A UNIFIED RATE–DISTORTION PERSPECTIVE ON VECTOR, PRODUCT, AND SCALAR QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Discrete visual tokenization relies mainly on vector, scalar, and product quantization, but lacks a unified conceptual framework for understanding their tradeoffs. In this paper, we propose a unified rate-distortion perspective on modern discrete visual tokenization. By viewing quantization as lossy compression, we characterize the nominal fixed-length coding rate through token count and codebook size, and quantization error as the distortion. Within this framework, we resolve three central questions. First, our theoretical and empirical results identify minimizing distortion, rather than maximizing codebook utilization, as the primary intrinsic objective for reconstruction fidelity and link distortion directly to the STE-induced gradient discrepancy. Second, we identify two conditions for fair intrinsic comparison: controlling latent feature statistics and matching nominal coding rates. Third, under these conditions, we recover the VQ-PQ-SQ distortion hierarchy in modern visual tokenization; modern VQ methods attain the lowest distortion in our experiments. The rate-distortion perspective clarifies quantizer evaluation by isolating intrinsic quantization effectiveness under controlled fixed-rate constraints.

1 INTRODUCTION

Discrete visual tokenization has advanced in recent years through improvements in vector quantization (VQ; Esser et al., 2021; Sun et al., 2024; Tian et al., 2024), product quantization (PQ; Jégou et al., 2011; Guo et al., 2024; Li et al., 2025), and scalar quantization (SQ; Mentzer et al., 2024; Yu et al., 2024a; Zhao et al., 2025; Han et al., 2025). A major line of VQ research has focused on increasing codebook utilization through better codebook updates, optimization, or parameterizations (Lee et al., 2022; Zheng & Vedaldi, 2023; Zhu et al., 2024a;b; Shi et al., 2025; Chang et al., 2026; Lu et al., 2026). Despite this progress, the relationship between utilization, distortion, and quantization effectiveness is less well understood. In particular, high or even complete codebook utilization does not necessarily imply faithful representation of the input. This motivates a framework for understanding the objectives and tradeoffs of different quantization algorithms.

In this paper, we propose a unified rate-distortion perspective on quantization to address these questions. Rate-distortion theory provides the classical framework for characterizing lossy compression, and has also motivated learned image compression methods that jointly optimize coding rate and reconstruction distortion (Ballé et al., 2017; Lei et al., 2025; Zhang & Wu, 2024; Jiang et al., 2026). We use rate-distortion as an analytical framework for comparing discrete visual quantizers. Specifically, we define distortion as quantization error and rate as the nominal fixed-length coding budget determined by token count and codebook size. From this perspective, we resolve three central open questions in discrete visual tokenization, which are summarized in Figure 1.

First, we study the primary optimization objective for quantization algorithms in modern visual tokenization. From a rate-distortion perspective, distortion provides a criterion for reconstruction fidelity and controls the local STE-induced gradient discrepancy. We give both theoretical and empirical evidence supporting distortion minimization as a more fundamental optimization principle than codebook utilization. Specifically, under the conditions of Proposition 1, every global distortion minimizer fully utilizes the codebook, whereas full utilization does not imply distortion optimality. We also show a direct connection between distortion and the gradient discrepancy induced by the straight-through estimator (STE) (Bengio et al., 2013). Empirically, we find substantially stronger correlations between distortion and reconstruction fidelity (measured by rFID) than those observed

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.

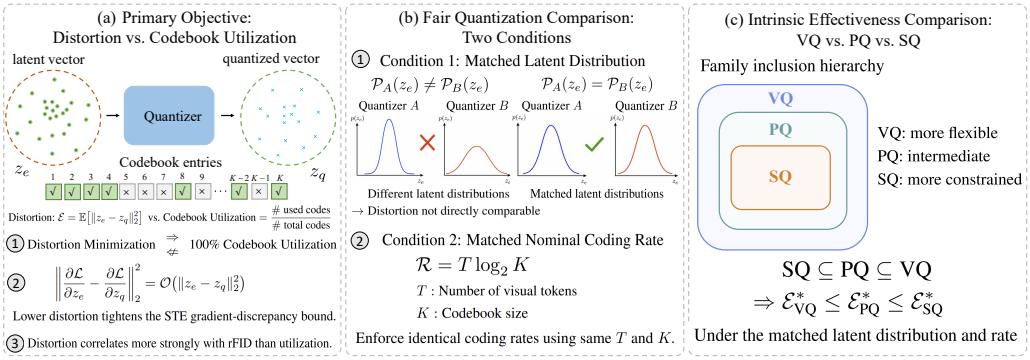

Figure 1: **Overview of our unified rate-distortion perspective on quantization.** (a) **Primary objective.** Distortion is a more fundamental criterion than codebook utilization, with direct links to utilization, STE-induced gradient discrepancy, and reconstruction fidelity. (b) **Fair comparison.** Intrinsic comparison requires matched latent feature distributions and the same nominal fixed-length rate. (c) **Intrinsic effectiveness.** VQ, PQ, and SQ satisfy SQ $\subseteq$ PQ $\subseteq$ VQ, implying $\mathcal{E}^*_{\text{VQ}} \leq \mathcal{E}^*_{\text{PQ}} \leq \mathcal{E}^*_{\text{SQ}}$ under matched latent distributions and fixed rate.

with codebook utilization. These results motivate a distortion-centric perspective on VQ optimization beyond codebook utilization, as summarized in Figure 1(a), complementing prior analyses of STE bias (Huh et al., 2023; Fifty et al., 2025) and recent efforts to address codebook collapse (Zhu et al., 2024b; Shi et al., 2025; Chang et al., 2026; Lu et al., 2026).

Second, we identify two conditions for fairly comparing the intrinsic effectiveness of quantization algorithms, as illustrated in Figure 1(b). First, the latent feature distribution must be controlled across compared algorithms. In particular, under global rescaling, the minimum achievable squared quantization distortion changes proportionally with the latent variance, so differences in feature statistics can confound comparisons of quantizer effectiveness. Second, all algorithms must operate at the same nominal fixed-length coding rate, $R = T \log_2 K$, which is jointly determined by the number of visual tokens T and the cardinality of the discrete code space K. Only under matched latent distributions and coding rates can the intrinsic effectiveness of different quantization algorithms be meaningfully compared.

Third, we examine the intrinsic effectiveness of quantization algorithms under controlled conditions. As illustrated in Figure 1(c), SQ, PQ, and VQ form a nested hierarchy in which VQ generalizes PQ and PQ generalizes SQ, implying greater modeling flexibility and no higher optimal distortion for VQ. Related work on neural image compression has similarly shown that vector or lattice quantization can achieve lower distortion than scalar quantization when exploiting cross-dimensional dependencies (Lei et al., 2025; Zhang & Wu, 2024). However, early visual-tokenizer studies reported severe codebook collapse for VQ, particularly with large codebooks (Dhariwal et al., 2020; Takida et al., 2022; Yu et al., 2022; Lee et al., 2022; Zheng & Vedaldi, 2023), leading to inferior empirical performance relative to SQ and PQ in prior works (Mentzer et al., 2024; Yu et al., 2024a; Zhao et al., 2025; Guo et al., 2024). To resolve this apparent conflict, we examine whether the classical distortion hierarchy among VQ, PQ, and SQ can be realized in modern visual tokenization under controlled fixed-rate conditions. Beyond this nested ordering, we show that VQ can exploit low-dimensional source structure to achieve strictly better distortion scaling than PQ and SQ. Our empirical study further shows that advanced VQ algorithms (Fang et al., 2025; 2026) consistently achieve lower distortion and better reconstruction under the two fairness conditions above.

Our key contributions are as follows:

1. **We establish distortion minimization as a fundamental intrinsic optimization principle for modern visual quantization.** Under the conditions of Proposition 1, every global distortion minimizer fully utilizes the codebook, whereas full utilization does not imply distortion optimality. We further connect quantization distortion to the gradient discrepancy induced by the straight-through estimator.
2. **We identify two conditions for fair intrinsic comparison of quantization algorithms.** We show that comparisons must control both the scale of latent features and the coding rate, since differences in either factor can obscure the intrinsic effectiveness of the quantizer itself.
3. **We characterize the VQ-PQ-SQ distortion hierarchy in modern visual tokenization and show that VQ can exploit low-dimensional source structure to achieve strictly better distortion**

scaling than PQ and SQ. In our controlled comparisons, modern VQ methods give the lowest distortion and, after decoder adaptation, the best reconstruction. The anonymous repository contains code and implementation details: Anonymous Repository

2 BACKGROUND

2.1 DISCRETE VISUAL TOKENIZATION

A discrete visual tokenizer produces a sequence of discrete symbols from an image for subsequent generative modeling (van den Oord et al., 2017; Esser et al., 2021; Sun et al., 2024; Tian et al., 2024). It typically contains an encoder $\mathcal{E}_\theta$, a quantization module $\mathcal{Q}_\phi$, and a decoder $\mathcal{D}_\varphi$. Given an image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times 3}$, the encoder produces continuous latent features

$$\boldsymbol{z}_e = \mathcal{E}_\theta(\boldsymbol{x}) \in \mathbb{R}^{h\times w\times d}, \qquad h = H/f, \;\; w = W/f,$$

where f denotes the spatial downsampling factor and d the latent dimension.

At each spatial position (i, j), the quantizer maps the continuous feature $\boldsymbol{z}_e^{ij} \in \mathbb{R}^d$ to a discrete representation,

$$r^{ij} = \mathcal{I}_\phi(\boldsymbol{z}_e^{ij}), \qquad \boldsymbol{z}_q^{ij} = \mathcal{Q}_\phi(\boldsymbol{z}_e^{ij}),$$

where r^{ij} denotes the discrete symbol and $\boldsymbol{z}_q^{ij}$ the quantized representation. The discrete symbols can be used as visual tokens for generative modeling. Reconstruction uses the quantized latent, decoded as $\hat{\boldsymbol{x}} = \mathcal{D}_\varphi(\boldsymbol{z}_q)$. We consider vector quantization (VQ), product quantization (PQ), and scalar quantization (SQ), three widely used families that impose different structures on the mapping from $\boldsymbol{z}_e$ to $\boldsymbol{z}_q$. We describe them below.

2.2 VECTOR QUANTIZATION

Vector quantization (VQ) discretizes an entire d-dimensional feature vector jointly (van den Oord et al., 2017). Given a learnable codebook $\phi = \{\boldsymbol{e}_k\}_{k=1}^K \subset \mathbb{R}^d$, containing K code vectors, VQ assigns each continuous feature $\boldsymbol{z}_e^{ij}$ to its nearest codebook entry:

$$r^{ij} = \underset{k\in\{1,\dots,K\}}{\arg\min} \|\boldsymbol{z}_e^{ij} - \boldsymbol{e}_k\|_2^2, \qquad \boldsymbol{z}_q^{ij} = \boldsymbol{e}_{r^{ij}}.$$

Thus, each spatial feature is represented by one of the K d-dimensional code vectors.

A longstanding optimization challenge in VQ is *codebook collapse*, where only a subset of the available code vectors is frequently selected (Dhariwal et al., 2020; Takida et al., 2022; Yu et al., 2022; Lee et al., 2022; Zheng & Vedaldi, 2023). The problem can become particularly pronounced for large codebooks (Zheng & Vedaldi, 2023; Mentzer et al., 2024). Many studies have therefore developed better codebook updates, optimization procedures, and parameterizations to increase codebook usage (Razavi et al., 2019; Williams et al., 2020; Zhang et al., 2023; Zhu et al., 2024a;b; Shi et al., 2025; Chang et al., 2026; Lu et al., 2026). Other work has studied difficulties caused by the non-differentiability of quantization, including the straight-through estimator and alternative gradient transformations (Huh et al., 2023; Fifty et al., 2025).

