken **F**iltering (**DTF**) algorithm based on cumulative conditional entropy, which adaptively selects the minimum number of patch tokens required for image reconstruction. Guided by rate-distortion theory (Berger, 2003), DTF introduces an information loss rate to balance compression efficiency and reconstruction quality, ensuring that the combined information of global tokens and selected patch tokens meets a predefined minimum information requirement. Specifically, we sort patch tokens by their conditional entropy, a metric quantifying the unique information each token contributes given the global token, and accumulate their information until the dual constraints of sufficient information and an acceptable loss rate are satisfied. This design ensures that only information-rich patch tokens are retained and redundant ones are filtered out, enabling adaptive token allocation that aligns with the inherent information density of each image.

Notably, TaTok is the first tokenizer that unifies global tokens and the dynamic filtering mechanism into an end-to-end trainable framework without introducing additional inference overhead. Extensive experiments demonstrate that TaTok achieves SOTA compression performance. The contributions of this work are summarized as follows:

- We theoretically prove that current tokenizers suffer from two critical flaws: global information deficiency and information redundancy in patch tokens.
- To address these two flaws, we propose TaTok, which unifies global tokens and the dynamic filtering mechanism into an end-to-end trainable framework, and theoretically proves its superiority.
- Extensive experiments show that TaTok achieves superior performance, striking a favorable balance between reconstruction speed and quality.

2 Motivation: Existing inherent flaws

Focusing on the core limitations of discrete visual tokenizers (based on ViT encoders) in image reconstruction tasks, this section rigorously proves two inherent flaws using information theory and rate-distortion (Berger, 2003) theory: information insufficiency and information redundancy.

2.1 Framework and Core Constraints of Discrete Visual Tokenizers

We consider a natural image represented as $x \in \mathcal{X} \subseteq \mathbb{R}^{H \times W \times C}$, where H and W denote the height and width of the image, respectively, and C denotes the number of channels. The probability distribution $P(x)$ of the image satisfies **local stationarity** (statistical correlation between any pixel and its neighboring pixels) and **non-local semantic correlation** (distributional association between different regions of the same semantic object) (Yu et al., 2021a). A discrete visual tokenizer consists of three deterministic mappings:

- **Patch Partitioning**: The image x is partitioned into $N = \lceil \frac{H-s(1-r)}{sr} \rceil \times \lceil \frac{W-s(1-r)}{sr} \rceil$ patches with size $s \times s$ and overlap rate $r \in [0, 1)$, denoted as $\{x_1, x_2, ..., x_N\}$ (where $x_i \in \mathbb{R}^{s^2 C}$);
- **ViT Encoding**: A global interaction mapping $\text{Enc} : \mathbb{R}^{N \times s^2 C} \to \mathbb{R}^{N \times D}$ (where D is the token dimension), with the pipeline: "Patch Embedding $\to$ Positional Encoding $\to$ Self-Attention $\to$ FFN", outputting a sequence of continuous tokens $z = \text{Enc}(x)$;
- **Quantization Discretization**: A quantization function $Q : \mathbb{R}^{N \times D} \to \{1, 2, ..., K\}^N$ (where K is the codebook size), outputting discrete token indices $q = Q(z)$. The reconstructed image $\hat{x} = \text{Dec}(q)$ is finally obtained via the decoder $\text{Dec} : \{1, ..., K\}^N \to \mathbb{R}^{H \times W \times C}$.

The core goal of tokenization is to minimize the reconstruction loss $\mathcal{L}(x, \hat{x}) = \mathbb{E}[\|x - \hat{x}\|_2^2]$ under a finite code rate $R = \frac{N \log_2 K}{HWC}$ (bits per pixel).

2.2 Defect I: Information Insufficiency

The essence of information insufficiency is that the irreversible mappings in the tokenization pipeline are constrained by rate-distortion (Berger, 2003) theory, making zero-distortion reconstruction impossible under finite code rates. First, both patch partitioning and ViT encoding are deterministic mappings. By the Data Processing Inequality (Beaudry and Renner, 2011), the differential mutual information of continuous variables satisfies $I(x; z) \leq I(x; \{x_1, ..., x_N\})$. Since patch partitioning is a full-coverage mapping ($x = \text{Concatenate}(x_1, ..., x_N)$), $I(x; \{x_1, ..., x_N\}) = h(x)$ (where $h(x)$ is the differential entropy of x), so $I(x; z) \leq h(x)$. When $D < s^2 C$, the dimension reduction in patch embedding is a many-to-one mapping—multiple $\{x_1, ..., x_N\}$ correspond to the same z—thus $I(x; z) < h(x)$, and information loss occurs during encoding.

Further, the quantization operation $Q(z) = q$ is a deterministic interval-partitioning mapping. The mutual information between continuous and discrete variables satisfies $I(x; q) = I(x; z) - I(x; z \mid q)$. Since $I(x; z \mid q) \geq 0$, we have $I(x; q) \leq I(x; z) < h(x)$, and quantization exacerbates information loss. According to **Shannon's Rate-Distortion Theorem** (Berger, 2003), for any continuous source, the rate-distortion function satisfies $R(D) \geq h(x) - \frac{1}{2} \log_2(2\pi e D)$. Specifically, for a Gaussian source, $R(D) = \frac{1}{2} \log_2 \left(\frac{\sigma_x^2}{D} \right)$ (when $D \leq \sigma_x^2$, where $\sigma_x^2 = \text{Var}(x)$ is the pixel variance). Natural images are not Gaussian distributions, and their rate-distortion functions have more complex forms, but they still satisfy similar lower-bound constraints—a finite code rate R corresponds to a positive minimum distortion $D_{\text{min}} > 0$.

Theorem 2.1 (Reconstruction Error Infimum of Discrete Visual Tokenizers)**.** *For a discrete visual tokenizer consisting of an encoder Enc, a quantizer Q, and a decoder Dec, the theoretical infimum of the reconstruction loss $\mathcal{L}$ (measuring discrepancy between original image x and reconstructed image $Dec(Q(Enc(x))))$ satisfies:*

$$\begin{aligned} \epsilon_{inf} = & \inf_{Enc, Q, Dec} \mathcal{L}\left(x, Dec\left(Q\left(Enc(x)\right)\right)\right) \\ & \geq D_{min} > 0 \end{aligned} \tag{1}$$

where D_{min} denotes a non-zero positive constant. This implies that high-frequency details and continuous grayscale information of the original image ***cannot be fully recovered using only patch tokens***. We present corresponding verification experiments to validate this in Experiment 4.2.

Note that this conclusion relies on the standard design of discrete visual tokenizers (finite code rate, non-invertible encoding). While zero-distortion reconstruction is theoretically feasible with invertible encoders (e.g., Flow-based Transformers ma2024sit) and infinite code rates ($K \to \infty$), this contradicts the core design goals of discrete tokenizers (e.g., "low code rate, discrete representation") and is thus impractical in real-world scenarios.

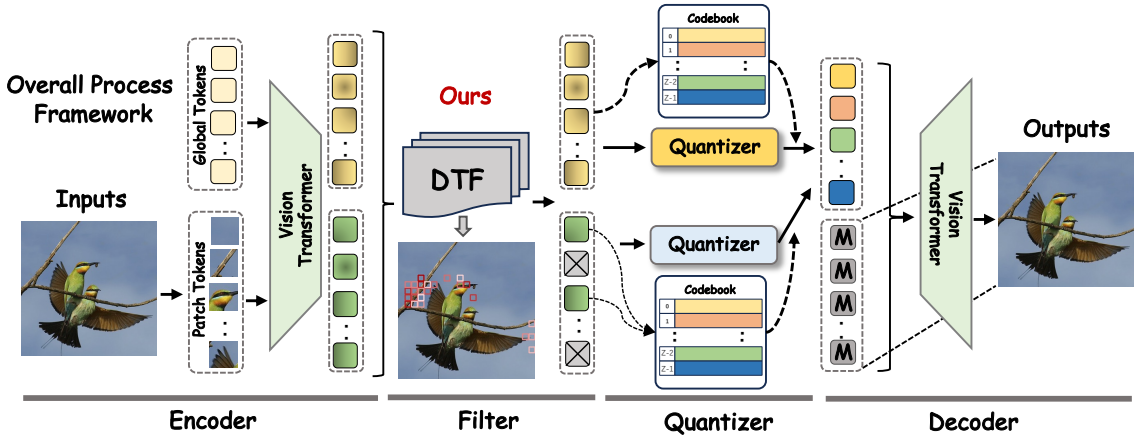

Figure 2: The overall process framework of TaTok. The input image is encoded into patch tokens via a vision transformer encoder, and we additionally introduce learnable global tokens. Our proposed DTF module then removes redundant patch tokens, which are further quantized by dual quantizers with codebooks. Finally, the decoder reconstructs the output image from the processed tokens.

