VP-VAE: RETHINKING VECTOR QUANTIZATION VIA ADAPTIVE VECTOR PERTURBATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vector-quantized autoencoders (VQ-VAEs) are the standard way to turn continuous signals into discrete tokens, yet their training is unstable and prone to codebook collapse. We argue that a central cause is the *coupling* between continuous representation learning and discrete codebook optimization, which forces the encoder to chase a moving codebook through non-differentiable, surrogate-gradient updates. To tackle this problem, we propose **VP-VAE**, a *perturbation-first* training paradigm that decouples representation learning from discretization by removing the dependence on codebook optimization during training. Our key insight is that, from the decoder's viewpoint, quantization primarily manifests as a bounded, local perturbation in the latent space. Instead of quantizing during training, we therefore train the decoder to be robust to latent perturbations that emulate the effect of future quantization, requiring those perturbations to be *scale-aligned* with the target codebook capacity and *distribution-consistent* with the learned latent manifold. We realize this paradigm through two instantiations. As a general prototype, we introduce an adaptive vector perturbation mechanism via Metropolis–Hastings sampling, which validates the paradigm without imposing any structural prior on the latent space. As a lightweight and practical solution, we further derive **FSP** (Finite Scalar Perturbation) under an approximately uniform latent prior, which offers a principled theoretical interpretation of, and a consistent empirical improvement over, FSQ-style fixed quantizers. Extensive experiments on image and audio tokenization show that both VP-VAE instantiations consistently improve reconstruction fidelity and maintain stable and balanced codebook utilization over competing baselines, across both in-domain and out-of-distribution benchmarks. These results demonstrate that our perturbation-first paradigm is a stable and effective alternative to conventional quantize-during-training pipelines. Code is available at `https://anonymous.4open.science/r/VP-VAE-4DD5/`.

1 INTRODUCTION

Discrete tokenization of continuous signals—such as images, audio, and video—has emerged as a cornerstone of modern representation learning and generative modeling. By mapping high-dimensional continuous data into compact discrete sequences, this paradigm enables scalable sequence architectures, predominantly Transformers (Chen et al., 2025; Ye et al., 2024; Esser et al., 2021), to be directly applied to diverse tasks ranging from high-fidelity reconstruction to cross-modal synthesis (Ramesh et al., 2021; Chang et al., 2022; Adiban et al., 2025; Yu et al., 2024; Borsos et al., 2023; Brendel et al., 2024a). Vector-Quantized VAEs (VQ-VAEs) (Van Den Oord et al., 2017) and their variants (Yu et al., 2021; Zeghidour et al., 2021; Chiu et al., 2022; Mentzer et al., 2023; Huh et al., 2023; Fifty et al., 2024; Zhu et al., 2025) remain the dominant way to learn such discrete representations.

Despite their success, optimizing VQ-VAEs is often unstable, and we argue that the root cause is the inherent *coupling* between representation learning and codebook learning. During training, encoder outputs are assigned to their nearest codes through a non-differentiable $\arg\min$ operator, while the codes themselves must simultaneously move to fit the evolving distribution of encoder outputs (Zheng et al., 2026). To allow gradient-based optimization, surrogate gradient estimators—most commonly the straight-through estimator (STE) (Bengio et al., 2013)—are widely adopted. While STE enables backpropagation, it introduces a gradient mismatch: the forward pass applies a non-differentiable discretization, whereas the backward pass relies on a continuous straight-through approximation.

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.

This gradient mismatch, together with non-stationary optimization targets induced by concurrent codebook updates, makes training sensitive and difficult to scale. A prominent failure mode is *codebook collapse* (or dead codes): a small subset of codes is selected repeatedly, while rarely used codes receive few or no updates and thus become even less likely to be selected later. The imbalance is self-reinforcing: it shrinks the usable discrete capacity, lowers the effective bitrate, and ultimately degrades reconstruction fidelity.

Existing efforts broadly follow two directions, each addressing the above issues only partially. The first stabilizes the joint optimization of the autoencoder and the codebook with additional regularization, reparameterizations, or training heuristics such as dead-code resets (Yu et al., 2021; Ramesh et al., 2021; Zeghidour et al., 2021; Takida et al., 2022; Huh et al., 2023; Zheng & Vedaldi, 2023; Fifty et al., 2024; Zhu et al., 2025). These measures mitigate collapse to some extent, but they introduce extra objectives and hyperparameters, and the underlying coupling remains. The second direction removes codebook learning by fixing the quantizer in advance, either as a scalar lattice (FSQ (Mentzer et al., 2023), LQ (Hsu et al., 2023), LFQ (Yu et al., 2024), BSQ (Zhao et al., 2024)) or as a predefined Gaussian grid (TokenBridge (Wang et al., 2024)). Fixed quantizers train stably, but they impose a rigid geometric prior: when the latent distribution a model would naturally learn deviates from the assumed grid geometry, the quantization capacity can be inefficiently used, e.g., many tokens carry redundant probability mass while other regions remain under-represented.

This paper revisits a more fundamental question: *is explicit quantization during training necessary at all?* More specifically, can one first learn a robust continuous encoder–decoder representation, without committing to a training-time codebook, and instantiate the discrete tokenizer only after the representation has largely converged? Our starting point is a simple observation: from the decoder's viewpoint, nearest-neighbor quantization primarily appears as the resulting *quantization error*, i.e., a bounded, local perturbation in latent space. This suggests a **perturbation-first** training paradigm. Instead of explicitly quantizing latent vectors during training—and inheriting the optimization difficulties of a moving codebook and surrogate gradients—we train the decoder to be robust to latent perturbations that emulate the effect of future quantization. If the perturbations are chosen appropriately, the model can be trained with no explicit codebook and still remain compatible with discretization at inference time.

For this perturbation-first paradigm to succeed, the perturbations must be both *scale-aligned* with the quantization error implied by the target codebook capacity and *distribution-consistent* with the learned latent manifold. Naive noise injection satisfies neither: isotropic Gaussian perturbation ignores the local geometry of the latent space and tends to push samples into unrealistic, low-density regions, forcing the decoder to model latent configurations that the data never produces. We therefore propose two instantiations of the paradigm, corresponding to different assumptions about the latent space.

In the general case, where no strong structural prior is imposed on the latent distribution, we instantiate VP-VAE using an adaptive vector perturbation mechanism based on Metropolis–Hastings (MH) sampling with non-parametric density estimation. This **MH-based VP-VAE** generates *scale-adaptive* and *distribution-consistent* perturbations, so that the perturbed latent vectors remain close to the empirical latent manifold while reflecting the expected magnitude of quantization error. As a result, the encoder–decoder can be trained stably without an explicit codebook, and a codebook can be generated only after convergence.