2.3 PRODUCT QUANTIZATION

Product quantization (PQ) factorizes a d-dimensional feature vector into M lower-dimensional subvectors and quantizes each subvector independently (Jégou et al., 2011; Guo et al., 2024; Li et al., 2025). Specifically,

$$\boldsymbol{z}_e^{ij} = \bigoplus_{m=1}^{M} \boldsymbol{z}_m^{ij}, \qquad \boldsymbol{z}_m^{ij} \in \mathbb{R}^{d_m}, \qquad \sum_{m=1}^{M} d_m = d,$$

where $\bigoplus$ denotes channel-wise concatenation. Each subvector is quantized using an independent subcodebook $\phi_m = \{\boldsymbol{e}_{m,k}\}_{k=1}^{n_m} \subset \mathbb{R}^{d_m}$ such that

$$r_m^{ij} = \underset{k\in\{1,\dots,n_m\}}{\arg\min} \|\boldsymbol{z}_m^{ij} - \boldsymbol{e}_{m,k}\|_2^2, \qquad \hat{\boldsymbol{z}}_m^{ij} = \boldsymbol{e}_{m,r_m^{ij}}.$$

The full quantized feature is then $\boldsymbol{z}_q^{ij} = \bigoplus_{m=1}^{M} \hat{\boldsymbol{z}}_m^{ij}$. The M subcodebooks jointly induce $K = \prod_{m=1}^{M} n_m$ possible composite codewords. Hence, PQ can represent a combinatorial discrete space using $\sum_{m=1}^{M} n_m$ explicitly stored subcodebook entries. For later analysis, we use K to denote this *composite code-space cardinality*, rather than the number of explicitly stored vectors.

2.4 Scalar Quantization

Scalar quantization (SQ) independently discretizes scalar components of a latent representation and can be viewed as the maximally factorized case of PQ with $M = d$. Writing

$$\boldsymbol{z}_e^{ij} = \bigoplus_{m=1}^{d} z_m^{ij}, \qquad z_m^{ij} \in \mathbb{R},$$

each scalar component is quantized independently using a scalar codebook $\phi_m = \{e_{m,k}\}_{k=1}^{n_m} \subset \mathbb{R}$. The corresponding discrete index and quantized value are

$$r_m^{ij} = \underset{k \in \{1,\dots,n_m\}}{\arg\min} |z_m^{ij} - e_{m,k}|^2, \qquad \hat{z}_m^{ij} = e_{m,r_m^{ij}}.$$

The resulting quantized vector is $\boldsymbol{z}_q^{ij} = \bigoplus_{m=1}^{d} \hat{z}_m^{ij}$, with composite code-space cardinality $K = \prod_{m=1}^{d} n_m$. Modern visual tokenizers use several forms of this scalar factorization. In FSQ, each latent dimension has a small finite set of scalar levels (Mentzer et al., 2024). LFQ uses binary values for each quantized coordinate (Yu et al., 2024a), while BSQ combines binary quantization with hyperspherical normalization (Zhao et al., 2025).

3 Toward a Controlled Comparison of Quantization Algorithms

Comparing quantization algorithms across independently trained visual tokenizers does not isolate the contribution of the quantizer itself. The quantization module is one of several factors that determine end-to-end tokenizer performance. Encoder-decoder capacity and architecture, discriminator design, training data, optimization schedule, and computational budget also matter, and they can differ substantially between tokenizer systems. Training every alternative quantizer from scratch in an otherwise identical setup is often computationally expensive. Consequently, empirical studies frequently compare separately developed systems with incompletely matched experimental conditions.

As illustrated in Table 1, representative tokenizers differ substantially along these dimensions (Esser et al., 2021; Lee et al., 2022; Zhu et al., 2024a; Sun et al., 2024; Tian et al., 2024; Li et al., 2025; Ma et al., 2025). A reconstruction difference can come from the quantization algorithm, model capacity, training objective, data, or optimization resources. Comparing these systems is informative about the complete tokenizers. It cannot isolate the *intrinsic effectiveness* of VQ, PQ, and SQ.

Table 1: Representative implementation differences among discrete visual tokenizers. Differences in model design, training setup, and computational budget can confound direct comparisons of quantization algorithms. '-' denotes unreported information, and "Para" denotes the number of encoder-decoder parameters.

Tokenizers	Para	Architectures	Discriminator	Training Datasets	Training Epochs	Training GPU Hours
VQGAN (Esser et al., 2021)	68M	CNN U-Net	PatchGAN	ImageNet-1k	-	-
RQVAE (Lee et al., 2022)	95M	CNN U-Net	PatchGAN	ImageNet-1k	50	-
VQGAN-LC (Zhu et al., 2024a)	68M	CNN U-Net	PatchGAN	ImageNet-1k	20	32 × V100 - Hours
LlamaGen (Sun et al., 2024)	68M	CNN U-Net	PatchGAN	ImageNet-1k	40	2 × A100 200 Hours
VAR (Tian et al., 2024)	104M	CNN U-Net	StyleGAN	OpenImages	16	16 × A100 60 Hours
ImageFolder (Li et al., 2025)	-	Transformer SEED	StyleGAN	ImageNet-1k	200	32 × A100 40 Hours
UniTok (Ma et al., 2025)	-	Transformer SEED	StyleGAN	OpenImages	-	256 × A100 50 Hours

These confounders motivate three questions about intrinsic quantization effectiveness: **(Q1)** What should a quantization algorithm optimize? **(Q2)** What conditions are required for a fair intrinsic comparison of different quantization algorithms? **(Q3)** Under these controlled conditions, how do VQ, PQ, and SQ compare in achievable distortion? In the next section, we develop a unified rate-distortion perspective to answer these three questions.

4 A Rate-distortion Perspective

We now develop a unified rate-distortion framework to address the three questions posed in Section 3. We first formalize the rate and distortion of discrete quantization, and then study its primary opti-

mization objective, the conditions required for fair intrinsic comparison, and the achievable distortion of VQ, PQ, and SQ under controlled conditions.

4.1 Fixed-Length Rate and Quantization Distortion

We formulate discrete quantization as a fixed-rate lossy compression problem. We compare the intrinsic effectiveness of quantization structures at a controlled representation budget. For this comparison, we use the nominal fixed-length coding rate for the budget and squared-error distortion for the information lost through quantization. Entropy-coded neural compression uses a different rate: the expected bitstream length under a learned entropy model (Ballé et al., 2017; Jiang et al., 2026).

Nominal fixed-length rate. Consider a tokenizer that produces T discrete tokens, each taking values in a discrete space of cardinality K. Its nominal fixed-length coding rate is

$$R = T \log_2 K \quad \text{bits.} \tag{1}$$

The nominal budget is $\log_2 K$ bits per token under fixed-length coding. This is the representation budget available to the quantizer. It should be distinguished from the expected entropy-coded rate, which depends on the distribution of the discrete symbols.

For VQ, K counts the available code vectors. For PQ and SQ, K is the size of the *composite* discrete code space, not the explicitly stored vectors. With M independently quantized components and sub-codebook sizes $\{n_m\}_{m=1}^{M}$, this count is $K = \prod_{m=1}^{M} n_m$. Using the composite count lets us match the nominal coding rate across the three families despite their different constraints on the discrete representation.

Quantization distortion. Let $X \sim P$ denote a latent feature vector in $\mathbb{R}^d$, and let $\mathcal{Q}_\phi$ denote a quantizer that maps X to a quantized representation $\mathcal{Q}_\phi(X)$. We measure quantization distortion using the expected squared Euclidean error

$$\mathcal{E}(P, \mathcal{Q}_\phi) = \mathbb{E}_{X \sim P}\left[\|X - \mathcal{Q}_\phi(X)\|_2^2\right]. \tag{2}$$

We approximate this expectation by averaging the squared quantization errors over the observed latent features. The analysis below uses this distortion as its main measure of information lost through quantization.

Remark 1 (Rate allocation)**.** *Under the nominal fixed-length rate $R = T \log_2 K$, doubling the token count is equivalent to squaring the code-space cardinality in terms of representation rate:*

$$2T \log_2 K = T \log_2 K^2. \tag{3}$$

Equal nominal budgets can give different distortions. Allocating the same rate differently between token count and code-space cardinality may change the representation structure. Whenever possible, our controlled comparisons match token count and composite code-space cardinality separately. This is a stronger control than matching R alone.

4.2 Distortion Minimization: The Primary Optimization Objective

Many studies of VQ have focused on reducing codebook collapse and improving codebook utilization (Zhu et al., 2024a; Dhariwal et al., 2020; Williams et al., 2020; Zheng & Vedaldi, 2023; Zhang et al., 2023; Ramesh et al., 2021). We argue that distortion minimization is a more fundamental optimization objective than codebook utilization. Intuitively, lower distortion promotes more stable and faithful representations in lossy compression systems (Touchette & Lloyd, 1999; Tomar et al., 2017; Touchette & Lloyd, 2001).

We first show that distortion controls the local gradient discrepancy induced by the straight-through estimator (STE). As derived in Appendix E, under standard local smoothness, the squared gradient discrepancy $\mathcal{G}$ satisfies

$$\mathcal{G} = \mathcal{O}\big(\|z_e - z_q\|_2^2\big). \tag{4}$$

Since $\|z_e - z_q\|_2^2$ is exactly the squared-error distortion, this gives a local connection between quantization distortion and STE gradient fidelity: as the distortion approaches zero, the gradient

discrepancy also vanishes. Thus, reducing distortion makes the gradient evaluated at the quantized representation locally closer to that evaluated at the continuous representation, providing a mechanism through which lower distortion can contribute to more stable optimization.

We further formalize the relationship between distortion minimization and codebook utilization. Let X be a random variable supported on $\mathcal{X} \subseteq \mathbb{R}^d$. A deterministic quantizer with codebook size $K \in \mathbb{N}_+$ is a mapping $f : \mathcal{X} \to \{1, \dots, K\}$. Define

$$\mathcal{E}(f) \triangleq \min_{\hat{x}:\{1,\dots,K\}\to\mathbb{R}^d} \mathbb{E}\left[\|X - \hat{x}(f(X))\|_2^2\right], \quad \mathrm{U}(f) \triangleq \frac{\left|\{k : \mathbb{P}(f(X) = k) > 0\}\right|}{K}.$$

Assumption 1 (Non-degenerate splittability)**.** *For any measurable set $A \subseteq \mathcal{X}$ with $\mathbb{P}(X \in A) > 0$ and nonzero conditional variance, there exists a partition of A into two measurable subsets $A = A_1 \dot{\cup} A_2$, each with positive probability, such that*

$$\mathbb{E}[X \mid X \in A_1] \neq \mathbb{E}[X \mid X \in A_2].$$

This mild condition is satisfied by standard continuous distributions with densities. The following proposition further excludes the trivial setting in which the support of the distribution can already be represented exactly using at most K codewords, since in that case zero distortion may be achieved without utilizing the entire codebook. For example, for an empirical distribution supported on n distinct samples, $K < n$ is sufficient to rule out such a zero-distortion solution.

Proposition 1. *(a) Let*

$$f^\star \in \operatorname*{arg\,min}_{f:\mathcal{X}\to\{1,\dots,K\}} \mathcal{E}(f).$$

Under Assumption 1, if no zero-distortion quantizer with at most K codewords exists, then every global distortion minimizer satisfies $\mathrm{U}(f^\star) = 1$.

(b) There exist quantizers $g : \mathcal{X} \to \{1, \dots, K\}$ with $\mathrm{U}(g) = 1$ that are not distortion minimizers:

$$\mathcal{E}(g) > \min_{f:\mathcal{X}\to\{1,\dots,K\}} \mathcal{E}(f).$$

Proposition 1 shows that, under these mild conditions, every global distortion minimizer fully utilizes the codebook, whereas full utilization does not imply distortion optimality. The proof is provided in Appendix A.

4.3 Conditions for Fair Intrinsic Comparison

A fair intrinsic comparison of quantization algorithms requires controlling both the latent source distribution and coding rate. First, all quantizers must operate on the same latent distribution, because quantization distortion depends not only on the quantization algorithm but also on the statistical properties of the source distribution. Even under the simplest change in source statistics, namely a global rescaling $X' = aX$, the optimal K-point squared-error distortion changes as

$$\mathcal{E}^\star(aX, K) = a^2 \mathcal{E}^\star(X, K). \tag{5}$$

Thus, differences in latent feature statistics can directly confound distortion comparisons. This result is proved in Proposition 4 in Appendix F.

Condition 1 (Matched latent distribution)**.** *All compared quantizers must be evaluated on the same latent distribution P.*

Second, distortion must be compared under the same coding budget. For T discrete tokens with code-space cardinality K, the nominal fixed-length coding rate is $R = T \log_2 K$. Since both T and K affect representational capacity and reconstruction behavior (Yu et al., 2024a; Zhu et al., 2024a; Yu et al., 2024b; Bachmann et al., 2025), comparisons at different rates do not isolate intrinsic quantization effectiveness.