2.3 Defect II: Information Redundancy

Based on multivariate mutual information (Lee and Kim, 2013), the redundancy of a continuous token sequence $z = [z_1, ..., z_N]$ is defined as:

$$\text{Red}(z) = \frac{\sum_{i=1}^{N} h(z_i) - h(z)}{\sum_{i=1}^{N} h(z_i)} \tag{2}$$

where the numerator is the total mutual information between tokens (quantifying information overlap), and the denominator is the sum of the differential entropies of individual tokens (quantifying the total information of tokens). It should be noted that the differential entropy $h(\cdot)$ of continuous variables can be negative, but this only affects the absolute value of the entropy and does not change the quantitative meaning of redundancy: when tokens are fully independent, $h(z) = \sum_{i=1}^{N} h(z_i)$ and $\text{Red}(z) = 0$; when tokens are fully correlated, $h(z) \approx h(z_1)$ (information of other tokens is completely contained in the first token) and redundancy approaches 1.

The local stationarity of natural images implies that adjacent patches x_i, x_j satisfy $I(x_i; x_j) = h(x_i) + h(x_j) - h(x_i, x_j) > 0$; non-local semantic correlation means non-adjacent patches x_k, x_l also satisfy $I(x_k; x_l) > 0$. The global interaction mechanism of ViT encoders typically preserves or enhances correlations between patches; thus, for relevant patch pairs where the original patches satisfy $I(x_i; x_j) > 0$, their corresponding tokens usually also satisfy $I(z_i; z_j) > 0$. Accordingly, the total mutual information between tokens $\sum_{i=2}^{N} I(z_i; z_1, ..., z_{i-1}) \geq \sum_{\text{relevant pairs}} I(z_i; z_j) > 0$, so substituting into the redundancy definition gives $\text{Red}(z) > 0$; for discrete tokens $q = Q(z)$, we similarly obtain $\text{Red}(q) > 0$. This indicates that ***there is information redundancy between patch tokens***. We present corresponding verification experiments to validate this in Experiment 4.3.

Redundancy has two negative impacts: (1) Increased Computational Overhead, the self-attention complexity of ViT is $O(N^2 D)$; redundancy prevents excessive reduction of N, leading to complexity growing with N^2; (2) Degraded Reconstruction Quality, superposition of redundant information in tokens causes blocking artifacts (non-overlapping partitioning) or over-smoothing (overlapping partitioning) in reconstructed images, while amplifying quantization noise.

Overall, the information insufficiency and redundancy flaws of discrete visual tokenizers originate from the inherent contradiction between their discretization design and the continuous correlation of natural images: information insufficiency stems from rate-distortion constraints of three irreversible mappings, while redundancy stems from the inherent correlation of natural images and information fusion via global interaction. These two defects have not been completely eliminated in previous work, but they inspired us to design a new tokenizer framework to eliminate them.

3 TaTok: Theoretically Grounded Adaptive Tokenizer

To address the dual defects of **information insufficiency** and **information redundancy** in discrete

visual tokenizers, we theoretically derive the necessity of introducing a learnable global token $g \in \mathbb{R}^D$ to augment the token sequence, and proposes an adaptive patch token filtering module to eliminate redundancy.

3.1 Necessity of Global Token: Information Completion under Practical Constraints

According to Theorem 2.1, the core of the information insufficiency problem lies in the fact that patch tokens cannot fully capture the global information $\mathcal{G}(x)$, resulting in a strict upper bound on mutual information $I(x; z) < h(x)$. Theoretical deduction shows that introducing global tokens has an enhancing effect on mutual information, and combined with rate distortion theory, it is proven that the global token enhancement mechanism can reduce the theoretical lower bound of reconstruction loss.

Lemma 3.1 (Mutual Information Enhancement Effect of Global Token)**.** *For the augmented token sequence $z' = [g; z]$ generated by the global token-augmented encoder Enc$'$, if g is optimized to approximate the global information $\mathcal{G}(x)$ (e.g., $\lim_{\text{training steps}\to\infty} h(\mathcal{G}(x) \mid g) = 0$), the mutual information between the augmented token sequence and the original image satisfies:*

$$I(x; z') = I(x; z) + I(x; g \mid z) \tag{3}$$

where $I(x; g \mid z) \approx I(x; \mathcal{G}(x) \mid z) > 0$. This lemma indicates that the global token augmentation mechanism enables the augmented token sequence to carry more image information, directly compensating for the information loss caused by patch partitioning, dimension reduction, and quantization operations. We provide detailed theoretical proof in the Appendix A.1.

Theorem 3.2 (Reconstruction Loss Reduction via Global Token)**.** *For a discrete visual tokenizer augmented with a learnable global token (consisting of encoder Enc$'$, quantizer Q, and decoder Dec$'$), the theoretical infimum of its reconstruction loss satisfies:*

$$\epsilon'_{inf} = \inf_{Enc',Q,Dec'} \mathcal{L}\left(x, Dec'\left(Q\left(Enc'(x)\right)\right)\right) < \epsilon_{inf} \tag{4}$$

where ϵ_{inf} is the reconstruction loss infimum of the patch-token-only scheme described in Theorem 2.1. We provide detailed theoretical proof in the Appendix A.2.

Notably, the global token scheme does not conflict with the core design goals of discrete visual tokenizers (low code rate, discrete representation, and efficient inference). Compared with other global information modeling methods (e.g., hierarchical patch aggregation, sparse attention), the global token scheme achieves a better trade-off among information modeling capability, computational complexity, and training stability under practical constraints. This provides a rigorous theoretical foundation for the subsequent design of improved tokenizer frameworks.

3.2 Dynamic Token Filtering Algorithm

In order to eliminate the redundancy of patch tokens, especially under the condition of global tokens, we introduce a **D**ynamic **T**oken **F**iltering (**DTF**) algorithm based on conditional entropy. Consistent with previous research, the entire tokenizer completes the image reconstruction task during the pre training process. Based on rate-distortion (Berger, 2003) theory, DTF finds the minimum number of Patch Tokens N such that the total information of "Global Tokens + N Patch Tokens" meets the minimum information requirement for image reconstruction tasks.

3.2.1 Core Constraint: Information Loss Rate

To balance "minimum number of Tokens" and "acceptable distortion", introduce the information loss rate ϵ ($\epsilon \in [0, 1)$), which indicates that the proportion of unique information lost after selection does not exceed ϵ. Meanwhile, we must satisfy Eq.27. Therefore, N needs to satisfy the following dual constraints: (1) Minimum Information Constraint: $H_N \geq C$. (2)Information Loss Rate Constraint: $H_N \geq (1 - \epsilon)H_{\text{total}}$ (controls the proportion of information loss). Combining these constraints:

$$H_N \geq \max\{T,\ (1 - \epsilon)H_{\text{total}}\} \tag{5}$$

where T represents the minimum unique information that needs to be supplemented by patch tokens, detailed definitions are provided in Appendix B.

3.2.2 Definition and Existence of N

The theoretical information lower bound N is the smallest positive integer satisfying:

$$N = \min\{k \in \mathbb{N}^+ \mid \sum_{i=1}^{k} H(\hat{p}_{\sigma(i)} \mid \hat{G}) \geq \max\{T,\ (1 - \epsilon)H_{\text{total}}\}\} \tag{6}$$

We provide detailed theoretical proof in the Appendix B.

3.3 TaTok De-Tokenization: Reconstruction via Augmented Token Sequence

To implement TaTok's image reconstruction, we design a de-tokenization pipeline using an augmented token sequence (replacing conventional 1D latent tokens).

3.3.1 Augmented Latent Token Sequence

Recall DTF: select top-N patch tokens (sorted by $H(\mathbf{z}_i \mid \mathbf{G})$) to form $\mathbf{z}_\tau \in \mathbb{R}^{N\times D}$. We use k learnable global tokens $\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_k] \in \mathbb{R}^{k\times D}$ (for multi-scale global info) and concatenate them with $\mathbf{z}_\tau$:

$$\mathbf{z}_{\text{aug}} = \mathbf{G} \oplus \mathbf{z}_\tau \in \mathbb{R}^{(k+N)\times D} \tag{7}$$

where $\oplus$ denotes token-dimension concatenation.