Under a stronger prior, the same principle admits a substantially simpler instantiation. When each latent dimension is constrained to be approximately uniform over a bounded interval, the perturbation mechanism reduces to a lightweight scalar form, which we call **FSP** (Finite Scalar Perturbation). We show that FSP connects naturally to Lloyd–Max optimal scalar quantization: instead of rounding to grid boundaries, it corresponds to training with bounded centered perturbations and decoding with interval centroids. This yields both a unified theoretical interpretation of FSQ-style fixed quantizers (Mentzer et al., 2023) and a practical improvement over them.

We evaluate both instantiations on image and audio tokenization tasks. Across in-domain evaluations and cross-dataset out-of-distribution (OOD) benchmarks, both the MH-based prototype and FSP improve reconstruction fidelity (for example, better rKID on images and higher STOI on audio) while achieving more balanced codebook utilization.

Our contributions are summarized as follows:

- We propose VP-VAE, a perturbation-first training paradigm that decouples representation learning from discretization by replacing training-time quantization with latent perturbations that emulate quantization error.
- We instantiate the paradigm in a general setting via an MH-based mechanism that injects *scale-adaptive* and *distribution-consistent* vector perturbations, enabling codebook-free training with post-training discretization.
- We derive FSP, a lightweight instantiation under approximately uniform bounded latents, which provides both theoretical insight into and consistent empirical improvements over FSQ-style fixed quantizers.
- Experiments on image and audio tokenization, in-domain and out-of-distribution, show that both instantiations improve reconstruction fidelity over coupled and fixed-quantizer baselines while keeping codebook utilization balanced—most notably on audio, where the coupled baselines we evaluate suffer severe codebook collapse.

2 Related Work

Existing approaches to neural discrete representation learning mainly fall into two categories: methods that mitigate the instability arising from joint model–codebook optimization, and methods that bypass this coupling by relying on fixed or predefined quantization targets.

Coupled codebook–network optimization. The foundational VQ-VAE (Van Den Oord et al., 2017) introduced the concept of learning a discrete latent bottleneck via nearest-neighbor lookup, utilizing the Straight-Through Estimator (STE) to approximate gradients for the encoder. As discussed in Section 1, this paradigm suffers inherently from the 'codebook collapse' phenomenon, since the codebook update (moving codes toward encoder outputs) and the encoder update (mapping inputs to codes) are coupled in a feedback loop that rewards already-frequent codes, often resulting in low effective bitrate and sub-optimal reconstruction (Takida et al., 2022; Mentzer et al., 2023).

To mitigate this, substantial efforts have focused on regularization, continuous relaxations, and heuristic stabilization. Early works employed exponential moving average (EMA) updates and restart heuristics—manually re-initializing dead codes with active encoder outputs—to maintain high codebook utilization (Razavi et al., 2019; Zheng & Vedaldi, 2023), while more recent approaches refine gradient estimation and update dynamics: SQ-VAE (Takida et al., 2022) replaces quantization with stochastic Gaussian smoothing, and SimVQ (Zhu et al., 2025) reparameterizes the codebook via a learnable linear layer so that gradients flow to all codes simultaneously.

Despite these advances, such methods still retain explicit training-time quantization and thus the coupled dynamics described above, which often necessitates additional losses (e.g., commitment, orthogonality, or entropy regularization) and careful hyperparameter tuning (Chang et al., 2022; Zhu et al., 2025). In contrast, VP-VAE does not attempt to further optimize a training-time quantizer: it does not depend on explicit quantization from the representation-learning phase and instead learns robustness to future quantization error.

Predefined codebook quantization. Recognizing the drawbacks of joint optimization, a parallel line of research explores decoupling representation learning from codebook generation by imposing fixed geometric priors. RandomVQ (Chiu et al., 2022) demonstrated that, in self-supervised speech learning, a randomly initialized and frozen codebook can still yield competitive representations, suggesting that the precise location of codes may matter less than the discretization mechanism itself.

Building on this, FSQ (Mentzer et al., 2023) and LFQ (Yu et al., 2024) eliminate the vector lookup entirely and instead project latent dimensions onto fixed scalar grids that form a hypercube. By removing the learnable codebook, these methods achieve high stability and near-100% utilization, but they impose a rigid bias, assuming the encoder can reshape the latent distribution into a factorial uniform grid; if the natural latent distribution is non-uniform or complex, such rigid grids lead to inefficient space allocation. Most recently, TokenBridge (Wang et al., 2024) discretizes the latent space of a pre-trained KL-regularized VAE using fixed Gaussian percentiles, which can be viewed as a *predefined Gaussian grid*. Like the scalar grid methods above, it shifts the burden of adaptation entirely onto the encoder, forcing the latent distribution to conform to a fixed prior.

VP-VAE offers a more flexible and theoretically grounded alternative: our general prototype, MH-based VP-VAE, does not need to impose such global distributional constraints on the latent space

(e.g., strict uniform or Gaussian priors), and instead utilizes local perturbations dynamically adapted to the empirical latent density. Furthermore, FSP arises as a lightweight instantiation when an approximately uniform bounded latent prior is intentionally imposed, from which the corresponding scalar discretization naturally follows.

3 METHODOLOGY

This section presents VP-VAE, a perturbation-based framework for discrete representation learning. We begin with the common setup and the core principle of replacing training-time quantization with latent perturbations (Section 3.1), then give a general MH-based instantiation (Section 3.2) and a lightweight scalar instantiation, FSP (Section 3.3), and finally describe the discretization procedures used at inference time (Section 3.4).

3.1 VP-VAE: FROM QUANTIZATION TO PERTURBATION

Classical quantization pipeline. Let x denote an input signal (e.g., image or audio). An encoder E maps x to a sequence of token features $\{h_t\}_{t=1}^{T}$, where $h_t \in \mathbb{R}^C$ and T is the number of tokens per sample. To make local perturbation design tractable, the perturbation or discretization is performed in a low-dimensional bottleneck. Specifically, each token is projected to a d-dimensional latent $z_t = P_{\downarrow}(h_t) \in \mathbb{R}^d$, quantized as $\tilde{z}_t = q(z_t)$, and then mapped back to the embedding space as $\tilde{h}_t = P_{\uparrow}(\tilde{z}_t)$, where $P_{\downarrow}$ and $P_{\uparrow}$ are learnable projections and $q(\cdot)$ denotes the quantization process. The decoder finally reconstructs the input as $\hat{x} = D(\{\tilde{h}_t\}_{t=1}^{T})$. In the remainder of the section, we omit the token index t when discussing a single latent vector.