Condition 2 (Matched nominal rate)**.** *The nominal fixed-length coding rate must be identical across all compared algorithms.*

In our controlled experiments, we adopt the stronger criterion $T_A = T_B$ and $K_A = K_B$, rather than matching only $T \log_2 K$, to avoid confounding different allocations of the same nominal rate.

4.4 Rate-Distortion Analysis of VQ, PQ, and SQ

In this section, we study the rate-distortion behavior of idealized VQ, PQ, and SQ, and compare their distortion under equal coding rates.

Given a probability distribution $\mathbb{P}$ over $\mathbb{R}^d$ and a codebook size $K \in \mathbb{N}_+$, we define the optimal distortions for VQ, PQ, and SQ as follows:

$$\mathcal{E}^*_{\text{VQ}}(\mathbb{P}, K) := \inf_{\phi} \Big\{ \mathbb{E}\big[\|X - \mathcal{Q}_\phi(X)\|_2^2\big] : |\phi| \le K \Big\},$$

$$\mathcal{E}^*_{\text{PQ}}\big(\mathbb{P}, K, M\big) := \inf_{\phi} \Big\{ \mathbb{E}\big[\|X - \mathcal{Q}_\phi(X)\|_2^2\big] : \phi = \bigoplus_{m=1}^{M} \phi_m, \phi_m \subseteq \mathbb{R}^{d_m}, |\phi_m| = n_m, \prod_{m=1}^{M} n_m \le K \Big\},$$

$$\mathcal{E}^*_{\text{SQ}}\big(\mathbb{P}, K\big) := \inf_{\phi} \Big\{ \mathbb{E}\big[\|X - \mathcal{Q}_\phi(X)\|_2^2\big] : \phi = \bigoplus_{m=1}^{d} \phi_m, \phi_m \subseteq \mathbb{R}, |\phi_m| = n_m, \prod_{m=1}^{d} n_m \le K \Big\}.$$

Here $|\phi|$ denotes the size of the codebook ϕ, and $\bigoplus$ denotes the Cartesian product of sets.

SQ is a special case of PQ with $M = d$, and PQ is a special case of VQ, a classical structural relationship in quantization theory. Thus, the optimal distortions satisfy:

$$\mathcal{E}^*_{\text{VQ}}(\mathbb{P}, K) \le \mathcal{E}^*_{\text{PQ}}\big(\mathbb{P}, K, M\big) \le \mathcal{E}^*_{\text{SQ}}\big(\mathbb{P}, K\big), \text{ for any } 2 \le M \le d.$$

Let $\mathcal{P}$ be the set of all probability distributions over $[-1, 1]^d$. The following results quantify the optimal distortions for VQ, PQ, and SQ.

Proposition 2. *For any $K \ge 2^d$, we have*

$$\frac{d}{8K^{2/d}} \le \sup_{\mathbb{P} \in \mathcal{P}} \mathcal{E}^*_{VQ}(\mathbb{P}, K) \le \sup_{\mathbb{P} \in \mathcal{P}} \mathcal{E}^*_{SQ}\big(\mathbb{P}, K\big) \le \frac{4d}{K^{2/d}}.$$

See Section B for the proof. Proposition 2 shows that for worst-case distributions, the optimal distortions of VQ, PQ, and SQ are the same up to universal constant factors. All three methods achieve a distortion that scales as $\Theta(d/K^{2/d})$.

When the data distribution has intrinsic low-dimensional structure, VQ can achieve strictly better distortion scaling than PQ and SQ, as the following proposition shows.

Proposition 3. *If the support of $\mathbb{P} \in \mathcal{P}$ is contained in a d_{eff}-dimensional subspace of $\mathbb{R}^d$ with $d_{eff} < d$, then for any $K \ge 2^{d_{eff}}$, we have*

$$\mathcal{E}^*_{VQ}(\mathbb{P}, K) \le \frac{4dd_{eff}}{K^{2/d_{eff}}}.$$

On the other hand, there exists a 1-dimensional linear subspace $L \subseteq \mathbb{R}^d$ and a distribution $\mathbb{P}$ supported on $L \cap [-1, 1]^d$ such that for any $K \ge 2^d$, we have

$$\mathcal{E}^*_{PQ}\big(\mathbb{P}, K, M\big) \ge \frac{M}{4} K^{-2/M}, \quad \textit{for any } 2 \le M \le d, \quad \textit{and} \quad \mathcal{E}^*_{SQ}\big(\mathbb{P}, K\big) \ge \frac{d}{4} K^{-2/d}.$$

See Section C for the proof. Proposition 3 shows that when the data distribution has intrinsic dimension $d_{\text{eff}} < d$, VQ can achieve distortion scaling as $\mathcal{O}(K^{-2/d_{\text{eff}}})$, significantly smaller than the worst-case rate $\Theta(d/K^{2/d})$ when $d_{\text{eff}} \ll d$. In contrast, for a simple 1-dimensional distribution, SQ still has a distortion lower bound of $\Omega(d/K^{2/d})$ while VQ can achieve $\mathcal{O}(dK^{-2})$. PQ interpolates between these extremes, with performance depending on the number of blocks M. The nested ordering alone does not describe this separation. VQ can use intrinsic low-dimensional structure; fixed coordinate-wise factorization may prevent PQ and SQ from doing so. Supports with covering-number growth $N(\varepsilon) \lesssim C\varepsilon^{-d_{\text{eff}}}$ also admit the VQ upper-bound argument, with constants that depend on covering regularity. The linear-subspace case above is the simplest special case. Appendix D discusses the relationship of these results to classical quantization theory and explains how our analysis differs.

Table 2: Quantization and reconstruction performance of VQ, PQ, and SQ on ImageNet-1K under the controlled latent-space setting. †: Results cited from VQ-Transplant (Fang et al., 2026). Within each quantization type and phase (Substitution/Adaptation), optimal values are underlined; overall best results per metric are **bold**.

Approaches	Types	Phase	Tokens	K	$\mathcal{E}$(↓)	U (↑)	PSNR(↑)	SSIM(↑)	LPIPS (↓)	r-FID(↓)	r-IS(↑)
Vanilla VQ†	VQ	Substitution	512	65536	0.422	0.2%	22.04	53.1	0.243	10.89	103.8
EMA VQ†	VQ	Substitution	512	65536	0.217	65.5%	24.94	65.9	0.127	1.78	185.8
Online VQ†	VQ	Substitution	512	65536	0.280	13.5%	24.42	63.2	0.147	2.28	174.2
Wasserstein VQ†	VQ	Substitution	512	65536	**0.201**	99.6%	25.22	**66.9**	**0.121**	1.76	186.0
MMD VQ†	VQ	Substitution	512	65536	**0.201**	99.9%	**25.24**	66.8	**0.121**	1.69	187.3
Vanilla VP2	PQ	Substitution	512	65536	0.233	59.2%	24.80	65.3	0.130	1.84	183.1
EMA VP2	PQ	Substitution	512	65536	0.209	**100%**	25.08	66.4	0.123	1.68	187.2
Online VP2	PQ	Substitution	512	65536	0.211	**100%**	25.09	66.3	0.124	1.79	185.7
Wasserstein VP2	PQ	Substitution	512	65536	0.217	**100%**	24.79	65.8	0.128	1.78	185.7
MMD VP2	PQ	Substitution	512	65536	0.212	**100%**	25.00	66.1	0.123	**1.61**	**189.5**
FSQ	SQ	Substitution	512	65536	0.272	100%	24.46	62.5	0.140	2.35	174.8
LFQ	SQ	Substitution	512	65536	0.279	29.8%	24.06	62.6	0.146	2.15	176.2
BSQ	SQ	Substitution	512	65536	0.231	**100%**	24.55	65.2	0.132	1.96	182.1
Vanilla VQ†	VQ	Adaptation	512	65536	0.422	0.2%	21.19	50.7	0.209	5.05	118.9
EMA VQ†	VQ	Adaptation	512	65536	0.217	65.5%	24.36	64.1	0.111	0.99	194.3
Online VQ†	VQ	Adaptation	512	65536	0.280	13.5%	23.84	61.6	0.130	1.38	182.9
Wasserstein VQ†	VQ	Adaptation	512	65536	**0.201**	99.6%	**24.68**	**65.4**	**0.106**	0.92	195.5
MMD VQ†	VQ	Adaptation	512	65536	**0.201**	99.9%	24.65	65.0	**0.106**	**0.86**	**197.1**
Vanilla VP2	PQ	Adaptation	512	65536	0.233	59.2%	24.28	64.0	0.114	1.07	191.7
EMA VP2	PQ	Adaptation	512	65536	0.209	**100%**	24.55	64.9	0.107	0.93	195.4
Online VP2	PQ	Adaptation	512	65536	0.211	**100%**	24.53	64.7	0.108	0.95	195.3
Wasserstein VP2	PQ	Adaptation	512	65536	0.217	**100%**	24.44	64.6	0.110	0.99	193.5
MMD VP2	PQ	Adaptation	512	65536	0.212	**100%**	24.43	64.5	0.109	0.95	196.1
FSQ	SQ	Adaptation	512	65536	0.272	**100%**	23.87	61.0	0.119	1.23	186.6
LFQ	SQ	Adaptation	512	65536	0.279	29.8%	23.42	60.7	0.130	1.30	183.2
BSQ	SQ	Adaptation	512	65536	0.231	**100%**	24.06	63.6	0.117	1.07	190.8

5 Empirical Evaluation

We empirically compare VQ, PQ, and SQ under matched latent distributions and nominal coding rates, with primary experiments conducted on ImageNet-1K (Deng et al., 2009). We further examine rate-distortion behavior across coding budgets and validate the conclusions across datasets and in a controlled pixel-space setting.

5.1 Experimental Setup

Controlled latent-space comparison. For our primary experiments, we use a pretrained VAR tokenizer (Tian et al., 2024) within the VQ-Transplant framework (Fang et al., 2026). During substitution, we replace the original quantizer with the method under evaluation and freeze the pretrained encoder and decoder. Decoder adaptation freezes the encoder and transplanted quantizer as the decoder adapts to the quantized representation. The shared pretrained encoder and representation pipeline give all methods the same latent source distribution, satisfying the matched-distribution condition in Section 4.3.

Quantization algorithms. We compare VQ, PQ, and SQ. For VQ, we evaluate Vanilla VQ (van den Oord et al., 2017), EMA VQ (Razavi et al., 2019), Online VQ (Zheng & Vedaldi, 2023), Wasserstein VQ (Fang et al., 2025), and MMD VQ (Fang et al., 2026). For PQ, we use the corresponding Vanilla VP2, EMA VP2, Online VP2, Wasserstein VP2, and MMD VP2 variants. For SQ, we evaluate FSQ (Mentzer et al., 2024), LFQ (Yu et al., 2024a), and BSQ (Zhao et al., 2025). All methods use $T = 512$ tokens and composite code-space cardinality $K = 65{,}536$ in the primary latent-space comparison, matching both components of the nominal rate $R = T \log_2 K$. Appendix H gives full implementation and training details for the latent- and pixel-space experiments.

Evaluation metrics. We report quantization distortion $\mathcal{E}$ and codebook utilization U. Reconstruction quality is evaluated using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), LPIPS (Zhang et al., 2018), reconstruction Fréchet Inception Distance (r-FID) (Heusel et al., 2017), and reconstruction Inception Score (r-IS) (Salimans et al., 2016).

5.2 Experimental Results

Controlled quantization comparison in latent space. Table 2 reports the primary ImageNet-1K comparison under matched latent feature statistics, token count, and code-space cardinality. The

results follow the distortion hierarchy: the best VQ method achieves lower distortion than the best PQ method, which in turn outperforms the best SQ method. MMD VQ gives $\mathcal{E} = 0.201$, versus 0.209 for PQ and 0.231 for SQ, following the theoretical order $\mathcal{E}^\star_{\mathrm{VQ}} \le \mathcal{E}^\star_{\mathrm{PQ}} \le \mathcal{E}^\star_{\mathrm{SQ}}$ in Section 4.4. Lower distortion also yields better reconstruction after decoder adaptation: MMD VQ achieves the best r-FID of 0.86, compared with 0.93 for PQ and 1.07 for SQ. Several PQ and SQ methods attain 100% codebook utilization despite higher distortion, showing that high utilization alone does not imply stronger quantization effectiveness.