3.3.2 Quantization and Decoder Input

Apply quantization $\text{Quant} : \mathbb{R}^{(k+N)\times D} \to \{1, \ldots, K\}^{(k+N)}$ (matching the $Q(\cdot)$ of Theorem 2.1) to get discrete tokens $\mathbf{q}_{\text{aug}} = \text{Quant}(\mathbf{z}_{\text{aug}})$. To align with initial patch count N_0, complement $\mathbf{q}_{\text{aug}}$ with $N_0-(k+N)$ replicated learnable mask tokens $\mathbf{M}$, forming the decoder input:

$$\mathbf{S} = \mathbf{q}_{\text{aug}} \oplus \mathbf{M} \tag{8}$$

The extended ViT decoder outputs the reconstructed image:

$$\hat{\mathbf{x}} = \text{Dec}'(\mathbf{S}) \tag{9}$$

3.3.3 Training Loss

The total loss $\mathcal{L}_{\text{total}}$ jointly optimizes tokenization (encoder+DTF+global tokens) and de-tokenization (decoder), comprising three weighted components: 1) Reconstruction Loss: $\mathcal{L}_{\text{rec}} = \mathbb{E}_{\mathbf{x}}[\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2]$ (pixel-wise discrepancy). 2) Quantization Loss: $\mathcal{L}_{\text{commit}} = \mathbb{E}_{\mathbf{x}} \left\|\mathbf{z}_{\text{aug}} - \mathbf{q}_{\text{aug}}\right\|_2^2$, where $\mathbf{z}_{\text{aug}} = [g; \mathbf{z}]$ (augmented token sequence) and $\mathbf{q}_{\text{aug}}$ is the quantized codebook $\mathbf{E}$ output, aligning augmented tokens with the codebook. 3) Global Semantic Alignment Loss: $\mathcal{L}_{\text{glob}} = \mathbb{E}_{\mathbf{x}}\left[h(\mathcal{G}(\mathbf{x}) \mid g)\right]$, minimizing conditional entropy between image holistic info $\mathcal{G}(\mathbf{x})$ and global token g (Lemma 3.1, Theorem 3.2). The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_1 \mathcal{L}_{\text{commit}} + \lambda_2 \mathcal{L}_{\text{glob}}$$

where $\lambda_1, \lambda_2 > 0$ are balancing hyperparameters, and both are set to 1 throughout all experiments.

4 Experiments

4.1 Setup

Following prior works, we set the codebook dimension for patch-level and global image tokens to 8 and 12, respectively. As discussed in , this configuration balances reconstruction quality and efficient codebook utilization. Both codebooks are fixed to a size of 4096, and the number of queries for capturing global features is set to 16. More details are provided in Appendix F.

4.2 Experiment I: Global Information is Missing

To verify global information deficiency in patch-only discrete visual tokenizers (Defect I), we performed two rFID-based controlled experiments (balanced total tokens, Fig.3). Setting 1 (fixed patch tokens=441, incremental global tokens): initial rFID=2.17, dropping to 1.51 (30.4% reduction) at 16 global tokens and plateauing at 1.48 (200 tokens). Setting 2 (fixed total tokens=441, global tokens replacing patch tokens): initial rFID=2.17, falling to 1.54 at 16 global tokens (higher than Setting 1), 1.50 at 65 tokens, and 1.52 at 200 tokens (all $>$ Setting 1).

Key insights: (1) Reconstruction gains stem from global information supplementation, not more total tokens; (2) Excessive patch token replacement fails to sustain improvements (patch tokens carry essential local information). Experiments confirm severe global deficiency in patch-only tokenizers: 16 global tokens maximize reconstruction quality, while excess tokens offer no benefits and may slightly degrade quality.

4.3 Experiment II: Patch Tokens Contain Redundant Information

To verify information redundancy in patch-only discrete visual tokenizers (Defect II), we conducted three controlled experiments (start/end/random truncation, global tokens=0, rFID as core metric, Fig.4). All groups started with 441 intact patch tokens (rFID=2.17), with rFID gently fluctuating early and surging exponentially later. Deleting 0$\to$211 tokens (230 remaining) caused $<$11% rFID growth: end truncation (2.32, slowest), start truncation (2.35, intermediate), random truncation (2.40, fastest). Beyond 211 deletions, rFID soared exponentially; deleting 440 tokens led to rFID=379 (complete reconstruction collapse).

Table 1: **ImageNet-1K** 256 × 256 **generation results evaluated.** †: Trained on OpenImages (Kuznetsova et al., 2020); ‡: Trained on OpenImages, LAION-Aesthetics/-Humans (Schuhmann et al., 2022). P: generator's parameters. S: sampling steps. T: throughput as samples per seconds on A100 with float32 precision.

tokenizer					**generator**				
Model	Ratio ↑	#Tokens	Codebook Size	rFID ↓ -	Model	gFID ↓	P ↓	S ↓	T ↑
diffusion-based generative models									
Taming-VQGAN†	1	1024	16384	1.14	LDM-8	7.76	258M	200	-
VAE†	1	4096×3	-	0.27	LDM-4	3.60	400M	250	0.4
					UViT-L/2	3.40	287M	50	1.1
VAE ‡	1	1024×4	-	0.62	UViT-H/2	2.29	501M	50	0.6
					DiT-XL/2	2.27	675M	250	0.6
transformer-based generative models									
Taming-VQGAN	1	256	1024	7.94	Taming-Transformer	15.78	1.4B	256	7.5
RQ-VAE	1	256	16384	3.20	RQ-Transformer	8.71	1.4B	64	16.1
ViT-VQGAN	1	1024	8192	1.28	VIM-Large	4.17	1.7B	1024	0.3
Hita	1	569	16384	1.50	MaskGIT-ViT	1.89	177M	8	10.8
SFTok-B-64	4.0x	64	8192	1.44	MaskGIT-ViT	2.06	177M	8	89.8
SFTok-B-57	4.4x	57	8192	1.53	MaskGIT-ViT	2.21	177M	8	94.0
TiTok-B-64	4.0x	64	4096	1.70	MaskGIT-ViT	2.48	177M	8	89.8
TiTok-B-57	4.4x	57	4096	1.75	MaskGIT-ViT	2.57	177M	8	94.0
TaTok(Ours)	**4.5x**	**56.4**	4096	1.51	MaskGIT-ViT	**1.89**	177M	8	**94.3**

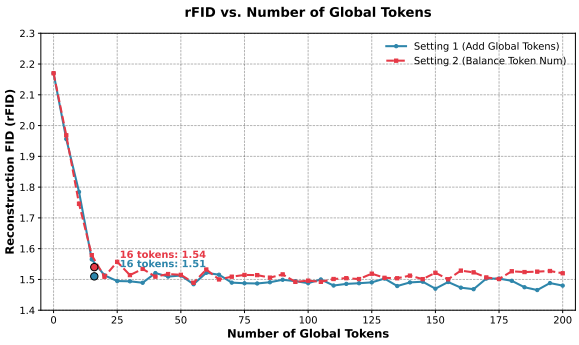

Figure 3: The plot of experimental results for reconstruction performance versus the number of added global tokens. The x-axis denotes the number of added global tokens, and the y-axis represents rFID.

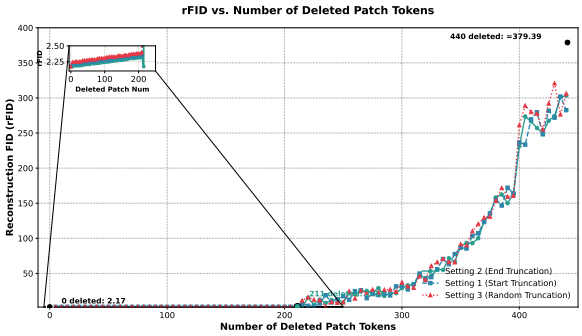

Figure 4: The experimental results of reconstruction performance versus the number of removed patch tokens. The x-axis represents the number of removed patch tokens, and the y-axis represents the rFID score.

Consistent trends confirm: 50% patch tokens are redundant (removal has negligible impact), while over 50% truncation triggers quality collapse (core information loss). Positional heterogeneity (minimal end truncation impact, maximal random) is observed, verifying significant redundancy in patch-only tokenizers and providing basis for precise redundant token screening in optimization.

4.4 Experiment III: Comparative Results

For ImageNet 256 × 256 (Table 1), TaTok outperforms other Transformer-based VQ tokenizers in rFID with fewer latent tokens. With 56.4 tokens (codebook size=4096), its rFID=1.51—surpassing MaskGIT-VQGAN (256 tokens, rFID=2.28) (Yu et al., 2021a) and TiTok-B-57 (57 tokens, rFID=1.75), proving superior compact token information preservation.

Paired with MaskGIT-ViT (Chang et al., 2022), TaTok also boosts generation FID (gFID=1.89), far outperforming MaskGIT-VQGAN (gFID=6.18) and TiTok-B-64 (gFID=2.48) with the same generator, validating more efficient generator training.

In efficiency, TaTok is lightweight: paired MaskGIT-ViT has 177M parameters, 8 sampling steps, and 94.3 samples/s throughput on A100 (float32). Vs. MaskGIT-VQGAN+MaskGIT-ViT (9.7 samples/s, 3.8B parameters), TaTok achieves 9.7x sampling speedup and drastically reduced parameters, balancing quality and efficiency better.