From quantization to perturbation. In the above vector quantization procedure, the latent vector z is mapped to a discrete representative $\tilde{z}$, inducing a quantization error $\epsilon = \tilde{z} - z$. From the decoder's viewpoint, the primary effect of quantization is therefore a bounded, local perturbation of the latent representation. This observation suggests a different training strategy: rather than quantizing during training and approximating its gradients, we train the model with a perturbation operator $\mathcal{T}$ that emulates the effect of future quantization, yielding $\tilde{z} = \mathcal{T}(z; G)$. Here, G denotes the *target codebook size*, i.e., the desired discrete capacity of the deployed tokenizer. Importantly, our goal is not to smooth a fixed training-time quantizer, but to shift the burden of training from explicit codebook optimization to appropriately designed perturbations; the deployed discretizer is then specified separately for inference.

For this perturbation-first principle to be effective, $\mathcal{T}$ should satisfy two critical requirements. (1) ***Scale alignment***: the perturbation magnitude should match the quantization error expected at the target codebook size G; if perturbations are too small, the decoder is under-prepared for inference-time discretization, and if they are too large, reconstruction quality is unnecessarily degraded. (2) ***Distribution consistency***: perturbed latents should remain in high-probability regions of the original latent distribution, since otherwise the decoder is forced to model unrealistic latent configurations that do not arise from the data.

We realize these two requirements in two instantiations—the general MH-based prototype (Section 3.2) and the simplified FSP approach (Section 3.3)—which differ in the perturbation operator $\mathcal{T}$ but are both trained with the objective $\mathcal{L} = \mathcal{L}_{\text{rec}}(x, \hat{x}) + \mathcal{L}_{\text{reg}}$, where $\mathcal{L}_{\text{rec}}$ is a task-specific reconstruction loss and $\mathcal{L}_{\text{reg}}$ an instantiation-specific regularizer that stabilizes the latent space. In the next two subsections, we specify $\mathcal{T}$ and $\mathcal{L}_{\text{reg}}$, respectively.

3.2 MH-BASED VP-VAE

We first consider the general setting where no global prior is imposed on the latent distribution. This serves as a conceptual prototype demonstrating the viability of our perturbation-first paradigm. In this case, both the perturbation scale and the accepted perturbations should adapt to the empirical local geometry of the latent space. MH-based VP-VAE achieves this by combining non-parametric radius estimation with a Metropolis–Hastings (MH) accept–reject step. Specifically, we maintain a first-in-first-out (FIFO) queue $\mathcal{S} = \{s_i\}_{i=1}^{|\mathcal{S}|}$ of recent latent vectors in $\mathbb{R}^d$ to approximate the current latent distribution. To reduce redundancy caused by strong within-sample token correlations, only a random subset of tokens from each minibatch is inserted into $\mathcal{S}$.

3.2.1 SCALE ALIGNMENT VIA ADAPTIVE RADIUS ESTIMATION

Let $D_m(z \mid \mathcal{S})$ denote the Euclidean distance from z to its m-th nearest neighbor in $\mathcal{S}$. For a target codebook size G, we define

$$M = \left\lceil \frac{|\mathcal{S}|}{G} \right\rceil, \qquad R(z) = \eta \, D_M(z \mid \mathcal{S}), \tag{1}$$

where $\eta > 0$ is a scale hyperparameter. The intuition is that, under a quantizer with G cells of comparable probability mass, each cell would contain roughly $|\mathcal{S}|/G$ samples from the queue, so the distance to the M-th nearest neighbor is the radius of a local region enclosing the same number of samples and therefore provides a data-adaptive estimate of the local quantization scale (made precise in Appendix B.1).

3.2.2 DISTRIBUTION CONSISTENCY VIA METROPOLIS–HASTINGS SAMPLING

A naive way to perturb z would be to sample uniformly from a ball of radius $R(z)$. However, such isotropic noise ignores the geometry of the latent distribution and may push samples into low-density regions. To retain local plausibility, we use a Metropolis–Hastings transition whose stationary distribution is defined by an empirical density estimator.

Target density estimation. We approximate the latent density by a k-nearest-neighbor estimator:

$$\pi(z) \propto \frac{1}{\big(D_k(z \mid \mathcal{S})\big)^d}, \tag{2}$$

where k is a small constant and d is the latent dimension. For brevity, we write $D_k(z) = D_k(z \mid \mathcal{S})$ and $D_M(z) = D_M(z \mid \mathcal{S})$ when the conditioning on $\mathcal{S}$ is clear.

Proposal distribution. Given the latent vector z, we propose $z' = z + u$ with $u \sim \mathrm{Unif}\big(\mathcal{B}(0, R(z))\big)$, i.e., z' is sampled uniformly from the d-dimensional ball centered at z with radius $R(z)$. With V_d denoting the volume of the unit ball in $\mathbb{R}^d$, the proposal density is equivalently given by

$$g(z' \mid z) = \frac{1}{V_d R(z)^d} \, \mathbf{1}[\|z' - z\|_2 \le R(z)] \, . \tag{3}$$

Acceptance probability. The MH acceptance probability is $\alpha(z, z') = \min\big(1, \ \pi(z')\, g(z \mid z') \,/\, [\pi(z)\, g(z' \mid z)]\big)$, which preferentially accepts moves toward higher-density regions while rejecting many moves into lower-density regions. For the uniform-ball proposal, the reverse density $g(z \mid z')$ is nonzero only if $\|z - z'\|_2 \le R(z')$; a proposal violating this condition is therefore rejected outright. When both support conditions hold, substituting Eqs. (1)–(3) makes the unit-ball volume V_d and the scale η cancel, and the acceptance probability reduces to the closed form (see derivation in Appendix B.3):

$$\alpha(z, z') = \min\left(1, \ \left(\frac{D_k(z) \cdot D_M(z)}{D_k(z') \cdot D_M(z')}\right)^d\right). \tag{4}$$

Perturbation operator. The final perturbed latent vector is

$$\tilde{z} = z' \ \text{ with probability } \alpha(z, z'), \qquad \tilde{z} = z \ \text{ otherwise.} \tag{5}$$

The acceptance rule corresponds to a one-step MH transition defined by the current empirical target density π, and therefore biases accepted perturbations toward high-density regions of the empirical latent space rather than low-density or unsupported areas. Algorithm 1 in Appendix A.2 summarizes the full per-token procedure.

Regularization for MH-based VP-VAE. For this instantiation, we use a latent normalization regularizer

$$\mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{norm}} = \lambda_1 \|\mu_{\text{batch}}(z)\|_2^2 + \lambda_2 \|\sigma^2_{\text{batch}}(z) - \mathbf{1}\|_2^2, \tag{6}$$

where $\mu_{\text{batch}}(z)$ and $\sigma^2_{\text{batch}}(z)$ are the minibatch mean and variance of the d-dimensional latents. This regularizer stabilizes distance-based scale estimation by preventing individual dimensions from dominating the Euclidean geometry. In implementation, kNN search, queue lookup, and MH accept–reject decisions are treated as detached operations during optimization.