Rate-distortion curves. Figure 2 compares the best-performing method within each quantization family across different coding rates. Consistent with our family-level theoretical comparison, at each nominal per-token rate $\log_2 K$, we report the minimum distortion achieved among the evaluated variants within each quantization family. Distortion decreases monotonically with rate in all three families. At every evaluated rate, VQ achieves the lowest distortion, followed by PQ and SQ, consistent with the theoretical hierarchy in Section 4.4. This advantage holds across the tested code-space cardinalities and is not specific to a single value of K.

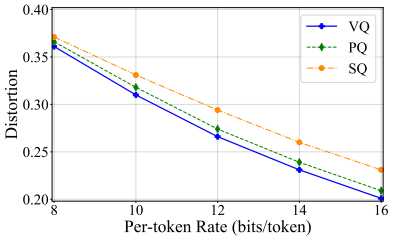

Figure 2: Rate–distortion curves on ImageNet-1K, reporting the best result within each quantization family at each rate.

Distortion versus codebook utilization. We also examine whether distortion or codebook utilization is more strongly associated with reconstruction fidelity. Using the adaptation-stage results in Table 2, we compute Spearman's rank correlations with r-FID. Distortion shows a near-perfect positive correlation ($\rho = 0.996$, $p < 10^{-8}$), whereas codebook utilization shows a substantially weaker negative correlation ($\rho = -0.540$, $p = 0.057$). Distortion is considerably more informative about reconstruction fidelity than utilization. High utilization alone cannot ensure faithful preservation of the source representation, supporting the distortion-minimization principle in Section 4.2.

Additional rate-distortion analyses. We analyze the effects of code-space cardinality K, token count T, and matched-rate allocations. Increasing either K or T reduces distortion, while equal-rate configurations yield similar but not identical performance. See Appendix I for full results.

Generalization across datasets and representation spaces. We also verify that our findings generalize beyond ImageNet-1K and the pretrained VAR latent space. Under the same controlled latent-space protocol, VQ retains its advantage over PQ and SQ on FFHQ and CelebA-HQ, with distortion remaining more strongly associated with reconstruction fidelity than codebook utilization (Appendix J). A controlled pixel-space comparison on CelebA-HQ yields the same qualitative conclusions, with VQ achieving the lowest distortion and best reconstruction fidelity (Appendix K). These results show that the findings hold across datasets and representation spaces.

6 Conclusion

In this work, we present a unified rate-distortion perspective for understanding and comparing vector, product, and scalar quantization. We characterize quantization error as distortion and the nominal fixed-length coding budget through token count and discrete code-space cardinality. This perspective yields three main conclusions. First, distortion minimization is a more fundamental optimization principle than codebook utilization: under the stated assumptions, every global distortion minimizer fully utilizes the codebook, whereas full utilization does not imply distortion optimality. Distortion also controls the local STE-induced gradient discrepancy. Second, fair intrinsic comparison requires matched latent distributions and nominal coding rates. Third, beyond the classical ordering $\mathcal{E}^\star_{\mathrm{VQ}} \le \mathcal{E}^\star_{\mathrm{PQ}} \le \mathcal{E}^\star_{\mathrm{SQ}}$, VQ can exploit low-dimensional source structure to achieve substantially better distortion scaling than fixed coordinate-factorized PQ and SQ. Our controlled experiments support these conclusions: distortion is more strongly associated with reconstruction fidelity than codebook utilization, and modern VQ methods achieve lower distortion and better reconstruction performance than PQ and SQ under matched representation budgets. Our results give a framework for evaluating intrinsic quantization effectiveness and guiding the analysis and design of discrete representation learning systems.

7 AI USE STATEMENT

In this work, we used generative AI tools only for limited writing and language editing assistance. Generative AI tools were not involved in the core methodology, experimental design, theoretical analysis, data generation, execution of experiments, or scientific contributions of this research. All AI-assisted text was reviewed and revised by the authors. We take full responsibility for the final content of this work, including text produced with the aid of generative AI.

8 ETHICS STATEMENT

This work does not involve crowdsourcing, human-subject studies, user studies, participant interaction, or the collection of human-subject data; therefore, Institutional Review Board (IRB) approval or an equivalent review was not required. The paper does not introduce or release new datasets, benchmarks, or pretrained models. It conducts theoretical analysis and experimental evaluation using existing publicly available datasets and model architectures. We do not identify additional ethical concerns related to participant risk, privacy, or data collection beyond those associated with the publicly available resources used in this work.

9 REPRODUCIBILITY STATEMENT

We provide materials to support reproduction and independent verification of our theoretical and empirical results. For the theoretical results, all assumptions, formal statements, derivations, and complete proofs are provided in the main paper and appendix.

For the empirical results, we provide an anonymous code repository linked in the introduction, together with extensive reproducibility materials. These resources include the training and evaluation source code for the reported experiments, execution scripts for the main experiments, configuration files specifying experimental settings, training logs recording optimization and model dynamics, and model outputs and evaluation artifacts underlying the reported results. The paper and supplementary material further document the datasets, data preprocessing procedures, model architectures, training configurations, optimization hyperparameters, quantization implementations, hardware settings, and evaluation protocols, with detailed experimental settings provided in Appendix H. All experiments are conducted using publicly available datasets, primarily ImageNet-1K, FFHQ, and CelebA-HQ.

To support independent verification and future research, we will publicly release the complete codebase and associated resources upon publication, including configuration files, execution scripts, and detailed documentation. Together, the anonymous repository, supplementary materials, experimental artifacts, and planned public release are intended to support reproduction of the main experiments, verification of the reported findings, and further research building upon this work.

REFERENCES

Roman Bachmann, Jesse Allardice, David Mizrahi, Enrico Fini, Ouguzhan Fatih Kar, Elmira Amirloo, Alaaeldin El-Nouby, Amir Zamir, and Afshin Dehghan. Flextok: Resampling images into 1d token sequences of flexible length. In *ICML*, 2025.

Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. In *ICLR*, 2017.

Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *ArXiv*, 2013.

James A. Bucklew and Gary L. Wise. Multidimensional asymptotic quantization theory with r-th power distortion measures. *IEEE Transactions on Information Theory*, 1982.

Yifan Chang, Jie Qin, Limeng Qiao, Xiaofeng Wang, Zheng Zhu, Lin Ma, and Xingang Wang. Scalable training for vector-quantized networks with 100% codebook utilization. In *ICLR*, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *ArXiv*, 2020.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.

Xianghong Fang, Litao Guo, Hengchao Chen, Yuxuan Zhang, Xiaofan Xia, Dingjie Song, Ye xin Liu, Hao Wang, Harry Yang, Yuan Yuan, and Qiang Sun. Enhancing vector quantization with distributional matching: A theoretical and empirical study. *ArXiv*, 2025.

Xianghong Fang, Yuan Yuan, Dehan Kong, and Tim G. J. Rudner. VQ-transplant: Efficient plug-and-play vq-module integration for pre-trained visual tokenizers. In *ICLR*, 2026.

Christopher Fifty, Ronald G. Junkins, Dennis Duan, Aniketh Iger, Jerry W. Liu, Ehsan Amid, Sebastian Thrun, and Christopher R'e. Restructuring vector quantization with the rotation trick. In *ICLR*, 2025.

Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. Optimized product quantization for approximate nearest neighbor search. In *CVPR*, 2013.

Allen Gersho and Robert M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992.

Robert M. Gray and David L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 1998.

Haohan Guo, Fenglong Xie, Dongchao Yang, Hui Lu, Xixin Wu, and Helen M. Meng. Addressing index collapse of large-codebook speech tokenizer with dual-decoding product-quantized variational auto-encoder. *Arxiv*, 2024.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *CVPR*, 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 *NeurIPS*, 2017.

Minyoung Huh, Brian Cheung, Pulkit Agrawal, and Phillip Isola. Straightening out the straight-through estimator: Overcoming optimization challenges in vector quantized networks. In *ICML*, 2023.

Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *TPAMI*, 2011.

Shiyin Jiang, Wei Long, Minghao Han, Zhenghao Chen, Ce Zhu, and Shuhang Gu. Differentiable vector quantization for rate-distortion optimization of generative image compression. In *CVPR*, 2026.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *ArXiv*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *CVPR*, 2022.

Eric Lei, Hamed Hassani, and Shirin Saeedi Bidokhti. Approaching rate-distortion limits in neural compression with lattice transform coding. In *ICLR*, 2025.

Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Jiuxiang Gu, Bhiksha Raj, and Zhe Lin. Imagefolder: Autoregressive image generation with folded tokens. In *ICLR*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Hao Lu, Onur Can Koyun, Yongxin Guo, Zhengjie Zhu, Abbas Alili, and Metin Nafi Gürcan. Beyond stationarity: Rethinking codebook collapse in vector quantization. *ArXiv*, 2026.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *ArXiv*, 2025.

Fabian Mentzer, David C. Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. In *ICLR*, 2024.

Mohammad Norouzi and David J. Fleet. Cartesian k-means. In *CVPR*, 2013.

Vivek Ramanujan, Kushal Tirumala, Armen Aghajanyan, Luke S. Zettlemoyer, and Ali Farhadi. When worse is better: Navigating the compression-generation tradeoff in visual tokenization. In *NeurIPS*, 2025.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021.

Ali Razavi, Aäron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *NeurIPS*, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016.

Fengyuan Shi, Zhuoyan Luo, Yixiao Ge, Yujiu Yang, Ying Shan, and Limin Wang. Taming scalable visual tokenizer for autoregressive image generation. In *ICCV*, 2025.

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*, 2024.

Yuhta Takida, Takashi Shibuya, Wei-Hsiang Liao, Chieh-Hsin Lai, Junki Ohmura, Toshimitsu Uesaka, Naoki Murata, Shusuke Takahashi, Toshiyuki Kumakura, and Yuki Mitsufuji. Sq-vae: Variational bayes on discrete representation with self-annealed stochastic quantization. In *ICML*, 2022.

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

Mahendra Singh Tomar, Matthias Rungger, and Majid Zamani. Invariance feedback entropy of uncertain control systems. *IEEE Transactions on Automatic Control*, 2017.

Hugo Touchette and Seth Lloyd. Information-theoretic limits of control. *Physical review letters*, 1999.

Hugo Touchette and Seth Lloyd. Information-theoretic approach to the study of control systems. *Physica A-statistical Mechanics and Its Applications*, 2001.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017.

Will Williams, Sam Ringer, Tom Ash, John Hughes, David Macleod, and Jamie Dougherty. Hierarchical quantized autoencoders. In *NeurIPS*, 2020.

Jingfeng Yao and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *CVPR*, 2025.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. In *ICLR*, 2022.

Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, 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 *ICLR*, 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 *NeurIPS*, 2024b.

Paul L. Zador. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory*, 1982.

Jiahui Zhang, Fangneng Zhan, Christian Theobalt, and Shijian Lu. Regularized vector quantization for tokenized image synthesis. In *CVPR*, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Xi Zhang and Xiaolin Wu. Learning optimal lattice vector quantizers for end-to-end neural image compression. In *NeurIPS*, 2024.

Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization. In *ICLR*, 2025.

Chuanxia Zheng and Andrea Vedaldi. Online clustered codebook. In *ICCV*, 2023.

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

Yongxin Zhu, Bocheng Li, Yifei Xin, and Linli Xu. Addressing representation collapse in vector quantized models with one linear layer. *ArXiv*, 2024b.

A Proof of Proposition 1

We prove the relationship between distortion minimization and codebook utilization. Let X be a random variable supported on $\mathcal{X} \subseteq \mathbb{R}^d$. A deterministic quantizer with codebook size K is a measurable mapping $f : \mathcal{X} \to \{1, \dots, K\}$. Given a reconstruction mapping $\hat{x} : \{1, \dots, K\} \to \mathbb{R}^d$, define the squared-error distortion as

$$\mathcal{E}(f, \hat{x}) \coloneqq \mathbb{E}\big[\|X - \hat{x}(f(X))\|_2^2\big] .$$

For a fixed quantizer f, the optimal reconstruction is given by the conditional mean $\hat{x}_f(k) = \mathbb{E}[X \mid f(X) = k]$, for every codeword k with $\mathbb{P}(f(X) = k) > 0$. Substituting the optimal reconstruction yields

$$\mathcal{E}(f) \coloneqq \min_{\hat{x}} \mathcal{E}(f, \hat{x}) = \mathbb{E}\big[\|X - \mathbb{E}[X \mid f(X)]\|_2^2\big] .$$

Equivalently, $\mathcal{E}(f) = \mathbb{E}[\mathrm{Var}(X \mid f(X))]$, where for a vector-valued random variable,

$$\mathrm{Var}(X \mid f(X)) \coloneqq \mathbb{E}\big[\|X - \mathbb{E}[X \mid f(X)]\|_2^2 \,\big|\, f(X)\big] .$$

Proof. **Part (a).** We prove by contradiction.