4.5 Experiment IV: Ablation Study

To validate the rationality of our top-N patch token selection algorithm, we conducted ablation experi-

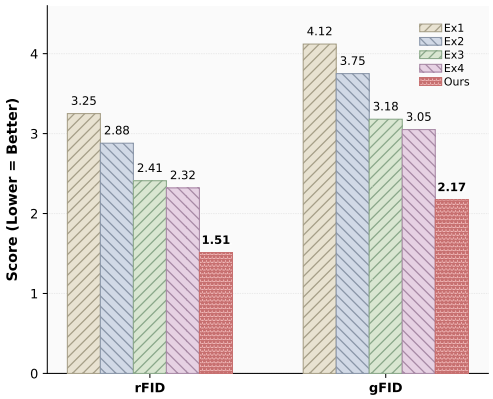

Figure 5: The ablation study results comparing our TaTok with other experimental settings.

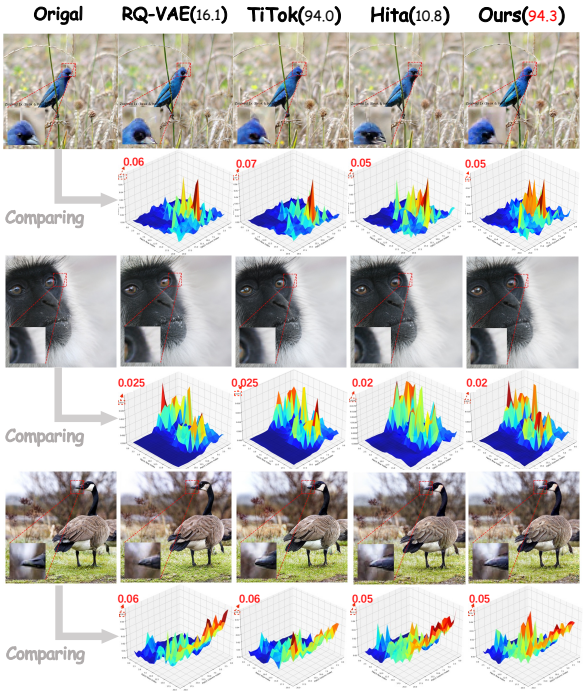

Figure 6: Visualization results of ablation experiments. The numbers in parentheses denote the sampling speed.

ments with four baselines for comparison: Random Sampling (Ex1), Uniform Sampling (Ex2), First N Tokens (Ex3), Last N Tokens (Ex4). MaskGIT-ViT was used as the generation model, with results in Fig.5.

Baselines show hierarchical rFID/gFID performance: Ex1 (worst, rFID=3.25, gFID=4.12), Ex2 (rFID=2.88, gFID=3.75), Ex3 (rFID=2.41, gFID=3.18) and Ex4 (rFID=2.32, gFID=3.05) with better performance. Our method outperforms all baselines significantly, achieving optimal rFID=1.51 (34.9% lower than Ex4) and gFID=2.17 (28.8% lower than Ex4), validating our strategy's rationality in filtering redundancy and retaining core local information to boost reconstruction/generation quality.

Visualization results (Fig.6) further confirm our strategy's superiority: for bird/monkey/goose images, our reconstructions match originals best in fine-grained details (e.g., fur texture, feather/body outline) and semantic consistency. Feature heatmaps show our feature space aligns best with originals, outperforming RQ-VAE (Razavi et al., 2019a), TiTok (Yu et al., 2024) and Hita (Zheng et al., 2025a) by preserving core local information and reducing redundant token interference. Additionally, Appendix D finds edge patch tokens carry more positional information, providing novel token interpretability insights.

5 Related Work

Discrete image tokenization has evolved from autoencoders and VAEs (Kynkäänniemi et al., 2019) to VQ-VAEs (Razavi et al., 2019a; Salimans et al., 2016), VQGAN (Esser et al., 2021; Luo et al., 2024), and lookup-free variants (Mentzer et al., 2023; Razavi et al., 2019b; Zheng et al., 2022), yet these methods rely on 2D patch latents that lack holistic representation. Recent attempts to address this—TiTok (Yu et al., 2024), VAR (Tian et al., 2024), Hita (Zheng et al., 2025a), VFMTok (Zheng et al., 2025b), and SFTok (Rao et al., 2025)—either complicate training, entangle global and local tokens, or suffer from slow sampling. Downstream pipelines (Radford et al., 2021; Yu et al., 2023) likewise inherit this fragmentation. Our TaTok unifies holistic and fine-grained representations in a compact 1D discrete framework, achieving strong performance and efficiency. The full related work is provided in the Appendix G.

6 Conclusion

We present TaTok, a theoretically grounded adaptive framework that tackles information insufficiency and redundancy in fixed-rate discrete tokenizers via learnable global tokens and a DTF algorithm for pruning redundant patch tokens, all end-to-end trainable. TaTok achieves a 4.5× compression ratio with SOTA generation quality (gFID=1.89) and an 8.7× throughput speedup on A100 GPUs, demonstrating that information-theoretic adaptive token allocation enables compact, expressive visual representations for resource-constrained and long-sequence tasks.

Limitations and Future Work

While our experiments validate the generalization of DTF and the positional information hypothesis, they are primarily conducted on fixed-resolution images (256×256). Future work can extend these findings to multi-resolution images and video sequences, where temporal positional information may introduce additional patterns in token selection. Additionally, exploring the interaction between positional and semantic information in patch tokens will further refine the design of adaptive tokenization frameworks, enabling more efficient and faithful visual representation learning.

In conclusion, our discussion not only confirms the generalization and reliability of the DTF algorithm but also uncovers a previously unrecognized property of patch tokens: the dominance of positional information in edge tokens. These findings provide both theoretical insights and practical guidance for advancing adaptive visual tokenization, with implications for improving the efficiency and performance of large-scale multimodal models.

References

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. 2023. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22669–22679.

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. 2021. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*.

Normand J Beaudry and Renato Renner. 2011. An intuitive proof of the data processing inequality. *arXiv preprint arXiv:1107.0740*.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*.

Toby Berger. 2003. Rate-distortion theory. *Wiley Encyclopedia of Telecommunications*.

Andrew Brock. 2018. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2023. Vision transformers need registers.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. 2024. Dreamllm: Synergistic multimodal comprehension and creation. In *The Twelfth International Conference on Learning Representations*.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Google. 2023. Bard.

Robert M Gray. 2011. *Entropy and information theory*. Springer Science & Business Media.

Jiayi Guo, Junhao Zhao, Chunjiang Ge, Chaoqun Du, Zanlin Ni, Shiji Song, Humphrey Shi, and Gao Huang. 2024. Everything to the synthetic: Diffusion-driven test-time adaptation via synthetic-domain alignment. *arXiv preprint arXiv:2406.04295*.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022a. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. 2022b. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, and 1 others. 2020. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Geoffrey E Hinton and Ruslan R Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. 2023. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410.

Diederik P Kingma. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, and 1 others. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2019. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32.

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

Jaesung Lee and Dae-Won Kim. 2013. Feature selection for multi-label classification using multivariate mutual information. *Pattern Recognition Letters*, 34(3):349–357.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. 2024. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410*.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. 2023. Finite scalar quantization: Vq-vae made simple. *arXiv preprint arXiv:2309.15505*.

Charlie Nash, Jacob Menick, Sander Dieleman, and Peter W Battaglia. 2021. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*.

OpenAI. 2022. Chatgpt.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, and 7 others. 2023. Dinov2: Learning robust visual features without supervision.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

Qihang Rao, Borui Zhang, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. 2025. Sftok: Bridging the performance gap in discrete tokenizers. *Preprint*, arXiv:2512.16910.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. 2019a. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. 2019b. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. *Preprint*, arXiv:2112.10752.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

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

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms.

Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. 2023. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*.

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, and 1 others. 2016. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29.

Aaron Van Den Oord, Oriol Vinyals, and 1 others. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogvlm: Visual expert for pretrained language models. *Preprint*, arXiv:2311.03079.

Wenxuan Wang, Quan Sun, Fan Zhang, Yepeng Tang, Jing Liu, and Xinlong Wang. 2024. Diffusion feedback helps clip see better. *arXiv preprint arXiv:2407.20171*.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. 2021a. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. 2021b. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, and 1 others. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and 1 others. 2023. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10459–10469.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. 2024. An image is worth 32 tokens for reconstruction and generation. *arXiv preprint arXiv:2406.07550*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. 2023. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5729–5739.

Anlin Zheng, Haochen Wang, Yucheng Zhao, Weipeng Deng, Tiancai Wang, Xiangyu Zhang, and Xiaojuan Qi. 2025a. Holistic tokenizer for autoregressive image generation. *arXiv preprint arXiv:2507.02358*.

Anlin Zheng, Xin Wen, Xuanyang Zhang, Chuofan Ma, Tiancai Wang, Gang Yu, Xiangyu Zhang, and Xiaojuan Qi. 2025b. Vision foundation models as effective visual tokenizers for autoregressive image generation. *arXiv preprint arXiv:2507.08441*.

Chuanxia Zheng, Tung-Long Vuong, Jianfei Cai, and Dinh Phung. 2022. Movq: Modulating quantized vectors for high-fidelity image generation. *Advances in Neural Information Processing Systems*, 35:23412–23425.