3.3 A LIGHTWEIGHT INSTANTIATION: FSP

The general MH-based construction makes minimal assumptions about the latent distribution, but its non-parametric density estimation introduces additional training cost. With a stronger structural prior, the same perturbation-first principle can be simplified substantially, leading to a highly efficient and practical instantiation. We derive **FSP** (Finite Scalar Perturbation), a lightweight method obtained when each latent dimension is constrained to be approximately uniform on a bounded interval.

Approximate uniformization via a CDF-like activation. Let $a \in \mathbb{R}^d$ denote pre-activation latent variables and define $z = g(a) \in [0,1]^d$, where g is a smooth, monotone, 'CDF-like' activation (e.g., sigmoid or normal CDF). In the idealized case where each a_i follows a distribution whose CDF equals g, the probability integral transform implies that each z_i is uniform on $[0,1]$. In practice, a highly expressive neural network can automatically learn to shape the pre-activations a to approximate this required distribution. To stabilize the latent space numerically without obstructing this automatic learning process, we apply a standard normalization loss on the pre-activations:

$$\mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{norm}}^{\text{FSP}} = \lambda_1 \|\mu_{\text{batch}}(a)\|_2^2 + \lambda_2 \|\sigma_{\text{batch}}^2(a) - \sigma_g^2 \mathbf{1}\|_2^2. \tag{7}$$

The target variance σ_g^2 is not an arbitrary hyperparameter; it should match the theoretical variance expected by the chosen activation g. If σ_g^2 conflicts with g, the regularization would incorrectly bias the pre-activation scale, damaging the uniformity of z and severely reducing codebook utilization (as empirically verified in Appendix C.3). For example, under this matching principle, we use $\sigma_g^2 = 1$ for the normal-CDF activation, $\sigma_g^2 \approx 0.8225$ for the tanh-based activation $g(a) = (\tanh(a) + 1)/2$, and $\sigma_g^2 \approx 3.29$ for sigmoid.

Simplified perturbation. Under the idealized uniform-support model, the density is constant inside $[0,1]^d$ and zero outside. Hence, distribution consistency reduces to a support check, and scale alignment is naturally controlled by the quantization bin width. For dimension i with L_i scalar levels, we define the perturbation proposal

$$z_i' = z_i + u_i, \qquad u_i \sim \mathcal{U}\left(-\frac{\eta}{2L_i}, \frac{\eta}{2L_i}\right), \tag{8}$$

and accept it iff the perturbed vector remains within the valid support:

$$\tilde{z} = z' \text{ if } z' \in [0,1]^d, \qquad \tilde{z} = z \text{ otherwise.} \tag{9}$$

Unlike MH-based VP-VAE, this instantiation requires neither kNN search nor an accept–reject probability beyond the box constraint, making it inherently scalable and computationally lightweight.

3.4 INFERENCE-TIME DISCRETIZATION

Because the perturbation-first paradigm separates representation learning from the final deployed discretizer, the inference-time discretization rule depends only on the chosen instantiation.

MH-based VP-VAE. No explicit codebook is constructed during training. After convergence, we simply apply K-Means clustering with K-Means++ initialization (Arthur & Vassilvitskii, 2006) on $\mathcal{S}$ to obtain G centroids $\mathcal{Q} = \{c_j\}_{j=1}^G$. At inference time, each latent is quantized by nearest-centroid assignment, $q(z) = \arg\min_{c \in \mathcal{Q}} \|z - c\|_2$, after which the quantized latent is mapped back through $P_\uparrow$ and decoded.

FSP. For FSP, no post-training clustering is required. For a uniform scalar source on $[0,1]$ with L levels, the Lloyd–Max conditions imply that the optimal reconstruction values are the centroids of equal-width intervals, $\mathcal{C} = \{(\ell + 1/2)/L\}_{\ell=0}^{L-1}$ (Appendix B.4). Accordingly, the natural deployed scalar discretizer for FSP is

$$q_i(z_i) = \frac{\ell_i + 1/2}{L_i}, \qquad \ell_i = \text{clip}\big(\lfloor L_i z_i \rfloor, 0, L_i - 1\big), \tag{10}$$

which quantizes each dimension to the centroid of its interval before up-projection and decoding. Crucially, this continuous training followed by centroid-based decoding distinguishes FSP from FSQ (Mentzer et al., 2023), bridging continuous representation learning with mathematically grounded discrete bounds.

Table 1: **In-domain results.** Models are trained and evaluated on COCO (image) and LibriSpeech (audio). For brevity, "VP-VAE" denotes the MH-based instantiation of our perturbation-first paradigm. rKID is reported in units of $\times 10^3$. "–" indicates training failure due to severe codebook collapse.

Method	Codebook Size	Image (COCO)						Audio (LibriSpeech)		
		CVU	LPIPS ↓	PSNR ↑	SSIM ↑	rFID ↓	rKID ↓	CVU	PESQ ↑	STOI ↑
VQ-VAE	256	0.3679	0.2053	23.1662	0.6704	4.5068	2.2691	-	-	-
SimVQ	256	0.9429	**0.1846**	23.5356	0.6918	**4.0689**	1.7271	0.0040	1.2524	0.4112
TokenBridge	256	0.9888	0.6075	11.5386	0.2026	32.630	310.89	0.9888	1.1802	0.2881
FSQ	240	0.8607	0.2101	22.9390	0.6689	4.6458	3.5197	0.0960	1.2763	0.7973
FSP	240	0.8947	0.1879	**23.7818**	0.6984	4.2133	**1.6917**	0.1372	2.2339	0.9076
VP-VAE	256	0.8532	0.1882	23.6154	**0.6994**	4.2212	1.8161	0.1352	**2.2534**	**0.9080**
VQ-VAE	1024	0.3557	0.1821	23.727	0.6946	3.9725	**1.4321**	0.0010	1.4005	0.3628
SimVQ	1024	0.8919	**0.1662**	24.0102	0.7176	**3.7080**	1.4413	0.0010	1.3213	0.4025
TokenBridge	1024	0.9887	0.6078	11.5145	0.1984	35.844	338.12	0.9888	1.1801	0.3124
FSQ	1000	0.8059	0.1839	23.6006	0.6955	4.1220	2.4979	0.0513	2.1738	0.9001
FSP	1000	0.8631	0.1674	**24.2073**	0.7228	3.8445	1.6591	0.0810	**2.4458**	**0.9211**
VP-VAE	1024	0.8515	0.1711	23.9648	**0.7242**	3.8772	1.4401	0.0687	2.4380	0.9185
VQ-VAE	4096	0.2401	0.1693	24.1303	0.7245	3.8032	1.5537	-	-	-
SimVQ	4096	0.8604	**0.1497**	24.8338	0.7455	**3.4587**	1.3357	0.0002	1.2519	0.3972
TokenBridge	4096	0.9887	0.6086	11.5931	0.2011	100.66	394.25	0.9887	1.2249	0.3160
FSQ	4375	0.7967	0.1647	24.4481	0.7273	3.8028	2.2483	0.0294	2.3821	0.9182
FSP	4375	0.8341	0.1506	**24.8914**	0.7473	3.4990	**0.8339**	0.0368	**2.5206**	**0.9257**
VP-VAE	4096	0.8493	0.1540	24.6418	**0.7481**	3.6066	0.9097	0.0564	2.4314	0.9214
VQ-VAE	16384	0.2826	0.1489	24.5994	0.7555	3.4806	**0.8440**	-	-	-
SimVQ	16384	0.8276	**0.1387**	25.2429	0.7605	**3.1894**	1.0329	0.0001	1.2477	0.4436
TokenBridge	16384	0.9887	0.6143	11.4920	0.1963	61.309	297.75	0.9887	1.2328	0.3008
FSQ	15360	0.7921	0.1552	24.5842	0.7354	3.6222	1.6288	0.0149	2.4356	0.9225
FSP	15360	0.7383	0.1460	**25.2882**	**0.7609**	3.3457	1.1196	0.0205	2.5560	0.9317
VP-VAE	16384	0.8862	0.1450	25.0065	0.7588	3.3575	0.9056	0.0455	**2.5940**	**0.9334**