Suppose that $f^\star$ minimizes distortion but does not use all codewords. Then there exists an unused codeword k_{new} such that

$$\mathbb{P}(f^\star(X) = k_{\mathrm{new}}) = 0.$$

Since no zero-distortion quantizer exists, we must have $\mathcal{E}(f^\star) > 0$. Therefore, there exists at least one quantization cell

$$A^\star = \{x \in \mathcal{X} : f^\star(x) = k^\star\}$$

with positive probability and nonzero conditional variance.

By Assumption 1, there exists a measurable partition

$$A^\star = A_1 \dot{\cup} A_2$$

such that both A_1 and A_2 have positive probability and

$$\mu_i = \mathbb{E}[X \mid X \in A_i], \qquad \mu_1 \neq \mu_2.$$

Construct a refined quantizer f' by splitting $A^\star$:

$$f'(x) = \begin{cases} k^\star, & x \in A_1, \\ k_{\mathrm{new}}, & x \in A_2, \\ f^\star(x), & x \notin A^\star. \end{cases}$$

For all cells other than $A^\star$, the distortion contribution remains unchanged. Within $A^\star$, the law of total variance gives

$$\mathrm{Var}(X \mid X \in A^\star) = \sum_{i=1}^{2} \mathbb{P}(A_i \mid A^\star)\, \mathrm{Var}(X \mid X \in A_i) + \sum_{i=1}^{2} \mathbb{P}(A_i \mid A^\star) \|\mu_i - \mu\|_2^2,$$

where $\mu = \mathbb{E}[X \mid X \in A^\star]$.

Since $\mu_1 \neq \mu_2$, the second term is strictly positive. Hence

$$\sum_{i=1}^{2} \mathbb{P}(A_i \mid A^\star)\, \mathrm{Var}(X \mid X \in A_i) < \mathrm{Var}(X \mid X \in A^\star).$$

Therefore, $\mathcal{E}(f') < \mathcal{E}(f^\star)$, contradicting the optimality of $f^\star$. Hence every distortion-minimizing quantizer must use all K codewords: $\mathrm{U}(f^\star) = 1$.

Part (b). We construct a full-utilization quantizer that is not distortion-optimal. Let $K = 2$ and let $X \sim \mathrm{Unif}[0, 1]$. Consider the quantizer

$$f_0(x) = \begin{cases} 1, & x \in [0, 1/2], \\ 2, & x \in (1/2, 1]. \end{cases}$$

Its optimal reconstructions are

$$\hat{x}_{f_0}(1) = \frac{1}{4}, \qquad \hat{x}_{f_0}(2) = \frac{3}{4}.$$

Hence

$$\mathcal{E}(f_0) = \int_0^{1/2} \left(x - \frac{1}{4}\right)^2 dx + \int_{1/2}^1 \left(x - \frac{3}{4}\right)^2 dx = \frac{1}{48}.$$

Now define another quantizer

$$g(x) = \begin{cases} 1, & x \in [0, 9/10], \\ 2, & x \in (9/10, 1]. \end{cases}$$

Both codewords are used, so $\mathrm{U}(g) = 1$.

The optimal reconstructions are

$$\hat{x}_g(1) = \frac{9}{20}, \qquad \hat{x}_g(2) = \frac{19}{20}.$$

Therefore,

$$\mathcal{E}(g) = \int_0^{9/10} \left(x - \frac{9}{20}\right)^2 dx + \int_{9/10}^1 \left(x - \frac{19}{20}\right)^2 dx = \frac{73}{1200}.$$

Since

$$\frac{73}{1200} > \frac{1}{48},$$

we obtain

$$\min_{f:\mathcal{X}\to\{1,2\}} \mathcal{E}(f) \le \mathcal{E}(f_0) < \mathcal{E}(g).$$

Thus g uses all codewords but is not distortion-optimal. □

B Proof of Proposition 2

We first prove the upper bound by constructing an SQ scheme. Let $n = \lfloor K^{1/d} \rfloor$. Since $K \ge 2^d$, we have $K^{1/d} \ge 2$ and hence $n \ge K^{1/d}/2$. For each coordinate $m \in [d]$, partition $[-1, 1]$ into n intervals of equal length $2/n$ and choose their midpoints as scalar reconstruction levels,

$$\phi_m = \left\{ -1 + \frac{2i-1}{n} : i = 1, \dots, n \right\}.$$

Let $\phi = \bigoplus_{m=1}^d \phi_m$. Then $|\phi| = n^d \le K$. For any $x \in [-1, 1]^d$, nearest-neighbor scalar quantization satisfies $|x_m - Q_{\phi_m}(x_m)| \le 1/n$ for every coordinate. Therefore,

$$\begin{aligned} \mathbb{E}\left[\|X - \mathcal{Q}_\phi(X)\|_2^2\right] &\le \sup_{x\in[-1,1]^d} \|x - \mathcal{Q}_\phi(x)\|_2^2 \\ &\le \frac{d}{n^2} \le \frac{4d}{K^{2/d}}. \end{aligned}$$

Since this construction is valid for every $\mathbb{P} \in \mathcal{P}$,

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathcal{E}^*_{\mathrm{SQ}}(\mathbb{P}, K) \le \frac{4d}{K^{2/d}}.$$

We next prove the lower bound using the uniform distribution $\mathbb{P} = \mathrm{Unif}([-1, 1]^d)$. Let $\phi = \{e_1, \dots, e_K\}$ be an arbitrary VQ codebook, and let $A_1, \dots, A_K$ be its nearest-neighbor cells intersected with $[-1, 1]^d$. Write $v_k = |A_k|$ and let V_d denote the volume of the Euclidean unit ball in $\mathbb{R}^d$. For any measurable set $A \subseteq \mathbb{R}^d$ of volume v and any $e \in \mathbb{R}^d$, the centered Euclidean ball of volume v minimizes the second moment about e, so

$$\int_A \|x - e\|_2^2 \, dx \ge \frac{d}{d+2} V_d^{-2/d} v^{1+2/d}.$$

Consequently,

$$\mathbb{E}\big[\|X - \mathcal{Q}_\phi(X)\|_2^2\big] = \frac{1}{2^d}\sum_{k=1}^{K}\int_{A_k}\|x - e_k\|_2^2\,dx$$
$$\geq \frac{1}{2^d}\frac{d}{d+2}V_d^{-2/d}\sum_{k=1}^{K}v_k^{1+2/d}.$$

Because $\sum_{k=1}^{K} v_k = 2^d$, convexity gives

$$\sum_{k=1}^{K} v_k^{1+2/d} \geq K\left(\frac{2^d}{K}\right)^{1+2/d} = 2^{d+2}K^{-2/d}.$$

Thus

$$\mathbb{E}\big[\|X - \mathcal{Q}_\phi(X)\|_2^2\big] \geq \frac{4d}{d+2}V_d^{-2/d}K^{-2/d}.$$

Finally, $V_d^{2/d} \leq 32/(d+2)$ for every integer $d \geq 1$. For $d = 1, 2$ this follows directly from $V_1 = 2$ and $V_2 = \pi$; for $d \geq 3$, it follows from the standard bound $V_d \leq (2\pi e/d)^{d/2}$ and $2\pi e/d \leq 32/(d+2)$. Hence

$$\mathbb{E}\big[\|X - \mathcal{Q}_\phi(X)\|_2^2\big] \geq \frac{d}{8}K^{-2/d}.$$

Since the codebook ϕ was arbitrary,

$$\sup_{\mathbb{P}\in\mathcal{P}}\mathcal{E}_{\text{VQ}}^*(\mathbb{P}, K) \geq \frac{d}{8K^{2/d}}.$$

Combining this with $\mathcal{E}_{\text{VQ}}^*(\mathbb{P}, K) \leq \mathcal{E}_{\text{SQ}}^*(\mathbb{P}, K)$ completes the proof.

C Proof of Proposition 3

We first prove the VQ upper bound. By assumption, there exists a d_{eff}-dimensional linear subspace $L \subseteq \mathbb{R}^d$ such that $\mathbb{P}$ is supported on $L \cap [-1, 1]^d$. Let $U \in \mathbb{R}^{d\times d_{\text{eff}}}$ have orthonormal columns spanning L. For $X \sim \mathbb{P}$, write $X = UZ$ with $Z = U^\top X$, and let $\widetilde{\mathbb{P}}$ denote the distribution of Z. Since U is an isometry on $\mathbb{R}^{d_{\text{eff}}}$,

$$\|Z\|_2 = \|UZ\|_2 = \|X\|_2 \leq \sqrt{d},$$

and hence $\|Z\|_\infty \leq \sqrt{d}$. Therefore, $\widetilde{\mathbb{P}}$ is supported on $[-\sqrt{d}, \sqrt{d}]^{d_{\text{eff}}}$.

Let $n = \lfloor K^{1/d_{\text{eff}}} \rfloor$. Because $K \geq 2^{d_{\text{eff}}}$, we have $n \geq K^{1/d_{\text{eff}}}/2$. Quantize each coordinate of Z independently using n equally spaced intervals on $[-\sqrt{d}, \sqrt{d}]$ and their midpoints as reconstruction levels. The resulting Cartesian-product codebook contains $n^{d_{\text{eff}}} \leq K$ points, and every coordinate incurs squared error at most d/n^2. Consequently,

$$\mathbb{E}_{Z\sim\widetilde{\mathbb{P}}}\big[\|Z - \mathcal{Q}_{\tilde{\phi}}(Z)\|_2^2\big] \leq \frac{d\,d_{\text{eff}}}{n^2} \leq \frac{4d\,d_{\text{eff}}}{K^{2/d_{\text{eff}}}}.$$

Mapping every reconstruction point $\tilde{e}$ back to $U\tilde{e} \in L$ preserves Euclidean distances. Thus the same distortion is achievable by a VQ codebook in $\mathbb{R}^d$, proving

$$\mathcal{E}_{\text{VQ}}^*(\mathbb{P}, K) \leq \frac{4d\,d_{\text{eff}}}{K^{2/d_{\text{eff}}}}.$$

We next prove the PQ and SQ lower bounds. Consider

$$L = \{\alpha\mathbf{1} : \alpha \in \mathbb{R}\}, \qquad \mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^d,$$

and let $X = A\mathbf{1}$, where $A \sim \text{Unif}[-1, 1]$. Then X is supported on $L \cap [-1, 1]^d$.

We use the following elementary one-dimensional fact. If $A \sim \mathrm{Unif}[-1,1]$, then every quantizer with at most n reconstruction values satisfies

$$\mathbb{E}\big[(A-\widehat{A})^2\big] \geq \frac{1}{3n^2}. \tag{6}$$

Indeed, for a nearest-neighbor scalar quantizer, let the lengths of its nonempty Voronoi cells be $\ell_1,\dots,\ell_r$, where $r \leq n$ and $\sum_{j=1}^{r} \ell_j = 2$. Replacing the reconstruction value in each cell by its midpoint can only decrease squared error, so

$$\mathbb{E}\big[(A-\widehat{A})^2\big] \geq \frac{1}{2}\sum_{j=1}^{r}\frac{\ell_j^3}{12} = \frac{1}{24}\sum_{j=1}^{r}\ell_j^3 \geq \frac{1}{24}\frac{\left(\sum_{j=1}^{r}\ell_j\right)^3}{r^2} \geq \frac{1}{3n^2},$$

where the penultimate inequality follows from convexity.