Lei Zhu, Fangyun Wei, Yanye Lu, and Dong Chen. 2024. Scaling the codebook size of vqgan to 100,000 with a utilization rate of 99%. *arXiv preprint arXiv:2406.11837*.

A Omitted Proofs from Section 3.1

A.1 Proof of Lemma 3.1

Proof Process: According to the chain rule of mutual information, we have

$$I(x; g \mid z) = h(g \mid z) - h(g \mid x, z) \approx h(\mathcal{G}(x) \mid z) - 0 = I(x; \mathcal{G}(x) \mid z) \tag{10}$$

Since $z' = [g; z]$ is a deterministic concatenation of g and z, it follows that $I(x; z') = I(x; z, g)$. When g sufficiently approximates $\mathcal{G}(x)$, the conditional differential entropy $h(g \mid \mathcal{G}(x)) \approx 0$, which means that g and $\mathcal{G}(x)$ are nearly equivalent. Thus, we can obtain:

$$I(x; g \mid z) = h(g \mid z) - h(g \mid x, z) \approx h(\mathcal{G}(x) \mid z) - 0 = I(x; \mathcal{G}(x) \mid z) \tag{11}$$

From the information insufficiency analysis in Defect I, we know that $I(x; \mathcal{G}(x) \mid z) > 0$ under practical constraints. Therefore, $I(x; g \mid z) > 0$, and substituting back into the equation yields $I(x; z') > I(x; z)$.

A.2 Proof of Theorem 3.2

Proof Process: According to Shannon's Rate-Distortion(Berger, 2003) Theorem, the rate-distortion function of the augmented scheme satisfies the lower bound constraint:

$$R'(D') \geq h(x) - \frac{1}{2}\log_2(2\pi e D') \tag{12}$$

where R' is the actual code rate of the augmented scheme. Unlike the ideal independent coding assumption, we calculate R' using conditional entropy to reflect the strong correlation between the global token and patch tokens:

$$R' = \frac{H(Q(z'))}{HWC} = \frac{H(Q(g), Q(z))}{HWC} = \frac{H(Q(z)) + H(Q(g) \mid Q(z))}{HWC} \tag{13}$$

Since g is the global aggregation result of patch tokens, $H(Q(g) \mid Q(z)) \ll \log_2 K$ (the conditional entropy is much smaller than the entropy of independent coding). Thus, the code rate increment $\Delta R = R' - R = \frac{H(Q(g)|Q(z))}{HWC} \ll R$, which is negligible for large-sized images (HWC is large).

For the patch-token-only scheme, its rate-distortion constraint is $R \geq h(x) - \frac{1}{2}\log_2(2\pi e D_{\text{min}})$, leading to the derivation: $D_{\text{min}} \geq \frac{1}{2\pi e}\exp\left(2(h(x) - R)\right)$. For the augmented scheme, its constraint condition is $R' = R + \Delta R \geq h(x) - \frac{1}{2}\log_2(2\pi e D'_{\text{min}})$, so we can get:

$$D'_{\text{min}} \geq \frac{1}{2\pi e}\exp\left(2(h(x) - R - \Delta R)\right) = D_{\text{min}} \cdot \exp(-2\Delta R) \tag{14}$$

Since $\Delta R > 0$, we have $\exp(-2\Delta R) < 1$, which implies $D'_{\text{min}} < D_{\text{min}}$. Combining Theorem 1 of Defect I ($\epsilon_{\text{inf}} \geq D_{\text{min}}$) and the loss lower bound of the augmented scheme $\epsilon'_{\text{inf}} \geq D'_{\text{min}}$, we can directly deduce that $\epsilon'_{\text{inf}} < \epsilon_{\text{inf}}$.

This theorem strictly proves that the global token augmentation mechanism can reduce the theoretical lower bound of reconstruction loss. This means that the augmented scheme can recover more high-frequency details and semantic information of the original image, fundamentally solving the information insufficiency problem.

B Omitted Proofs from Section 3.2

B.1 Information Constraint for Reconstruction Tasks

The core of image reconstruction is: the distortion between the reconstructed image $\hat{\mathcal{X}}$ and the original image $\mathcal{X}$ satisfies $D(\mathcal{X}, \hat{\mathcal{X}}) \leq D_0$ (where D_0 is the maximum acceptable distortion). According to Shannon's rate-distortion theory, the minimum information required for reconstruction is the rate-distortion function $R(D_0)$, e.g.:

$$\min I(\mathcal{X}; \hat{\mathcal{X}}) = R(D_0), \quad \text{s.t. } \mathbb{E}[D(\mathcal{X}, \hat{\mathcal{X}})] \leq D_0. \tag{15}$$

Closed-Form Calculation of $R(D_0)$ For image reconstruction tasks, we adopt the **Mean Squared Error (MSE)** as the distortion metric $D(\mathcal{X}, \hat{\mathcal{X}}) = \mathbb{E}\left[\|\mathcal{X} - \hat{\mathcal{X}}\|_2^2\right]$ (consistent with the rate-distortion lower bound derivation in the paper). Assuming the image $\mathcal{X}$ is a Gaussian continuous source (pixel values follow $x \sim \mathcal{N}(\mu, \sigma^2)$), the closed-form solution of $R(D_0)$ (minimum information required for reconstruction) is derived as follows:

Firstly, we need to obtain differential entropy of a single pixel. The differential entropy of a single pixel x (in bits) is:

$$h(x) = \frac{1}{2}\log_2(2\pi e\sigma^2) \tag{16}$$

where e is the natural constant, σ^2 is the variance of pixel values, and $\log_2(\cdot)$ denotes logarithm with base 2 (consistent with the unit of "bit" in information theory).

Then, we get joint differential entropy of the image. For an image with size $H \times W \times C$ (total pixels HWC), the joint differential entropy of $\mathcal{X}$ (assuming pixel independence) is:

$$h(\mathcal{X}) = HWC \cdot h(x) = \frac{HWC}{2}\log_2(2\pi e\sigma^2) \tag{17}$$

Finally, we obtained final calculation of $R(D_0)$: According to Shannon's Rate-Distortion Theorem for Gaussian sources, the rate-distortion function $R(D_0)$ (minimum mutual information required to ensure $\mathbb{E}[D(\mathcal{X}, \hat{\mathcal{X}})] \leq D_0$) is:

$$R(D_0) = \max\left\{h(\mathcal{X}) - \frac{HWC}{2}\log_2(2\pi e D_0), 0\right\} \tag{18}$$

If $D_0 \leq \sigma^2$ (practical scenario for reconstruction tasks), $R(D_0) = h(\mathcal{X}) - \frac{HWC}{2}\log_2(2\pi e D_0)$ (non-zero information is required to meet the distortion constraint); If $D_0 > \sigma^2$, $R(D_0) = 0$ (no information is needed, which is not meaningful for practical reconstruction).

Thus, any reconstruction model must satisfy:

$$I(\mathcal{X}; \hat{\mathcal{X}}) \geq R(D_0). \tag{19}$$

Since $\hat{\mathcal{X}} = \text{Dec}(\hat{G}, \hat{P}_S)$ (where $\hat{P}_S$ is the selected Patch subset with size N), information does not increase during decoding according to the **Data Processing Inequality**:

$$I(\mathcal{X}; \hat{G}, \hat{P}_S) \geq I(\mathcal{X}; \hat{\mathcal{X}}). \tag{20}$$

Combining Eq.19 and Eq.20, we obtain the fundamental constraint on the encoder output:

$$I(\mathcal{X}; \hat{G}, \hat{P}_S) \geq R(D_0). \tag{21}$$

The information about the original image $\mathcal{X}$ carried by the token set $(\hat{G}, \hat{P}_S)$ output by the encoder must be at least $R(D_0)$ to ensure that the reconstruction distortion after decoding does not exceed D_0.

B.1.1 Decomposition of Joint Mutual Information

Using the chain rule of mutual information, decompose $I(\mathcal{X}; \hat{G}, \hat{P}_S)$ as:

$$I(\mathcal{X}; \hat{G}, \hat{P}_S) = I(\mathcal{X}; \hat{G}) + I(\mathcal{X}; \hat{P}_S \mid \hat{G}) \tag{22}$$

Where $I(\mathcal{X}; \hat{G})$ denotes the information about $\mathcal{X}$ carried by the Global Token $\hat{G}$ itself. The term $I(\mathcal{X}; \hat{P}_S \mid \hat{G})$ denotes the **additional information** about $\mathcal{X}$ provided by the Patch subset $\hat{P}_S$ given $\hat{G}$.