4 EXPERIMENTS

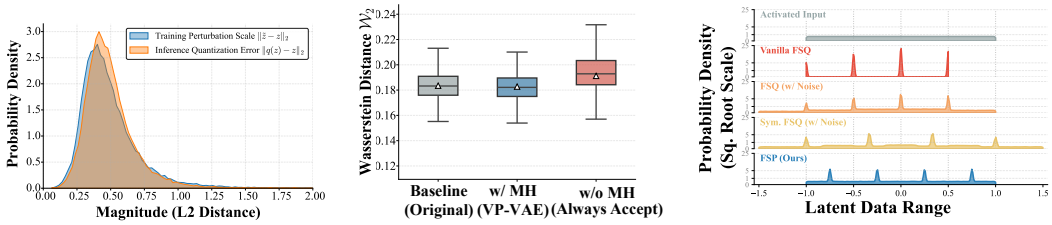
4.1 EXPERIMENTAL SETUP

We evaluate both instantiations of our perturbation-first paradigm—MH-based VP-VAE (denoted simply VP-VAE below) and FSP—on two representative tokenization tasks: image reconstruction on COCO 2017 (Lin et al., 2014) with the VQGAN encoder–decoder (Esser et al., 2021), and audio compression on LibriSpeech (Panayotov et al., 2015) with the SQCodec encoder–decoder (Zhai et al., 2025). Out-of-distribution (OOD) behavior is assessed by evaluating the same trained models, without any adaptation, on the ImageNet validation set (Deng et al., 2009) and the Common Voice English test set (Mozilla). We compare against four representative baselines spanning the two families reviewed in Section 2—VQ-VAE (Van Den Oord et al., 2017), SimVQ (Zhu et al., 2025), FSQ (Mentzer et al., 2023) (with its stronger noise-injected variant (Parker et al., 2024; Brendel et al., 2024b)), and TokenBridge (Wang et al., 2024)—at four target capacities $G \in \{256,\ 1024,\ 4096,\ 16384\}$.

Within each modality, all methods share the same encoder–decoder, reconstruction losses, optimizer, and training budget and differ only in the quantization or perturbation module; no method uses a discriminator, and the radius scale of VP-VAE is derived from the scale-alignment requirement of Section 3.1 rather than tuned. For reconstruction quality we report LPIPS (Zhang et al., 2018), PSNR, SSIM (Wang et al., 2004), rFID (Heusel et al., 2017), and rKID (Bińkowski et al., 2018) on images and PESQ (Rix et al., 2001) and STOI (Taal et al., 2010) on audio; for codebook utilization we report the Codebook Valid Usage (CVU), the perplexity of the empirical code-usage distribution normalized by G, which measures how uniformly the discrete capacity is used rather than merely how many codes are touched at least once. Datasets, per-method configurations, metric definitions, and the remaining protocol details are given in Appendix A.

4.2 MAIN BENCHMARK RESULTS

Cross-modality consistency and stability. A main finding is that VP-VAE and FSP remain effective across two markedly different modalities, whereas the baselines tend to over-specialize on one modality or fail outright on the other. As shown in Table 1, SimVQ is highly competitive on images—it attains the lowest LPIPS at every codebook size—yet this efficacy does not transfer to audio. On LibriSpeech, SimVQ exhibits near-collapse in code usage (CVU ≤ 0.004) and poor reconstruction (PESQ ≈ 1.3), on par with the only VQ-VAE setting that did not fail outright. We hypothesize that this instability in audio stems from the statistical nature of the audio signal: silent and low-amplitude segments produce highly concentrated encoder outputs early in training, and coupled codebook learning can amplify small early assignment biases into a self-reinforcing imbalance (cf.



(a) Scale alignment (b) Distribution consistency (c) Output distributions of FSQ/FSP

Figure 1: Mechanism-level evidence for the two design requirements and for the FSP decoding rule. **(a)** Scale alignment: the distribution of perturbation scales $\|\tilde{z} - z\|_2$ used during training against the inference-time quantization errors $\|q(z) - z\|_2$ induced by the deployed K-Means discretizer. **(b)** Distribution consistency: Wasserstein distance W_2 between the perturbed and the original latent distributions; "Baseline (Original)" reflects the intrinsic finite-sample discrepancy (same number of samples drawn from the original pool). **(c)** Latent output distributions induced by different quantization schemes (square-root scale, $L = 4$ levels): starting from the uniform activated input (top), we compare vanilla FSQ, noise-injected FSQ, symmetric noise-injected FSQ, and FSP.

Section 1), in the most severe cases leading to outright training failure (as observed for VQ-VAE at $G{\in}\{256, 4096, 16384\}$).

FSQ behaves differently: because its quantization grid is fixed, it remains stable on audio and avoids outright collapse, but the same rigid prior is less well matched to image latents. For example, at $G = 1024$ on COCO, FSQ reaches 23.60 dB PSNR, whereas FSP reaches 24.21 dB. Overall, the two instantiations of our paradigm offer the most consistent cross-modality behavior: on COCO, FSP achieves the best PSNR at *every* codebook size and either VP-VAE or FSP attains the best SSIM at every codebook size, while on the audio side VP-VAE and FSP jointly occupy the top two positions in both PESQ and STOI at every target capacity.

The two distribution-level image metrics show a consistent split. SimVQ attains the best rFID on COCO at all four codebook sizes, in line with its lowest LPIPS. On rKID, however, our perturbation-first methods are clearly stronger: FSP or VP-VAE attain the best rKID at $G{\in}\{256, 4096\}$ and remain within a narrow margin of the top at the other sizes. Since rKID is an unbiased (Bińkowski et al., 2018), kernel-based estimator that does not rely on the Gaussianity assumption underlying rFID (Heusel et al., 2017), this consistent advantage on rKID supports the distributional fidelity of our reconstructions.