Now consider an arbitrary coordinate-block PQ quantizer with M blocks of dimensions $d_1,\dots,d_M$, where $d_m \geq 1$ and $\sum_{m=1}^{M} d_m = d$. Let the m-th sub-codebook contain n_m points, with $\prod_{m=1}^{M} n_m \leq K$. The input to block m is $A\mathbf{1}_{d_m}$. For any reconstruction point in $\mathbb{R}^{d_m}$, orthogonally projecting it onto the line $\mathrm{span}(\mathbf{1}_{d_m})$ cannot increase its distance to $A\mathbf{1}_{d_m}$. Hence an n_m-point vector quantizer for this block cannot have smaller distortion than the optimal n_m-level scalar quantizer for A, scaled by $\|\mathbf{1}_{d_m}\|_2^2 = d_m$. Using equation 6,

$$\mathbb{E}\big[\|A\mathbf{1}_{d_m} - \mathcal{Q}_{\phi_m}(A\mathbf{1}_{d_m})\|_2^2\big] \geq \frac{d_m}{3n_m^2} \geq \frac{d_m}{4n_m^2}.$$

Summing over the mutually orthogonal coordinate blocks gives

$$\begin{aligned}
\mathbb{E}\big[\|X - \mathcal{Q}_\phi(X)\|_2^2\big] &\geq \frac{1}{4}\sum_{m=1}^{M}\frac{d_m}{n_m^2} \\
&\geq \frac{M}{4}\left(\prod_{m=1}^{M}\frac{d_m}{n_m^2}\right)^{1/M} \\
&\geq \frac{M}{4}\left(\prod_{m=1}^{M} n_m\right)^{-2/M} \\
&\geq \frac{M}{4}K^{-2/M}.
\end{aligned}$$

Here the second line is AM-GM, and the third uses $d_m \geq 1$. Because the PQ codebooks and rate allocation $(n_1,\dots,n_M)$ were arbitrary, taking the infimum proves

$$\mathcal{E}^*_{\mathrm{PQ}}(\mathbb{P},K,M) \geq \frac{M}{4}K^{-2/M}, \qquad 2 \leq M \leq d.$$

Finally, SQ is the case $M = d$ and $d_m = 1$ for every block, yielding

$$\mathcal{E}^*_{\mathrm{SQ}}(\mathbb{P},K) \geq \frac{d}{4}K^{-2/d}.$$

This completes the proof.

D Relationship to Classical Quantization Theory

The $K^{-2/d}$ distortion scaling in Proposition 2 has the same high-rate exponent as Zador's asymptotic quantization theorem (Zador, 1982) and subsequent multidimensional extensions (Bucklew & Wise, 1982), as surveyed by Gray & Neuhoff (1998). The classical results describe optimal quantization distortion in the high-rate regime for regular d-dimensional sources. We use Proposition 2 for a different question: worst-case distortion at finite K over distributions supported on a bounded domain. Similarly, under the quantizer families defined in Section 4.4, the nested ordering $\mathcal{E}^*_{\mathrm{VQ}} \leq \mathcal{E}^*_{\mathrm{PQ}} \leq \mathcal{E}^*_{\mathrm{SQ}}$ follows directly from the structural inclusion $\mathrm{SQ} \subseteq \mathrm{PQ} \subseteq \mathrm{VQ}$: SQ is the fully factorized special case of PQ, while a PQ codebook is a Cartesian-product-structured VQ codebook. Classical treatments

of vector and product quantization already give this relationship (Gersho & Gray, 1992; Gray & Neuhoff, 1998). We include Proposition 2 as a self-contained finite-rate reference for our framework; we do not claim a new asymptotic quantization result.

The main theoretical distinction of our analysis lies in Proposition 3, which provides the explicit, non-asymptotic lower bound $\mathcal{E}^*_{\text{PQ}}(P, K, M) \geq \frac{M}{4} K^{-2/M}$ for the fixed coordinate-block PQ family on an axis-misaligned one-dimensional source. The scaling of this quantitative VQ-PQ separation depends explicitly on the number of blocks M. In approximate nearest-neighbor search, prior work has shown that the choice of subspace decomposition can strongly affect PQ distortion. Optimizing the orientation or decomposition of the feature space can reduce that distortion (Ge et al., 2013; Norouzi & Fleet, 2013). To our knowledge, that literature does not give the explicit finite-K lower bound with block-count dependence proved here.

Two parts of the study go beyond classical distortion analysis. Appendix E connects quantization distortion to the STE-induced gradient gap, relating quantization error to the optimization dynamics of learned discrete representations. We also examine the theoretical relationships above in visual-tokenization experiments: VQ, PQ, and SQ receive matched source representations, with token counts and composite code-space cardinalities held fixed across quantizers.

E Distortion Minimization and STE Gradient Discrepancy

In this section, we derive a local relationship between quantization distortion and the gradient discrepancy induced by the Straight-Through Estimator (STE) (Bengio et al., 2013). The analysis shows that, under standard local smoothness, the gradient discrepancy vanishes quadratically with the quantization distance, providing theoretical support for distortion minimization as a means of improving the local fidelity of the STE gradient.

Problem Setup: Let $\mathcal{L}$ denote the loss as a function of the latent representation. In the quantization process, the encoder outputs a continuous latent vector z_e, which is discretized into a quantized latent representation z_q. During backpropagation, the non-differentiable quantization operation is bypassed using the STE, which propagates the gradient evaluated at the quantized representation:

$$\text{STE Approximation:} \quad \frac{\partial \mathcal{L}}{\partial z_e} \leftarrow \frac{\partial \mathcal{L}}{\partial z_q}. \tag{7}$$

We define the **gradient gap** $\mathcal{G}$ as the squared discrepancy between the gradients evaluated at the continuous and quantized latent representations:

$$\mathcal{G} \triangleq \left\| \frac{\partial \mathcal{L}}{\partial z_e} - \frac{\partial \mathcal{L}}{\partial z_q} \right\|_2^2. \tag{8}$$

Local Taylor Analysis: Assume that $\mathcal{L}$ is twice differentiable in a neighborhood of z_q. Let

$$\mathbf{H}(z_q) = \left. \frac{\partial^2 \mathcal{L}}{\partial x^2} \right|_{x=z_q}$$

denote the Hessian at z_q. A first-order expansion of the gradient around z_q gives

$$\frac{\partial \mathcal{L}}{\partial z_e} = \frac{\partial \mathcal{L}}{\partial z_q} + \mathbf{H}(z_q)(z_e - z_q) + o(\|z_e - z_q\|_2). \tag{9}$$

Therefore,

$$\frac{\partial \mathcal{L}}{\partial z_e} - \frac{\partial \mathcal{L}}{\partial z_q} = \mathbf{H}(z_q)(z_e - z_q) + o(\|z_e - z_q\|_2), \tag{10}$$

and hence

$$\mathcal{G} = \|\mathbf{H}(z_q)(z_e - z_q) + o(\|z_e - z_q\|_2)\|_2^2 = \mathcal{O}(\|z_e - z_q\|_2^2). \tag{11}$$

The leading-order term is

$$\mathcal{G} = \|\mathbf{H}(z_q)(z_e - z_q)\|_2^2 + o(\|z_e - z_q\|_2^2). \tag{12}$$

Since $\|z_e - z_q\|_2^2$ is exactly the squared quantization distortion, Equation 11 gives a local connection between distortion and the STE-induced gradient discrepancy: as the quantization distortion

approaches zero, the gradient gap also vanishes at least quadratically in the quantization distance. With smaller distortion, the gradient evaluated at the quantized representation is locally closer to the gradient at the continuous representation. This local relation explains why reducing distortion can improve the fidelity of gradient propagation through the STE.

F Effect of Latent Scale on Quantization Distortion

We provide a theoretical analysis of the effect of latent feature scale on optimal quantization distortion. The result motivates the matched-distribution condition introduced in Section 4.3.

Theoretical analysis. Let the optimal K-point vector-quantization distortion under squared Euclidean distance be

$$\mathcal{E}^\star(X, K) = \inf_{\mathcal{C}:\,|\mathcal{C}|=K} \mathbb{E}\left[\min_{c\in\mathcal{C}} \|X - c\|_2^2\right]. \tag{13}$$

Proposition 4 (Scale dependence of optimal distortion)**.** *For any scalar $a \in \mathbb{R}$,*

$$\mathcal{E}^\star(aX, K) = a^2\mathcal{E}^\star(X, K). \tag{14}$$

Proof. For $a \neq 0$, the mapping

$$\mathcal{C}' \mapsto a\mathcal{C}' = \{ac' : c' \in \mathcal{C}'\}$$

is a bijection over the set of all K-point codebooks. Hence,

$$\begin{aligned}
\mathcal{E}^\star(aX, K) &= \inf_{\mathcal{C}:\,|\mathcal{C}|=K} \mathbb{E}\left[\min_{c\in\mathcal{C}} \|aX - c\|_2^2\right] && (15)\\
&= \inf_{\mathcal{C}':\,|\mathcal{C}'|=K} \mathbb{E}\left[\min_{c'\in\mathcal{C}'} \|aX - ac'\|_2^2\right] && (16)\\
&= a^2 \inf_{\mathcal{C}':\,|\mathcal{C}'|=K} \mathbb{E}\left[\min_{c'\in\mathcal{C}'} \|X - c'\|_2^2\right] && (17)\\
&= a^2\mathcal{E}^\star(X, K). && (18)
\end{aligned}$$

For $a = 0$, both sides are zero, completing the proof. □

Under global rescaling of the source distribution, the optimal squared quantization distortion is homogeneous of degree two (Proposition 4). No Gaussian assumption or high-rate approximation is needed for this result.

A useful special case follows by writing a location-scale family as

$$X = \mu + \sigma Z, \tag{19}$$

where Z has a fixed reference distribution. Since translation can be absorbed by an equal translation of the codebook,

$$\mathcal{E}^\star(\mu + \sigma Z, K) = \sigma^2\mathcal{E}^\star(Z, K). \tag{20}$$

Thus, when the shape of the latent distribution is fixed and only its scale varies, the optimal distortion grows linearly with the source variance σ^2.

G Limitations

Our results have several limitations. First, we study the intrinsic effectiveness of quantization under a nominal fixed-length representation rate, $R = T \log_2 K$, rather than the expected bitstream length produced by entropy coding. The probabilities of the discrete symbols also affect the effective coding rate in entropy-constrained quantization, variable-length coding, and adaptive bit allocation. Our analysis does not directly characterize these settings. Including them would connect the analysis more closely to classical rate-distortion optimization in learned compression.

Second, our theoretical analysis adopts squared Euclidean quantization distortion as the intrinsic information-loss measure. It lets us characterize source scaling and the structural relationships

among VQ, PQ, and SQ. Perceptual, semantic, and task-dependent distortion measures are not directly covered. The hierarchy is about the optimal distortion achievable by each family. It does not guarantee that every practical VQ algorithm outperforms every PQ or SQ method with finite data and finite optimization.

Finally, our conclusions concern *intrinsic quantization and reconstruction effectiveness*, rather than downstream generative modeling performance. Generation quality need not improve when reconstruction fidelity improves (Ramanujan et al., 2025; Yao & Wang, 2025). Downstream performance also depends on the properties of the discrete representation and the generative model's capacity and optimization. A question for future work is how intrinsic quantization distortion interacts with representation modelability and downstream generation.

H Experimental Details

Data Preprocessing. For all three datasets, ImageNet-1k (Deng et al., 2009), FFHQ (Karras et al., 2019), and CelebA-HQ (Karras et al., 2017), we follow LlamaGen (Sun et al., 2024) by applying iterative box downsampling to resize all images to a 256×256 resolution.

Encoder-Decoder Architecture. For experiments conducted in the latent space, we adopt the VQ-Transplant framework (Fang et al., 2026) and initialize all quantization methods with the same pretrained VAR tokenizer (Tian et al., 2024). This setup fixes the encoder-decoder architecture and latent representation across methods, enabling a controlled comparison of different quantization algorithms under the same latent space. The encoder is a U-Net (Ronneberger et al., 2015) that downsamples input images by a factor of 16, producing latent features $\boldsymbol{z}_e = \mathcal{E}_\theta(\boldsymbol{x}) \in \mathbb{R}^{16\times16\times32}$ with a spatial resolution of 16×16. In addition to these latent-space experiments, we also consider a pixel-space setting constructed directly from image pixels without relying on a pretrained encoder-decoder representation.

Training Details. Following Fang et al. (2026), all latent-space models were trained on two NVIDIA H100 GPUs using the AdamW optimizer (Loshchilov & Hutter, 2019) with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. During VQ module substitution, we used an initial learning rate of 10^{-4} with linear decay to 10^{-5}, while the learning rate was fixed at 10^{-5} during decoder adaptation. For ImageNet-1k, VQ module substitution and decoder adaptation were performed for 2 and 5 epochs, respectively. For both FFHQ and CelebA-HQ, VQ module substitution and decoder adaptation were each performed for 30 epochs. The batch size was fixed at 32 per GPU for all latent-space experiments. We additionally conducted pixel-space experiments on CelebA-HQ using a single-stage training procedure for 30 epochs. We used the same AdamW optimizer and batch size as in the latent-space experiments, with an initial learning rate of 10^{-4} linearly decayed to 10^{-5}.