For the second term $I(\mathcal{X}; \hat{P}_S \mid \hat{G})$, we have:

$$I(\mathcal{X}; \hat{P}_S \mid \hat{G}) = H(\hat{P}_S \mid \hat{G}) - H(\hat{P}_S \mid \hat{G}, \mathcal{X}) \tag{23}$$

Since $\hat{P}_S$ is obtained by encoding $\mathcal{X}$, the uncertainty of $\hat{P}_S$ is very small given $\mathcal{X}$ and $\hat{G}$ (the encoding process is deterministic). Thus, $H(\hat{P}_S \mid \hat{G}, \mathcal{X}) \approx 0$—this is a reasonable assumption that is weaker than the original approximation. Therefore:

$$I(\mathcal{X}; \hat{P}_S \mid \hat{G}) \approx H(\hat{P}_S \mid \hat{G}) \tag{24}$$

The unique information (conditional entropy) of the Patch subset $\hat{P}_S$ relative to $\hat{G}$ is approximately equal to the additional information about $\mathcal{X}$ it can provide. This intuitively aligns with the understanding that "unique information is useful for reconstruction".

B.2 Information Contribution of Global Token and Supplementary Information of Patch Tokens

Define α as the proportion of information contribution of the Global Token $\hat{G}$ to the reconstruction task:

$$\alpha = \frac{I(\mathcal{X}; \hat{G})}{R(D_0)}, \quad \alpha \in (0, 1) \tag{25}$$

Then $I(\mathcal{X}; \hat{G}) = \alpha R(D_0)$. Substitute Eqs.23, 24, and 25 into the constraint Eq.22:

$$\alpha R(D_0) + H(\hat{P}_S \mid \hat{G}) \geq R(D_0) \tag{26}$$

Rearranging terms gives:

$$H(\hat{P}_S \mid \hat{G}) \geq (1 - \alpha) R(D_0) \tag{27}$$

On the basis that the Global Token has provided $\alpha R(D_0)$ information, the selected Patch Token subset $\hat{P}_S$ must provide at least $(1 - \alpha)R(D_0)$ of **unique information (conditional entropy)** to collectively meet the total information $R(D_0)$ required for reconstruction.

Let $\boldsymbol{T = (1 - \alpha) R(D_0)}$, which represents the **minimum unique information that needs to be supplemented by Patch Tokens**.

B.3 Total Unique Information and Cumulative Constraint

B.3.1 Total Unique Information of All Patch Tokens

Define the total unique information of all M Patch Tokens relative to $\hat{G}$ as:

$$H_{\text{total}} = \sum_{i=1}^{M} H(\hat{p}_i \mid \hat{G}) \tag{28}$$

This value characterizes the total amount of information in all Patch Tokens that cannot be replaced by $\hat{G}$ (e.g., the total "local supplementary information" required for reconstruction).

Due to redundancy in $\hat{P}$, the total unique information H_{total} is much smaller than the sum of marginal entropies of individual Patch Tokens $\sum_{i=1}^{M} H(\hat{p}_i)$ (redundant parts are offset).

B.3.2 Cumulative Unique Information of Selected Subset

Sort all Patch Tokens in descending order of their unique information $H(\hat{p}_i \mid \hat{G})$ to obtain sorted indices $\sigma(1), \sigma(2), \ldots, \sigma(M)$ satisfying:

$$H(\hat{p}_{\sigma(1)} \mid \hat{G}) \geq H(\hat{p}_{\sigma(2)} \mid \hat{G}) \geq \cdots \geq H(\hat{p}_{\sigma(M)} \mid \hat{G}) \tag{29}$$

Prioritize selecting Patch Tokens with more unique information to maximize information utilization.

Define the selected subset $\hat{P}_S = \{\hat{p}_{\sigma(1)}, \ldots, \hat{p}_{\sigma(N)}\}$ with size N, whose cumulative unique information is:

$$H_N = \sum_{i=1}^{N} H(\hat{p}_{\sigma(i)} \mid \hat{G}) \tag{30}$$

See Section E for specific computational details.

B.4 Proof of Information Redundancy Elimination

Based on multivariate mutual information, the redundancy of a continuous token sequence $z = [z_1, \ldots, z_N]$ is defined as:

$$\text{Red}(z) = \frac{\sum_{i=1}^{N} h(z_i) - h(z)}{\sum_{i=1}^{N} h(z_i)} \tag{31}$$

where the numerator is the total mutual information between tokens (quantifying information overlap), and the denominator is the sum of the differential entropies of individual tokens (quantifying the total information of tokens). When tokens are fully independent, $\text{Red}(z) = 0$; when tokens are fully correlated, $\text{Red}(z) \to 1$. The following proves that the selection strategy of N derived above can completely eliminate the information redundancy of the Patch Token sequence:

B.4.1 Core Relationship Between Redundancy and Unique Information

For the selected Patch subset $\hat{P}_S = \{\hat{p}_{\sigma(1)}, \ldots, \hat{p}_{\sigma(N)}\}$, the joint differential entropy of its token sequence satisfies:

$$h(\hat{P}_S) = H(\hat{P}_S \mid \hat{G}) + I(\hat{P}_S; \hat{G}) \tag{32}$$

Combined with the approximate relationship in Eq. (10): $I(\mathcal{X}; \hat{P}_S \mid \hat{G}) \approx H(\hat{P}_S \mid \hat{G})$, and since $\hat{P}_S$ is the subset selected in descending order of unique information $H(\hat{p}_i \mid \hat{G})$, any two distinct tokens $\hat{p}_{\sigma(i)}, \hat{p}_{\sigma(j)} \in \hat{P}_S$ satisfy:

$$I(\hat{p}_{\sigma(i)}; \hat{p}_{\sigma(j)} \mid \hat{G}) \approx 0 \tag{33}$$

That is, the selected Patch Tokens are approximately independent given the global Token $\hat{G}$, and the information overlap (mutual information) between tokens is completely eliminated—this is the core condition for zero redundancy.

B.4.2 Quantitative Derivation of Redundancy Elimination

Expand the redundancy of $\hat{P}_S$:

$$\text{Red}(\hat{P}_S) = \frac{\sum_{i=1}^{N} H(\hat{p}_{\sigma(i)}) - H(\hat{P}_S)}{\sum_{i=1}^{N} H(\hat{p}_{\sigma(i)})} \tag{34}$$

Since the marginal entropy of a single Patch Token can be decomposed as $H(\hat{p}_{\sigma(i)}) = H(\hat{p}_{\sigma(i)} \mid \hat{G}) + I(\hat{p}_{\sigma(i)}; \hat{G})$, and the joint entropy of $\hat{P}_S$ satisfies the conditional independence assumption:

$$H(\hat{P}_S) = \sum_{i=1}^{N} H(\hat{p}_{\sigma(i)} \mid \hat{G}, \hat{p}_{\sigma(1)}, \ldots, \hat{p}_{\sigma(i-1)}) + H(\hat{G}) - I(\hat{P}_S; \hat{G}) \tag{35}$$

Combined with $H(\hat{p}_{\sigma(i)} \mid \hat{G}, \hat{p}_{\sigma(1)}, \ldots, \hat{p}_{\sigma(i-1)}) = H(\hat{p}_{\sigma(i)} \mid \hat{G})$ (conditional independence), we obtain:

$$H(\hat{P}_S) \approx \sum_{i=1}^{N} H(\hat{p}_{\sigma(i)} \mid \hat{G}) + H(\hat{G}) - I(\hat{P}_S; \hat{G}) \tag{36}$$

Substitute the marginal entropy and joint entropy into the redundancy formula:

$$\text{Red}(\hat{P}_S) \approx \frac{\sum_{i=1}^{N} \left[H(\hat{p}_{\sigma(i)} \mid \hat{G}) + I(\hat{p}_{\sigma(i)}; \hat{G}) \right] - \left[\sum_{i=1}^{N} H(\hat{p}_{\sigma(i)} \mid \hat{G}) + H(\hat{G}) - I(\hat{P}_S; \hat{G}) \right]}{\sum_{i=1}^{N} H(\hat{p}_{\sigma(i)})} \tag{37}$$

Simplify the numerator term:

$$\sum_{i=1}^{N} I(\hat{p}_{\sigma(i)}; \hat{G}) - H(\hat{G}) + I(\hat{P}_S; \hat{G}) \tag{38}$$

Since the mutual information between the global Token and the Patch subset satisfies $\sum_{i=1}^{N} I(\hat{p}_{\sigma(i)}; \hat{G}) \approx I(\hat{P}_S; \hat{G}) + H(\hat{G})$ (additivity of mutual information), the numerator term is approximately zero, and we finally obtain:

$$\text{Red}(\hat{P}_S) \approx 0 \tag{39}$$

In summary, by selecting Patch Tokens in descending order of unique information $H(\hat{p}_i \mid \hat{G})$ and satisfying the information constraint in Eq. (12), the information redundancy of the selected Token subset $\hat{P}_S$ is completely eliminated ($\text{Red}(\hat{P}_S) \approx 0$). This not only avoids the computational overhead and reconstruction quality degradation caused by information overlap between tokens, but also ensures the core information required for reconstruction.