TokenBridge provides an instructive point of comparison. Its CVU is nearly perfect in every setting (≈ 0.99), yet its reconstruction quality is far worse than all competitive methods (e.g., rKID exceeds 290 on COCO). This demonstrates that balanced token usage alone is insufficient: if the discretization geometry is mismatched to the latent distribution, high utilization does not translate into useful discrete capacity.

Out-of-distribution generalization. Table 2 in Appendix C.1 reports the same comparison under distribution shift, which is most informative when read jointly with the in-domain results. On PSNR, SimVQ and VP-VAE swap places across the domain boundary: in-domain on COCO, SimVQ matches or exceeds VP-VAE at most codebook sizes, whereas on ImageNet it is VP-VAE that matches or exceeds SimVQ at most sizes (e.g., 25.4548 vs. 25.3304 at $G{=}16384$). The audio results reinforce this picture: on Common Voice, VP-VAE and FSP again jointly occupy the top two positions in both PESQ and STOI at every codebook size, and both degrade only mildly under the shift (at $G{=}16384$, STOI drops by 0.034 for VP-VAE and 0.043 for FSP). We attribute this generalization advantage to our training objective: rather than fitting a specific set of codebook vectors, VP-VAE trains the decoder to tolerate a whole distribution of local perturbations, which regularizes the latent space so that inference-time quantization errors—even when slightly shifted by OOD inputs—remain within the decoder's learned tolerance.

FSP consistently improves over FSQ. Across *all* experimental conditions—both modalities, all four codebook sizes, and both in-domain and OOD evaluations—FSP outperforms FSQ on every reconstruction metric. Since FSP and FSQ share an identical fixed scalar grid and differ only in the training-time perturbation and the decoding rule, this cleanly isolates the benefit of the perturbation-first treatment, whose mechanism we analyze in Section 4.4.

4.3 Validating the Perturbation-First Principle

We dive into the MH-based prototype to verify that its two design requirements—*scale alignment* and *distribution consistency*—are indeed the operative mechanisms enabling codebook-free training.

Scale alignment. Figure 1a tests the core hypothesis that our designed perturbations can *replace* explicit training-time quantization, by overlaying the magnitude distribution of training-time perturbations $\|\tilde{z} - z\|_2$ onto that of the inference-time quantization errors $\|q(z) - z\|_2$. The close agreement between the two distributions indicates that our adaptive radius estimation places perturbations in the same local-error regime as the deployed discretizer, thereby satisfying the scale-alignment requirement.

Distribution consistency. The MH accept–reject step is designed to reject moves that would push a latent vector into low-density or unsupported regions. Figure 1b confirms its effect: removing the MH step (i.e., always accepting the proposal) substantially increases the Wasserstein distance between the perturbed and the original latent distributions, i.e., the perturbed latents drift off the empirical manifold.

Component-wise ablation. A component-wise ablation on COCO (G=4096), reported in full in Appendix C.2, confirms that both requirements are necessary: removing the perturbations altogether is by far the most damaging (PSNR drops from 24.64 to 22.22) even though post-training K-Means still yields a high CVU, replacing them with tuned fixed Gaussian noise degrades reconstruction fidelity, always accepting the proposals lowers CVU from 0.85 to 0.80, and dropping the latent normalization regularizer degrades every reconstruction metric.

4.4 Analyzing FSP: Mechanism and Efficiency

While MH-based VP-VAE validates our paradigm in a generalized setting, FSP shows that, under an approximately uniform latent prior, the same principle yields a highly efficient, lightweight tokenizer.

Performance gains over FSQ. Figure 1c gives an intuitive account of the gain over FSQ reported above. The top panel shows the ideal uniform input distribution after the activation, drawn on the symmetric range $[-1, 1]$ conventional for FSQ-style quantizers, which is affinely equivalent to the $[0, 1]$ formulation of Section 3.3. Boundary-based rounding, as in FSQ, turns this uniform input into an output distribution skewed toward the interval edges, whereas FSP decodes to interval centroids—precisely the Lloyd–Max optimal reconstruction values for a uniform scalar source—yielding a more uniform output distribution and lower expected distortion. Combining bounded centered perturbation with centroid decoding therefore improves on the rounding-to-grid heuristic of FSQ while retaining its computational simplicity. The choice of the uniformizing activation g, and the need to pair it with the pre-activation variance σ_g^2 that the activation itself implies, are studied in Appendix C.3.

Computational overhead. A micro-benchmark isolating the quantization or perturbation module (Appendix C.4) shows that the extra training cost of MH-based VP-VAE grows with G (from 6.42 to 79.83 ms per batch of $1{,}024$ latent tokens, against roughly 90.62 ms for one forward–backward pass of the image backbone), whereas FSP remains close to FSQ because its perturbation rule is fully factorized. This overhead is moreover confined entirely to training: at inference, MH-based VP-VAE reduces to a single nearest-centroid lookup and FSP to per-dimension scalar discretization, both on par with the lightest baselines.

Further qualitative reconstructions and a preliminary downstream generation experiment are reported in Appendix C.

5 Conclusion

We propose VP-VAE, a perturbation-first perspective on discrete representation learning that treats quantization primarily as a *structured latent perturbation* and, crucially, removes the need for an explicit codebook during training. Through two instantiations—a general MH-based prototype and the lightweight FSP—VP-VAE demonstrates that designing training-time perturbations to faithfully emulate inference-time quantization error is a principled and effective alternative to coupled codebook learning, offering a promising path toward stable, scalable discrete tokenizers for modern generative modeling.

Ethics Statement

This work studies discrete representation learning for images and audio. All experiments use publicly available research benchmarks—COCO (Lin et al., 2014), ImageNet (Deng et al., 2009), LibriSpeech (Panayotov et al., 2015), and Common Voice (Mozilla)—in accordance with their licenses and intended research use; we collected no new data and conducted no studies involving human subjects. The speech corpora contain recordings of human speakers, which we use solely to measure reconstruction quality; we perform no speaker identification and no other analysis of individuals.

Reproducibility Statement

The core implementation code is available at `https://anonymous.4open.science/r/VP-VAE-4DD5/`, covering MH-based VP-VAE (adaptive radius estimation, the FIFO queue, the MH accept–reject step, and post-training K-Means codebook construction) and FSP (the uniformizing activation, the bounded perturbation, and centroid decoding). The remaining experimental components are reused rather than newly designed: within each modality, all methods share the encoder–decoder backbone, optimizer, schedule, batch size, training budget, and reconstruction losses of the original backbones (Esser et al., 2021; Zhai et al., 2025), and each baseline follows the hyperparameters specified in its original paper. Datasets and preprocessing, level configurations, queue and kNN settings, the derivation of the radius scale η, and the regularization weights are documented in Section 4.1 and Appendix A; the theoretical claims underlying the design are proved in Appendix B. The complete experimental code will be released publicly upon acceptance.