Loss Weights. For the latent-space experiments, we follow the loss formulation and notation of VQ-Transplant (Fang et al., 2026). Here, λ_P and λ_G denote the perceptual and adversarial loss weights, respectively, while γ controls the weight of the distribution-matching objective. We set $\lambda_P = 1$ and $\lambda_G = 0.4$. For configurations using the Wasserstein distance (Wasserstein VQ and Wasserstein VP2), we set $\gamma = 0.2$, while for configurations using the MMD distance (MMD VQ and MMD VP2), we set $\gamma = 0.5$. For the pixel-space experiments, the training objective consists of the reconstruction, perceptual, and quantization losses, with the perceptual loss weight fixed to $\lambda_P = 1$. For Wasserstein- and MMD-based quantizers, the corresponding distribution-matching weight γ follows the same setting as in the latent-space experiments.

Latent-Space VQ Implementation. We implement five VQ variants within the VQ-Transplant framework: Vanilla VQ (van den Oord et al., 2017), EMA VQ (Razavi et al., 2019), Online VQ (Zheng & Vedaldi, 2023), Wasserstein VQ (Fang et al., 2025), and MMD VQ (Fang et al., 2026). We first use the pretrained VAR encoder to extract latent features $\boldsymbol{z}_e = \mathcal{E}_\theta(\boldsymbol{x}) \in \mathbb{R}^{16\times16\times32}$. A three-layer convolutional projector then processes $\boldsymbol{z}_e$ while preserving the 32-dimensional feature space. Each 32-dimensional feature vector is partitioned into two 16-dimensional vectors, each of which is treated as an independent quantization token and processed by a separate VQ module. Consequently, the 16×16 latent grid yields a total of $T = 512$ quantization tokens. The two quantized 16-dimensional vectors at each spatial location are subsequently concatenated to recover a 32-dimensional representation, as

illustrated in Figure 3. A second three-layer convolutional projector maps the concatenated quantized features to the decoder input. The code-space cardinality of each token is fixed to $K = 65{,}536$, yielding a nominal representation rate of $R = T \log_2 K$ with $T = 512$.

Latent-Space PQ Implementation. We implement five PQ variants within the VQ-Transplant framework: Vanilla VP2 (van den Oord et al., 2017), EMA VP2 (Razavi et al., 2019), Online VP2 (Zheng & Vedaldi, 2023), Wasserstein VP2 (Fang et al., 2025), and MMD VP2 (Fang et al., 2026). The overall architecture and tokenization scheme are identical to those used for VQ, with the only change being the internal quantization structure. Specifically, each 16-dimensional quantization token is further partitioned into $M = 2$ 8-dimensional sub-vectors. We quantize the two sub-vectors independently, each with a separate codebook of size 256. Each token then has composite code-space cardinality $256^2 = 65{,}536$. Therefore, PQ retains the same number of quantization tokens, $T = 512$, and the same effective code-space cardinality, $K = 65{,}536$, as VQ, ensuring a matched nominal representation rate.

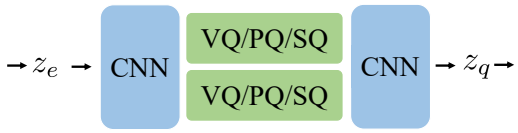

Figure 3: Illustration of the implementation details.

Latent-Space SQ Implementation. We implement three SQ variants within the VQ-Transplant framework: FSQ (Mentzer et al., 2024), LFQ (Yu et al., 2024a), and BSQ (Zhao et al., 2025). We replace the quantization module with the corresponding scalar quantizer and keep the surrounding architecture consistent with the VQ and PQ implementations. LFQ quantizes each 16-dimensional token using binary scalar levels, giving code-space cardinality $2^{16} = 65{,}536$. BSQ uses the same 16-dimensional tokens but normalizes them before binary spherical quantization; it has $2^{16} = 65{,}536$ possible codes. For FSQ, we project the 32-dimensional features to 16 dimensions with a three-layer convolutional projector, then arrange them as two 8-dimensional quantization tokens. Each scalar dimension has four fixed levels, so each token has code-space cardinality $4^8 = 65{,}536$. A symmetric three-layer convolutional projector maps the quantized features back to 32 dimensions. The resulting $T = 512$ tokens and effective per-token code-space cardinality $K = 65{,}536$ match those of VQ and PQ, giving the same nominal representation rate.

Pixel-Space Implementation. In addition to the latent-space experiments, we conduct pixel-space experiments on CelebA-HQ, where quantization is performed without a pretrained encoder-decoder. We apply PixelUnshuffle with factor 4 to the input image $\boldsymbol{x} \in \mathbb{R}^{256\times256\times3}$. It deterministically moves each 4×4 spatial block into the channels, giving a $64 \times 64 \times 48$ representation without discarding information. A three-layer convolutional projector maps the 48 channels to the quantization feature dimension, and the corresponding quantization module quantizes the projected features. The 64×64 grid gives $T = 4096$ tokens. We fix the code-space cardinality of each token at $K = 65{,}536$, giving nominal representation rate $R = T \log_2 K$. To reconstruct the image, a symmetric three-layer convolutional projector maps the quantized features back to 48 channels. PixelShuffle with factor 4 restores the original 256×256 resolution. We keep the PixelUnshuffle/PixelShuffle operations, convolutional projectors, token count, and code-space cardinality the same across methods. The quantization module is the primary component that varies.

Remark. We control the comparisons within each setting separately. Every latent-space method receives the same pretrained VAR encoder output and uses $T = 512$ tokens, with code-space cardinality $K = 65{,}536$ per token. In the pixel-space experiments, all methods share the same PixelUnshuffle operation and convolutional projector, and use $T = 4096$ quantization tokens with the same code-space cardinality $K = 65{,}536$. Therefore, the nominal representation rate $R = T \log_2 K$ is matched across quantization methods within each setting, satisfying Condition 2. With respect to Condition 1, all compared methods within each setting receive the same source representation before quantizer-specific processing, providing a controlled basis for comparing intrinsic quantization effectiveness.

Table 3: Quantization and reconstruction performance under varying per-token code-space cardinality K with the token count fixed at $T = 512$.

Methods	Phase	Tokens	Codebook Size K	$\mathcal{E}(\downarrow)$	U (↑)	PSNR(↑)	SSIM(↑)	LPIPS (↓)	r-FID(↓)	r-IS(↑)
MMD VQ	Substitution	512	1024	0.318	99.6%	23.75	60.8	0.162	2.68	167.6
MMD VQ	Substitution	512	2048	0.296	99.4%	24.12	62.4	0.159	2.59	168.8
MMD VQ	Substitution	512	4096	0.273	99.4%	24.41	63.5	0.141	1.96	178.0
MMD VQ	Substitution	512	8192	0.252	99.5%	24.67	64.6	0.135	1.85	181.0
MMD VQ	Substitution	512	16384	0.234	99.8%	24.89	65.4	0.130	1.84	183.7
MMD VQ	Substitution	512	32768	0.215	99.7%	25.06	66.3	0.126	1.79	184.8
MMD VQ	Substitution	512	65536	0.201	99.9%	25.24	66.8	0.121	1.69	187.3
MMD VQ	Adaptation	512	1024	0.318	99.6%	23.06	58.8	0.148	1.90	169.8
MMD VQ	Adaptation	512	2048	0.296	99.4%	23.58	61.2	0.137	1.63	176.6
MMD VQ	Adaptation	512	4096	0.273	99.4%	23.89	61.8	0.128	1.28	185.1
MMD VQ	Adaptation	512	8192	0.252	99.5%	24.11	62.9	0.121	1.18	187.9
MMD VQ	Adaptation	512	16384	0.234	99.8%	24.31	63.7	0.115	1.05	191.2
MMD VQ	Adaptation	512	32768	0.215	99.7%	24.53	64.7	0.110	0.97	194.1
MMD VQ	Adaptation	512	65536	0.201	99.9%	24.65	65.0	0.106	0.86	197.1

Table 4: Quantization and reconstruction performance under varying token count T with the per-token code-space cardinality fixed at $K = 16{,}384$.

Methods	Phase	Tokens	Codebook Size K	$\mathcal{E}(\downarrow)$	U (↑)	PSNR(↑)	SSIM(↑)	LPIPS (↓)	r-FID(↓)	r-IS(↑)
MMD VQ	Substitution	256	16384	0.369	99.6%	22.97	57.2	0.194	4.91	141.4
MMD VQ	Substitution	512	16384	0.234	99.8%	24.89	65.4	0.130	1.84	183.7
MMD VQ	Substitution	1024	16384	0.098	100%	26.40	71.0	0.100	2.01	191.7
MMD VQ	Substitution	2048	16384	0.035	100%	27.16	73.1	0.089	2.36	192.2
MMD VQ	Adaptation	256	16384	0.369	99.6%	22.41	55.9	0.171	3.06	148.9
MMD VQ	Adaptation	512	16384	0.234	99.8%	24.31	63.7	0.115	1.05	191.2
MMD VQ	Adaptation	1024	16384	0.098	100%	26.03	69.6	0.079	0.54	210.1
MMD VQ	Adaptation	2048	16384	0.035	100%	27.31	73.2	0.060	0.42	217.0

Table 5: Comparison of different token-count and code-space allocations under matched nominal rate.

Methods	Phase	Tokens	Codebook Size K	Rate R	$\mathcal{E}(\downarrow)$	U (↑)	PSNR(↑)	SSIM(↑)	LPIPS (↓)	r-FID(↓)	r-IS(↑)
MMD VQ	Substitution	512	16384	7,168 bits	0.234	99.8%	24.89	65.4	0.130	1.84	183.7
MMD VQ	Adaptation	512	16384	7,168 bits	0.234	99.8%	24.31	63.7	0.115	1.05	191.2
MMD VQ	Substitution	1024	128	7,168 bits	0.243	100%	24.72	64.9	0.132	1.80	184.5
MMD VQ	Adaptation	1024	128	7,168 bits	0.243	100%	24.01	63.0	0.118	1.10	190.4
MMD VQ	Substitution	512	65536	8,192 bits	0.201	99.9%	25.24	66.8	0.121	1.69	187.3
MMD VQ	Adaptation	512	65536	8,192 bits	0.201	99.9%	24.65	65.0	0.106	0.86	197.1
MMD VQ	Substitution	1024	256	8,192 bits	0.212	100%	25.11	66.4	0.124	1.67	187.6
MMD VQ	Adaptation	1024	256	8,192 bits	0.212	100%	24.51	64.6	0.108	0.96	195.3

I Rate-Distortion Analysis of Code-Space Cardinality and Token Count

We first vary the code-space cardinality K while fixing the token count at $T = 512$, increasing the representation budget under the nominal fixed-length rate $R = T \log_2 K$. The per-token cardinality doubles successively from $K = 1{,}024$ to $K = 65{,}536$. As shown in Table 3, quantization distortion $\mathcal{E}$ decreases monotonically from 0.318 to 0.201 across this range, while r-FID after decoder adaptation improves from 1.90 to 0.86.

We next vary the token count from $T = 256$ to $T = 2{,}048$ while fixing $K = 16{,}384$. As shown in Table 4, increasing T reduces distortion from 0.369 to 0.035, while adapted r-FID improves from 3.06 to 0.42. Increasing token count can also substantially reduce information loss. The improvements differ when we vary T and K, suggesting that the two may change the representation structure differently even though each contributes to the nominal rate.

We also compare allocations that keep the nominal representation rate fixed. Under $R = T \log_2 K$, doubling the token count and squaring the per-token code-space cardinality produce the same nominal rate:

$$2T \log_2 K = T \log_2 K^2. \tag{21}$$

This identity concerns only the representation budget and does not imply identical distortion. Table 5 compares matched-rate configurations, including $(T, K) = (512, 16{,}384)$ versus $(1{,}024, 128)$ and $(512, 65{,}536)$ versus $(1{,}024, 256)$. Distortion and reconstruction performance are close within each pair, but not identical. Consistent with the rate accounting in Section 4.1, different allocations of the same nominal budget between token count and code-space cardinality can induce different representation structures.