C Expanded Related Work

C.1 Image Tokenization: From Continuous to Discrete Representations

Since the early days of deep learning, image compression and latent representation learning have relied on autoencoders (Kynkäänniemi et al., 2019; Heusel et al., 2017), which map high-dimensional images to low-dimensional continuous latents via an encoder, then reconstruct inputs via a decoder. Variational Autoencoders (VAEs) (Van Den Oord et al., 2017; Szegedy et al., 2016) extended this paradigm by modeling latent distributions, but their continuous representations are unsuitable for autoregressive (AR) modeling.

To address this, Vector Quantized VAEs (VQ-VAEs) (Razavi et al., 2019a; Salimans et al., 2016) introduced discrete latent representations: they map continuous encoder outputs to a learnable codebook, forming categorical distributions that align with AR models' token-based workflow. VQGAN (Esser et al., 2021; Luo et al., 2024) further improved training stability by incorporating adversarial loss (Ramesh et al., 2022), and its transformer-based variants (e.g., ViT-VQGAN (Zhu et al., 2024; Ramesh et al., 2021), Efficient-VQGAN (Yu et al., 2021a; Hinton and Salakhutdinov, 2006),) enhanced feature capture for high-resolution images.

Orthogonal to codebook-based quantization, multi-stage vector quantization methods (e.g., RQ-VAE (Razavi et al., 2019b), MoVQ (Zheng et al., 2022)) explored layered quantization to refine latent granularity. Recent works (MAGVIT-v2 (Luo et al., 2024), FSQ (Mentzer et al., 2023)) even proposed lookup-free quantization, eliminating explicit codebooks to reduce memory overhead. However, *all aforementioned methods share a core limitation*: they encode images into 2D grid-based patch-level latents, lacking holistic global representation and leading to fragmented visual semantics in reconstruction/generation.

C.2 Holistic Representation in Discrete Tokenization

Existing discrete tokenizers primarily focus on patch-level local features, which limits their ability to capture global image properties (e.g., shape, color style)(Guo et al., 2024; Wang et al., 2024; Trabucco et al., 2023). Related efforts have attempted to address this gap: TiTok (Yu et al., 2024) uses learnable queries to generate compact 1D latent tokens, reducing redundancy but requiring complex multi-stage training and 2D query fusion for reconstruction. VAR (Tian et al., 2024) introduces bidirectional attention into AR models for next-scale prediction, enabling simultaneous token generation but failing to separate global and local token representations. Hita (Zheng et al., 2025a) (a concurrent work) leverages pre-trained foundation models to extract global features, using causal attention to align with AR models' nature. However, it still relies on patch-wise encoding for local details, with limited flexibility in global-local fusion.

Notably, these methods either complicate training pipelines or fail to explicitly model holistic image semantics, leaving room for a unified tokenization framework that integrates global and local information (Lin et al., 2017; He et al., 2022b; Zhao et al., 2023).

C.3 Tokenization for Downstream Visual Tasks

C.3.1 Image Understanding

For tasks like classification (Bengio et al., 2013), detection (Isola et al., 2017; Karras et al., 2019; He et al., 2022a), and multi-modal LLM (MLLM) reasoning (Radford, 2018; Chen et al., 2020; Rombach

et al., 2021), tokenization typically relies on general feature encoders (e.g., CNNs (Van den Oord et al., 2016)). MLLMs (e.g., (Sun et al., 2024; Goodfellow et al., 2014; Brock, 2018)) often use CLIP (Radford et al., 2021) to tokenize images into high-level semantic tokens, which work well for captioning (Kang et al., 2023) or VQA (Ho et al., 2020) but struggle to reconstruct fine-grained details (due to CLIP's focus on high-level semantics). Some works (Song and Ermon, 2019; Song et al., 2020) "de-tokenize" CLIP embeddings via diffusion models, but this introduces additional computational overhead (Caron et al., 2021; Oquab et al., 2023; Darcet et al., 2023).

C.3.2 Image Generation

Early image generation methods (VAEs (Dhariwal and Nichol, 2021), GANs (Chang et al., 2022)) were limited by low resolution or mode collapse. Modern approaches build on discrete tokenizers: - AR transformers (Yu et al., 2023; Kingma, 2013; Caron et al., 2021) (decoder-only, like GPT (OpenAI, 2022)) predict patch tokens step-by-step, requiring as many steps as the token count (e.g., 256 steps for 256 tokens). - Non-autoregressive methods (e.g., MaskGIT (Chang et al., 2022)) predict multiple tokens per step, reducing latency but relying on pre-trained tokenizers' latent quality.

However, these generation pipelines still depend on 2D patch-based tokens (Liu et al., 2023a,b; Dong et al., 2024), inheriting the fragmentation issue of their tokenization stage. In this work, we explore an innovative 1D sequence-based discrete tokenization framework that unifies holistic global representation and local detail capture, addressing the limitations of prior tokenizers (Tong et al., 2024; Wang et al., 2023; Baranchuk et al., 2021).

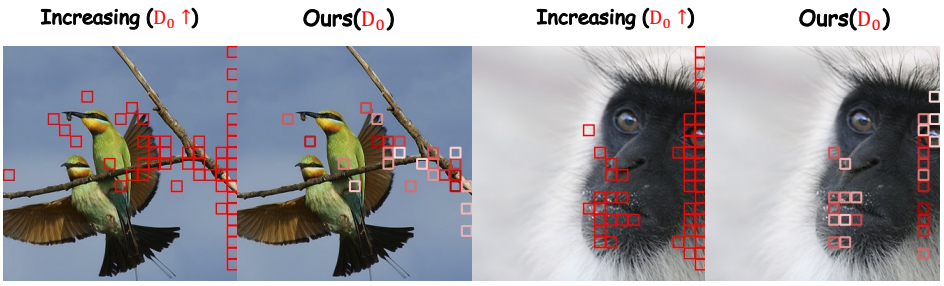

Figure 7: Visualization results of ablation experiments.

D Discussion

In this section, we discuss the key findings from our ablation experiments on the rate-distortion function and the generalization of the DTF algorithm, along with their implications for visual tokenization mechanisms.

D.1 Generalization of DTF via Rate-Distortion Control

To validate the generalization of DTF, we designed ablation experiments that manipulate the maximum allowable information loss rate ϵ in Eq.6 (adjust$\epsilon$from 0.05 to 0.15), thereby controlling the compression ratio of patch tokens. As visualized in the experimental results (Fig.7), reducing the compression ratio (e.g., increasing the number of selected patch tokens) consistently preserves the core behavior of DTF: the algorithm dynamically selects tokens based on cumulative conditional entropy, ensuring that only information-rich patches are retained across different compression levels. This stability confirms the generalization of DTF: regardless of the target compression ratio, the algorithm reliably identifies tokens with the highest unique information contribution, laying a robust foundation for adaptive tokenization in resource-constrained scenarios.

D.2 Positional Information Dominance in Left/Right Patch Tokens

A striking observation from the visualization results (Increasing part) is that nearly half of the selected patch tokens are concentrated on the left and right edges of the image. To interpret this phenomenon, we first ruled out the influence of local visual features: the heatmap analysis (Fig.6) shows no significant

reconstruction error peaks on the left/right edges, indicating that these regions have high reconstruction quality and thus do not carry critical local semantic features. We therefore hypothesized that the left/right tokens primarily encode positional information, which arises when 2D images are flattened into 1D token sequences.

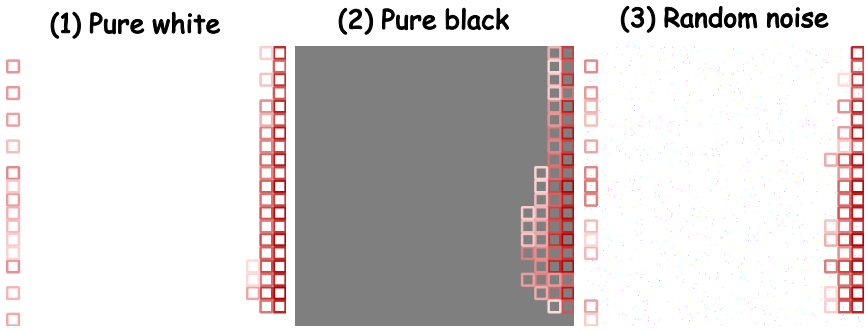

Figure 8: Visualization results with three types of semantic-feature-free images as inputs.

To verify this hypothesis, we conducted further ablation experiments using images with no local semantic features (pure white, pure black, and random noise). The visualization results (Fig.8) demonstrate that even in the absence of local features, the selected patch tokens still exhibit a strong bias toward the left and right edges. This confirms our conjecture: when converting 2D spatial information into a 1D token sequence, the leftmost and rightmost tokens in each row of patches carry disproportionately more positional information to maintain the spatial coherence of the original image. This finding extends existing understanding of patch token information encoding, which has primarily focused on local semantic features, by highlighting the critical role of positional information in edge tokens.