AI Use Statement

We used AI assistants (large language models) in three of the ways covered by the ICLR policy, and disclose each of them here. First, *to aid or polish writing*: text drafted by the authors was revised with AI assistance for grammar, wording, and readability. Second, *for research ideation and execution*: the general MH-based instantiation of the perturbation-first paradigm was formulated by the authors and iteratively refined and clarified through discussions with AI assistants. Third, *for proving mathematical claims*: the derivations supporting that instantiation—the adaptive radius as a local quantization-scale estimator, the detailed balance of the one-step MH perturbation, and the closed-form acceptance ratio (Appendix B)—were likewise worked out in collaboration with AI assistants. The authors designed and ran all experiments, checked every derivation and reported result, and take full responsibility for the content of this paper.

References

Mohammad Adiban, Kalin Stefanov, Sabato Marco Siniscalchi, and Giampiero Salvi. S-HR-VQVAE: Sequential hierarchical residual learning vector quantized variational autoencoder for video prediction. *IEEE Transactions on Multimedia*, 27:4321–4332, 2025. ISSN 1941-0077. doi: 10.1109/TMM.2025.3535370.

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

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

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *Proceedings of the International Conference on Learning Representations (ICLR)*, February 2018.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. AudioLM: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2023. ISSN 2329-9304. doi: 10.1109/TASLP.2023.3288409.

Andreas Brendel, Nicola Pia, Kishan Gupta, Lyonel Behringer, Guillaume Fuchs, and Markus Multrus. Neural speech coding for real-time communications using constant bitrate scalar quantization. *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–15, 2024a. ISSN 1941-0484. doi: 10.1109/JSTSP.2024.3491575.

Andreas Brendel, Nicola Pia, Kishan Gupta, Lyonel Behringer, Guillaume Fuchs, and Markus Multrus. Neural speech coding for real-time communications using constant bitrate scalar quantization. *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–15, 2024b. ISSN 1941-0484. doi: 10.1109/JSTSP.2024.3491575.

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

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718, 2025. ISSN 2998-4173. doi: 10.1109/TASLPRO.2025.3530270.

Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu. Self-supervised learning with random-projection quantizer for speech recognition. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 3915–3924. PMLR, June 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Christopher Fifty, Ronald Guenther Junkins, Dennis Duan, Aniketh Iyengar, Jerry Weihong Liu, Ehsan Amid, Sebastian Thrun, and Christopher Re. Restructuring vector quantization with the rotation trick. In *Proceedings of the International Conference on Learning Representations (ICLR)*, October 2024.

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 *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, pp. 6629–6640, Red Hook, NY, USA, December 2017. Curran Associates, Inc.

Kyle Hsu, William Dorrell, James Whittington, Jiajun Wu, and Chelsea Finn. Disentanglement via latent quantization. *Advances in Neural Information Processing Systems*, 36:45463–45488, December 2023.

Minyoung Huh, Brian Cheung, Pulkit Agrawal, and Phillip Isola. Straightening out the straight-through estimator: Overcoming optimization challenges in vector quantized networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 14096–14113. PMLR, July 2023.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, number 353, pp. 3927–3936, Red Hook, NY, USA, December 2019. Curran Associates Inc.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. doi: 10.1007/978-3-319-10602-1_48.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *Proceedings of the International Conference on Learning Representations (ICLR)*, October 2023.

Mozilla. Mozilla common voice. https://commonvoice.mozilla.org/.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015.

Julian D. Parker, Anton Smirnov, Jordi Pons, C. J. Carr, Zack Zukowski, Zach Evans, and Xubo Liu. Scaling transformers for low-bitrate high-quality speech coding. In *Proceedings of the International Conference on Learning Representations (ICLR)*, October 2024.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8821–8831. PMLR, July 2021.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc., 2019.

A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2001.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010.

Yuhta Takida, Takashi Shibuya, Weihsiang 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 *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 20987–21012. PMLR, June 2022.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, NIPS'17, pp. 6309–6318, Red Hook, NY, USA, December 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4.

Yuqing Wang, Zhijie Lin, Yao Teng, Yuanzhi Zhu, Shuhuai Ren, Jiashi Feng, and Xihui Liu. Bridging continuous and discrete tokens for autoregressive visual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2024.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004. ISSN 1941-0042. doi: 10.1109/TIP.2003.819861.

Dongjie Ye, Baoliang Chen, Shiqi Wang, and Sam Kwong. CodedBGT: Code bank-guided transformer for low-light image enhancement. *IEEE Transactions on Multimedia*, 26:9880–9891, 2024. ISSN 1941-0077. doi: 10.1109/TMM.2024.3400668.

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 *Proceedings of the International Conference on Learning Representations (ICLR)*, October 2021.

Lijun Yu, José 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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An End-to-End Neural Audio Codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, November 2021. ISSN 2329-9290. doi: 10.1109/TASLP.2021.3129994.

Linwei Zhai, Han Ding, Cui Zhao, fei wang, Ge Wang, Wang Zhi, and Wei Xi. One quantizer is enough: Toward a lightweight audio codec, April 2025.

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

Yue Zhao, Yuanjun Xiong, and Philipp Kraehenbuehl. Image and video tokenization with binary spherical quantization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, October 2024.

Chuanxia Zheng and Andrea Vedaldi. Online clustered codebook. In *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22798–22807, 2023.

Zhaolin Zheng, Liqi Xue, Linghao Li, Chenwei Gong, Xiantong Zhen, and Jun Xu. Rethinking the role of vector quantization for blind image restoration. *IEEE Transactions on Multimedia*, pp. 1–12, 2026. ISSN 1941-0077. doi: 10.1109/TMM.2026.3688406.

Yongxin Zhu, Bocheng Li, Yifei Xin, Zhihua Xia, and Linli Xu. Addressing representation collapse in vector quantized models with one linear layer. In *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22968–22977, 2025.

APPENDIX

The appendix is organized into three parts. Appendix A gives the complete experimental setup that was condensed in the main text, together with the pseudocode of the MH-based perturbation operator. Appendix B provides the theoretical derivations underlying the design of VP-VAE, including the adaptive radius estimator, the stationarity of the Metropolis–Hastings (MH) perturbation, the closed-form acceptance ratio, and the optimality of centroid decoding in FSP. Appendix C reports the results deferred from the main text, including the full out-of-distribution benchmark, the component-wise ablation of the MH-based perturbation, the study of the uniformizing activation in FSP, a timing micro-benchmark of the quantization module, qualitative reconstructions, and downstream autoregressive generation results.

A DETAILED EXPERIMENTAL SETUP

This section provides the full experimental setup that was condensed in Section 4.1: datasets and preprocessing, the per-method configurations and training details, and the evaluation protocol and metrics.