Table 6: Quantization and reconstruction performance of VQ, PQ, and SQ on FFHQ under matched token count ($T = 512$) and code-space cardinality ($K = 65{,}536$). Within each quantization type and phase (Substitution/Adaptation), optimal values are underlined; overall best results per metric are **bold**.

Approaches	Types	Phase	Tokens	K	$\mathcal{E}(\downarrow)$	U (↑)	PSNR(↑)	SSIM(↑)	LPIPS (↓)	r-FID(↓)
Vanilla VQ	VQ	Substitution	512	65536	0.286	0.2%	24.20	66.7	0.159	7.16
EMA VQ	VQ	Substitution	512	65536	0.146	52.9%	27.66	77.4	0.075	2.13
Online VQ	VQ	Substitution	512	65536	0.186	12.4%	26.98	75.2	0.088	2.41
Wasserstein VQ	VQ	Substitution	512	65536	0.132	99.5%	28.05	**78.3**	0.070	2.13
MMD VQ	VQ	Substitution	512	65536	**0.131**	99.3%	**28.06**	**78.3**	**0.069**	2.14
Vanilla VP2	PQ	Substitution	512	65536	0.159	40.3%	27.33	76.7	0.078	2.38
EMA VP2	PQ	Substitution	512	65536	0.136	**100%**	27.94	77.9	0.071	2.14
Online VP2	PQ	Substitution	512	65536	0.137	**100%**	27.89	77.9	0.072	2.43
Wasserstein VP2	PQ	Substitution	512	65536	0.146	99.9%	27.64	77.6	0.079	2.95
MMD VP2	PQ	Substitution	512	65536	0.139	**100%**	27.83	77.9	0.072	2.47
FSQ	SQ	Substitution	512	65536	0.170	**100%**	27.27	75.4	0.081	3.18
LFQ	SQ	Substitution	512	65536	0.224	7.2%	25.62	71.6	0.110	3.70
BSQ	SQ	Substitution	512	65536	0.155	**100%**	27.04	76.7	0.080	**2.09**
Vanilla VQ	VQ	Adaptation	512	65536	0.286	0.2%	23.63	63.8	0.142	2.31
EMA VQ	VQ	Adaptation	512	65536	0.146	52.9%	27.35	75.9	0.071	1.25
Online VQ	VQ	Adaptation	512	65536	0.186	12.4%	26.57	73.3	0.085	1.77
Wasserstein VQ	VQ	Adaptation	512	65536	0.132	99.5%	27.69	76.5	**0.067**	0.89
MMD VQ	VQ	Adaptation	512	65536	**0.131**	99.3%	**27.75**	**76.9**	**0.067**	**0.85**
Vanilla VP2	PQ	Adaptation	512	65536	0.159	40.3%	27.11	75.2	0.074	1.33
EMA VP2	PQ	Adaptation	512	65536	0.136	**100%**	27.57	76.1	0.070	1.05
Online VP2	PQ	Adaptation	512	65536	0.137	**100%**	27.54	75.8	0.069	1.10
Wasserstein VP2	PQ	Adaptation	512	65536	0.146	99.9%	27.28	75.5	0.072	1.11
MMD VP2	PQ	Adaptation	512	65536	0.139	**100%**	27.58	76.4	0.068	1.16
FSQ	SQ	Adaptation	512	65536	0.170	**100%**	26.83	73.5	0.076	1.61
LFQ	SQ	Adaptation	512	65536	0.224	7.2%	25.23	69.9	0.102	1.74
BSQ	SQ	Adaptation	512	65536	0.155	**100%**	26.73	75.1	0.078	1.54

Table 7: Quantization and reconstruction performance of VQ, PQ, and SQ on CelebA-HQ under matched token count ($T = 512$) and code-space cardinality ($K = 65{,}536$). Within each quantization type and phase (Substitution/Adaptation), optimal values are underlined; overall best results per metric are **bold**.

Approaches	Types	Phase	Tokens	K	$\mathcal{E}(\downarrow)$	U (↑)	PSNR(↑)	SSIM(↑)	LPIPS (↓)	r-FID(↓)
Vanilla VQ	VQ	Substitution	512	65536	0.269	0.1%	23.39	64.9	0.176	16.2
EMA VQ	VQ	Substitution	512	65536	0.131	31.9%	27.87	77.4	0.071	3.17
Online VQ	VQ	Substitution	512	65536	0.180	11.6%	26.82	74.4	0.093	4.44
Wasserstein VQ	VQ	Substitution	512	65536	**0.111**	99.3%	28.32	**78.7**	**0.064**	2.63
MMD VQ	VQ	Substitution	512	65536	**0.111**	99.1%	**28.33**	78.5	**0.064**	2.66
Vanilla VP2	PQ	Substitution	512	65536	0.168	11.6%	26.67	75.1	0.086	3.78
EMA VP2	PQ	Substitution	512	65536	0.115	**100%**	28.25	78.5	0.065	3.03
Online VP2	PQ	Substitution	512	65536	0.119	99.9%	28.14	78.3	0.066	3.01
Wasserstein VP2	PQ	Substitution	512	65536	0.123	99.9%	27.82	77.7	0.068	2.81
MMD VP2	PQ	Substitution	512	65536	0.120	**100%**	27.99	78.1	0.069	2.92
FSQ	SQ	Substitution	512	65536	0.147	99.9%	27.58	75.9	0.075	4.64
LFQ	SQ	Substitution	512	65536	0.211	2.9%	25.40	71.1	0.115	6.56
BSQ	SQ	Substitution	512	65536	0.133	**100%**	27.28	77.0	0.074	**2.58**
Vanilla VQ	VQ	Adaptation	512	65536	0.269	0.1%	22.95	62.4	0.159	5.13
EMA VQ	VQ	Adaptation	512	65536	0.131	31.9%	27.53	75.6	0.069	2.34
Online VQ	VQ	Adaptation	512	65536	0.180	11.6%	26.30	72.0	0.090	3.46
Wasserstein VQ	VQ	Adaptation	512	65536	**0.111**	99.3%	**27.98**	76.6	0.064	**1.73**
MMD VQ	VQ	Adaptation	512	65536	**0.111**	99.1%	**27.98**	**76.7**	**0.063**	1.76
Vanilla VP2	PQ	Adaptation	512	65536	0.168	11.6%	26.37	72.8	0.083	2.36
EMA VP2	PQ	Adaptation	512	65536	0.115	**100%**	27.78	76.0	0.068	1.96
Online VP2	PQ	Adaptation	512	65536	0.119	99.9%	27.67	75.3	0.068	2.11
Wasserstein VP2	PQ	Adaptation	512	65536	0.123	99.9%	27.67	75.9	0.068	2.02
MMD VP2	PQ	Adaptation	512	65536	0.120	**100%**	27.68	75.8	0.066	2.07
FSQ	SQ	Adaptation	512	65536	0.147	99.9%	27.16	73.6	0.072	2.24
LFQ	SQ	Adaptation	512	65536	0.211	2.9%	24.88	68.1	0.111	3.35
BSQ	SQ	Adaptation	512	65536	0.133	**100%**	26.93	74.9	0.074	2.47

J GENERALIZATION ACROSS DATASETS

To examine whether the main empirical findings generalize across data distributions, we also evaluate VQ, PQ, and SQ on FFHQ (Karras et al., 2019) and CelebA-HQ (Karras et al., 2017). We use the controlled latent-space protocol from Appendix H. Every method has $T = 512$ and $K = 65{,}536$; the results are in Tables 6 and 7.

Generalization of Quantization Comparisons. The results are consistent with the main experimental findings reported in Table 2. Modern VQ methods retain their advantage on FFHQ and CelebA-HQ under controlled feature statistics and matched nominal rate. The best distortions for VQ, PQ, and SQ, respectively, are 0.131, 0.136, and 0.155 on FFHQ, and 0.111, 0.115, and 0.133 on CelebA-HQ. After decoder adaptation, the best r-FID values in the same family order are 0.85, 1.05, and 1.54 on FFHQ, and 1.73, 1.96, and 2.24 on CelebA-HQ.

Generalization of Distortion-Reconstruction Correlation. We compute Spearman's rank correlations of r-FID with quantization distortion $\mathcal{E}$ and codebook utilization U. The association is much stronger for distortion on FFHQ: $\rho = 0.979$, $p = 5.49\times10^{-9}$, compared with $\rho = -0.492$, $p = 0.088$ for utilization. On CelebA-HQ, the corresponding correlations are $\rho = 0.944$ ($p = 1.29 \times 10^{-6}$) and $\rho = -0.559$ ($p = 0.047$). In all these datasets, including ImageNet-1K in the main text, distortion is more strongly associated with reconstruction fidelity. This favors distortion minimization over utilization as a reliable criterion for intrinsic quantization effectiveness.

K Controlled Quantization Comparison in Pixel Space

To examine whether our conclusions depend on the particular latent representation induced by a pretrained encoder, we conduct an additional controlled comparison directly from image pixels on CelebA-HQ (Karras et al., 2017). Quantizing from pixels removes dependence on a pretrained latent feature distribution and reduces this potential source of confounding. We use the same PixelUnshuffle operation and convolutional projectors for every method, as described in Appendix H. Token count is fixed at $T = 4096$ and code-space cardinality at $K = 65{,}536$. With this common representation pipeline and nominal rate, the quantization modules are the primary difference between methods. Table 8 reports the results.

Quantization and Reconstruction Performance. MMD VQ has the lowest distortion in pixel space, as in the latent-space experiments: $\mathcal{E} = 0.0021$, compared with 0.0023 for the best PQ variant and 0.0032 for the best SQ variant. It also achieves the best r-FID of 3.64, compared with 3.77 for the best PQ method and 4.54 for the best SQ method. More broadly, modern VQ variants achieve strong reconstruction performance despite PQ and SQ methods often attaining higher codebook utilization. These results indicate that the relative advantage of VQ is not specific to the pretrained latent space used in the main experiments.

Distortion-Reconstruction Correlation. We further compute Spearman's rank correlations between r-FID and quantization distortion $\mathcal{E}$ or codebook utilization U. Distortion shows a strong positive correlation with r-FID ($\rho = 0.947$, $p = 2.91 \times 10^{-6}$), whereas codebook utilization shows a substantially weaker and statistically insignificant negative correlation ($\rho = -0.413$, $p = 0.182$). The stronger association of distortion with reconstruction fidelity also holds in this setting, which does not use a pretrained encoder. The pixel-space results qualitatively match the latent-space results and support the same conclusions.

Table 8: Quantization and reconstruction performance of VQ, PQ, and SQ methods on CelebA-HQ under the controlled pixel-space setting with matched token count ($T = 4096$) and code-space cardinality ($K = 65{,}536$). Within each quantization type, the best results are underlined, while the overall best result for each metric is shown in **bold**. Results for Online VQ are omitted because the experiment could not be completed due to out-of-memory errors.

Approaches	Types	Tokens	K	$\mathcal{E}(\downarrow)$	U (↑)	PSNR(↑)	SSIM(↑)	LPIPS (↓)	r-FID(↓)
Vanilla VQ	VQ	4096	65536	0.0055	0.2%	30.99	86.1	0.085	8.45
EMA VQ	VQ	4096	65536	0.0033	4.2%	33.33	89.4	0.057	6.51
Online VQ	VQ	4096	65536	-	-	-	-	-	-
Wasserstein VQ	VQ	4096	65536	0.0022	96.0%	35.32	92.9	0.026	3.75
MMD VQ	VQ	4096	65536	**0.0021**	98.5%	**35.37**	92.9	**0.023**	**3.64**
Vanilla VP2	PQ	4096	65536	0.0024	**100%**	34.71	**93.0**	0.025	3.93
EMA VP2	PQ	4096	65536	0.0027	**100%**	34.12	90.7	0.057	6.37
Online VP2	PQ	4096	65536	0.0023	**100%**	34.96	92.7	0.027	4.00
Wasserstein VP2	PQ	4096	65536	0.0024	99.2%	34.88	92.5	0.027	4.21
MMD VP2	PQ	4096	65536	0.0023	**100%**	35.15	92.9	0.025	3.77
FSQ	SQ	4096	65536	0.0032	**100%**	33.40	91.8	0.031	4.54
LFQ	SQ	4096	65536	0.0050	10.1%	31.27	87.2	0.071	7.99
BSQ	SQ	4096	65536	0.0041	99.8%	32.15	90.1	0.039	4.75