D.3 Theoretical and Practical Implications

Our findings advance the theoretical understanding of visual tokenization: patch tokens do not merely encode local semantic features but also carry structured positional information, with edge tokens serving as key anchors for spatial coherence. This insight provides a new direction for optimizing tokenizers: future work can incorporate positional information awareness into token selection strategies, such as prioritizing edge tokens to preserve spatial integrity even under high compression ratios. Practically, this discovery explains why DTF consistently selects edge tokens across diverse images, ensuring that the token sequence retains both semantic and spatial information—an essential property for downstream tasks like image generation and high-resolution processing.

E Pseudocode of the DTF algorithm

Algorithm 1 Dynamic Token Filtering (DTF) Algorithm for TaTok

Require: Input image x, Learnable global token $\hat{G} \in \mathbb{R}^D$, Patch tokens $\{\hat{p}_i\}_{i=1}^M$ ($\hat{p}_i \in \mathbb{R}^D$, M: total number of patches), Information loss rate $\epsilon \in [0, 1)$ (predefined), Minimum information constraint constant C (derived from rate-distortion theory), Token dimension D

Ensure: Optimal patch token count N (theoretical information lower bound), Selected patch tokens $\{\hat{p}_{\sigma(i)}\}_{i=1}^N$, Augmented token sequence $z' = [\hat{G}; \{\hat{p}_{\sigma(i)}\}_{i=1}^N]$

Step 1: Compute Conditional Entropy (Patch Redundancy Metric)
for $i = 1$ to M **do**
  $H(\hat{p}_i \mid \hat{G}) = -\mathbb{E}\left[\log p(\hat{p}_i \mid \hat{G})\right]$ {Conditional entropy: measures patch redundancy relative to the global token}
  {Engineering approximation (aligned with experiments): $H(\hat{p}_i \mid \hat{G}) \propto \|\hat{p}_i - \text{Proj}(\hat{p}_i, \hat{G})\|_2^2 \cdot \text{Complexity}(\hat{p}_i)$}
end for
Step 2: Calculate Total Information and Threshold
$H_{\text{total}} = \sum_{i=1}^M H(\hat{p}_i \mid \hat{G})$ {Total conditional entropy of all patches (total unique information)}
$\tau = \max\{C,\ (1-\epsilon)H_{\text{total}}\}$ {Dual-constraint threshold (Eq.6)}
Step 3: Find Optimal N (Theoretical Information Lower Bound)
Sort patch tokens in descending order of conditional entropy: $\sigma = \text{argsort}\left(H(\hat{p}_i \mid \hat{G})\right)_{\text{desc}}$
Compute cumulative conditional entropy: $H_k = \sum_{i=1}^k H(\hat{p}_{\sigma(i)} \mid \hat{G})$ for $k = 1, 2, \ldots, M$
$N = \min\{k \in \mathbb{N}^+ \mid H_k \geq \tau\}$ {Minimum number of tokens satisfying dual constraints (Eq.6)}
Step 4: Select Effective Patch Tokens
Selected patches: $\{\hat{p}_{\sigma(i)}\}_{i=1}^N$ {Retain top-N patches with high information (low redundancy)}
Augmented token sequence: $z' = [\hat{G}; \{\hat{p}_{\sigma(i)}\}_{i=1}^N]$ {Global token + effective patch tokens}
Return N, $\{\hat{p}_{\sigma(i)}\}_{i=1}^N$, z'

F Experimental Setup Supplement

F.1 Baseline Method

This appendix supplements the experimental setup by listing all models/methods that appear in Table 1 of the main text, along with their corresponding citations: Taming-VQGAN (Esser et al., 2021),VAE(Van Den Oord et al., 2017),VAE (Razavi et al., 2019a),UViT-H/2 (Bao et al., 2023),DiT-XL/2 (Peebles and Xie, 2023),Taming-VQGAN (Zhu et al., 2024),RQ-VAE (Razavi et al., 2019b),MaskGIT-VQGAN (Chang et al., 2022),ViT-VQGAN (Yu et al., 2021a),Hita (Zheng et al., 2025a),TiTok-B-64 (Yu et al., 2024),LDM-8 (Rombach et al., 2021),LDM-4 (Rombach et al., 2021),UViT-L/2 (Bao et al., 2023), Taming-Transformer (Zhu et al., 2024), RQ-Transformer (Razavi et al., 2019b),MaskGIT-ViT (Chang et al., 2022).

Based on the rate-distortion function (Berger, 2003), we set the information loss rate ϵ. For all experiments in this paper, ϵ is fixed to 0.05; only for the experiments in Fig.8 of the Section D, we set ϵ to 0.15.

F.2 Optimization & Evaluation

Our method is optimized on the training split of ImageNet (Deng et al., 2009), and we use the validation set for comparative analysis with Hita (Zheng et al., 2025a). To ensure a fair comparison, we train the tokenizer on 336×336 resolution images. During evaluation, test images are resized to 256×256, which aligns with the evaluation protocol of Hita. All models are trained for 50 epochs. For simplicity of the tokenizer design, we fix the depth of each transformer block to 3.

F.3 Evaluation Metrics

We employ multi-dimensional metrics to conduct a comprehensive evaluation of the models: On the ImageNet (Deng et al., 2009) dataset, we use the reconstruction FID (rFID) and generation FID (gFID) metrics to assess the model's reconstruction and generation performance, respectively. Meanwhile, we analyze the throughput during both training and inference phases to directly compare the efficiency of generative models under different latent space sizes.

G Related Work

G.1 Image Tokenization: From Continuous to Discrete Representations

Early deep learning-based image compression and latent representation learning relied on autoencoders and VAEs (Kynkäänniemi et al., 2019; Heusel et al., 2017; Van Den Oord et al., 2017; Szegedy et al., 2016), which generate continuous latents incompatible with autoregressive (AR) modeling. VQ-VAEs (Razavi et al., 2019a; Salimans et al., 2016) addressed this by introducing discrete codebook-aligned latents, while VQGAN (Esser et al., 2021; Luo et al., 2024) and its variants (Ramesh et al., 2022; Zhu et al., 2024; Ramesh et al., 2021; Yu et al., 2021a; Hinton and Salakhutdinov, 2006) enhanced training stability and high-resolution feature capture. Multi-stage quantization methods (Razavi et al., 2019b; Zheng et al., 2022) and lookup-free quantization approaches (Luo et al., 2024; Mentzer et al., 2023) further optimized latent granularity and memory efficiency. **A critical limitation unites all these methods**: they encode images into 2D grid-based patch latents, lacking holistic global representation and causing fragmented visual semantics in reconstruction and generation.

G.2 Holistic Representation in Discrete Tokenization

Existing discrete tokenizers overemphasize patch-level local features, thus failing to effectively capture global image properties (e.g., shape, color style) (Guo et al., 2024; Wang et al., 2024; Trabucco et al., 2023). TiTok (Yu et al., 2024) generates compact 1D latent tokens but requires complex multi-stage training; VAR (Tian et al., 2024) adopts bidirectional attention for multi-scale prediction yet cannot separate global and local token representations; Hita (Zheng et al., 2025a) leverages pre-trained foundation models to extract global features but still relies on patch-wise encoding for local details. VFMTok (Zheng et al., 2025b) excels in performance but is slow due to long token sequences; SFTok (Rao et al., 2025) compresses to 64 tokens yet still suffers from slow iterative sampling. Our TaTok (56.4 tokens) achieves both performance and efficiency. These methods either complicate the training pipeline or fail to explicitly model holistic image semantics, leaving a clear gap for a unified tokenization framework that seamlessly integrates global and local information (Lin et al., 2017; He et al., 2022b; Zhao et al., 2023).

G.3 Tokenization for Downstream Visual Tasks

For image understanding tasks (e.g., classification, detection, multi-modal LLM (MLLM) reasoning) (Bengio et al., 2013; Isola et al., 2017; Karras et al., 2019; He et al., 2022a; Radford, 2018; Chen et al., 2020; Rombach et al., 2021), mainstream tokenizers (e.g., CLIP (Radford et al., 2021)) generate high-level semantic tokens that perform well in captioning and VQA (Kang et al., 2023; Ho et al., 2020) but struggle to capture fine-grained details; supplementary denoising methods (Song and Ermon, 2019; Song et al., 2020) introduce extra computational overhead (Caron et al., 2021; Oquab et al., 2023; Darcet et al., 2023). For image generation (Dhariwal and Nichol, 2021; Chang et al., 2022), AR transformers (Yu et al., 2023; Kingma, 2013; Caron et al., 2021) suffer from high latency due to step-by-step token prediction, while non-autoregressive methods face performance dependence on pre-trained latent quality. **All these downstream pipelines inherit the core fragmentation issue of 2D patch-based tokenization** (Liu et al., 2023a,b; Dong et al., 2024). Our work fills these gaps with an innovative 1D sequence-based discrete tokenization framework, which unifies holistic global representation and fine-grained local detail capture to overcome the inherent limitations of previous tokenizers (Tong et al., 2024; Wang et al., 2023; Baranchuk et al., 2021).