A.1 DATASETS AND PREPROCESSING

For image tokenization, models are trained on the COCO 2017 training set (Lin et al., 2014) and evaluated in-domain on the COCO 2017 validation set. During preprocessing, all images are resized such that the shorter side is 256 pixels, followed by a center-crop to yield a fixed resolution of 256×256. For out-of-distribution (OOD) evaluation, the ImageNet validation set (Deng et al., 2009) is processed using the identical pipeline. For audio tokenization, models are trained on the LibriSpeech (Panayotov et al., 2015) 'train-clean-460' and 'train-other-500' splits and evaluated in-domain on the 'test-clean' and 'test-other' splits, with the Common Voice English test set (Mozilla) serving as the OOD benchmark. All audio signals are resampled to a 16 kHz sampling rate before being fed into the network.

A.2 PER-METHOD CONFIGURATIONS AND TRAINING DETAILS

Unless stated otherwise, the remaining hyperparameters of each baseline follow the settings specified in its original paper.

Backbones For images we use the VQGAN encoder–decoder (Esser et al., 2021) with a downsampling factor of 8; for audio we use the SQCodec encoder–decoder (Zhai et al., 2025).

Algorithm 1 MH-based Vector Perturbation

Require: Latent token $z \in \mathbb{R}^d$, queue $\mathcal{S}$, target codebook size G, scale η, kNN parameter k.
$M \leftarrow \lceil |\mathcal{S}|/G \rceil$
Compute $D_k(z), D_M(z)$ via kNN search on $\mathcal{S}$
$R(z) \leftarrow \eta \cdot D_M(z)$
Sample $u \sim \mathrm{Unif}(\mathcal{B}(0, R(z)))$; $\quad z' \leftarrow z + u$
Compute $D_k(z'), D_M(z')$ via kNN search on $\mathcal{S}$
$R(z') \leftarrow \eta \cdot D_M(z')$
if $\|z - z'\|_2 > R(z')$ **then**
    return $\tilde{z} \leftarrow z$
end if
ratio $\leftarrow \big(D_k(z) \cdot D_M(z)/(D_k(z') \cdot D_M(z'))\big)^d$
$\alpha \leftarrow \min(1, \text{ratio})$
Sample $b \sim \mathrm{Bernoulli}(\alpha)$
if $b = 1$ **then**
    return $\tilde{z} \leftarrow z'$
else
    return $\tilde{z} \leftarrow z$
end if

Fixed-Grid Level Configurations For FSQ and FSP, we use the standard level configurations recommended in Mentzer et al. (2023) to instantiate the four capacity regimes: $[8, 6, 5]$, $[8, 5, 5, 5]$, $[7, 5, 5, 5, 5]$, and $[8, 8, 8, 6, 5]$. The corresponding actual Cartesian vocabulary sizes are 240, 1000, 4375, and 15360, respectively. For MH-based VP-VAE, we match the bottleneck dimension d to the corresponding FSQ/FSP configuration at each target capacity. Our FSQ baseline is not vanilla FSQ but its stronger symmetric noise-injected variant (Parker et al., 2024; Brendel et al., 2024b), so that FSP is compared against the strongest published form of the fixed-grid quantizer.

FSP Configuration FSP reuses the level configurations listed above, and the perturbation scale is set to $\eta = 1.0$, so that the perturbation width coincides exactly with the quantization bin width. For the uniformizing activation we use the tanh-based $g(a) = (\tanh(a) + 1)/2$, which reproduces the latent range of FSQ, even though Appendix C.3 shows that a normal-CDF activation performs slightly better on the reconstruction metrics; this keeps the FSP–FSQ comparison focused on the perturbation and decoding rules alone. Following Brendel et al. (2024b), FSP is trained with a hybrid *perturb-or-quantize* protocol: with probability $1/2$ the latent is perturbed as in Eqs. (8)–(9), and otherwise quantized to interval centroids with the straight-through estimator as in Eq. (10).

TokenBridge Configuration Following the official implementation, we first train a KL-regularized VAE with a standard normal prior, after which the continuous latent space is discretized using Gaussian percentiles with 4 bins per dimension.

MH-based VP-VAE Configuration The FIFO queue size is set proportionally to the target discrete capacity as $|\mathcal{S}| = 32\,G\,d$, so that the number of queue samples per nominal quantization cell, $M = \lceil |\mathcal{S}|/G \rceil = 32\,d$, is independent of G and the distance estimates remain comparably reliable across capacities. The queue is initialized by drawing samples from a standard normal distribution $\mathcal{N}(0, \mathbf{I})$, which provides a well-conditioned starting geometry for the early-stage distance estimation before the encoder distribution stabilizes. At each step, we insert a random subset of the current minibatch tokens into $\mathcal{S}$, with the insertion probability set proportionally to $|\mathcal{S}|$ such that the entire queue is refreshed approximately one to two times per epoch, ensuring that $\mathcal{S}$ tracks the evolving latent distribution without excessive redundancy from strongly correlated within-sample tokens. The kNN density-estimation parameter is fixed to $k = 8$ throughout.

Per-Token Perturbation Procedure For completeness, Algorithm 1 summarizes the per-token perturbation operator described in Section 3.2.

Principled Choice of the Radius Scale η The radius scale η (Eq. (1)) is the only scale-related hyperparameter of MH-based VP-VAE, and we set it in a principled, codebook-size-dependent manner rather than by tuning. Concretely, for a standard normal reference latent, we compute (i) the expected magnitude of a uniform-ball perturbation of radius $D_M(z)$ and (ii) the expected quantization

error obtained by running K-Means with G centroids on samples from the same distribution; η is then chosen so that the former matches the latter, thereby aligning the training-time perturbation scale with the deployed quantization error (cf. the scale-alignment requirement in Section 3.1). This procedure yields $\eta \in \{1.126, 1.110, 1.093, 1.088\}$ for $G \in \{256, 1024, 4096, 16384\}$, respectively, where the mild decrease with G reflects that finer partitions make the local density increasingly uniform, driving the ratio toward its asymptotic value.

Weight for Latent Normalization Loss For the latent normalization regularizer (Eqs. (6) and (7)), we set the mean and variance regularization weights to $\lambda_1 = 1.0$ and $\lambda_2 = 0.07$, respectively.

Training Budget and Hardware All models are trained for 50 epochs using the optimizer and learning-rate schedules of the original backbones (Esser et al., 2021; Zhai et al., 2025), on 2 NVIDIA RTX 4090 GPUs.

A.3 Evaluation Protocol and Metrics

Within each modality, all methods share an identical encoder–decoder, optimizer, learning-rate schedule, batch size, training budget, and the reconstruction losses of the corresponding backbone, and differ *only* in the quantization or perturbation module, so that the discretization mechanism is the sole independent variable. We deliberately train every method without a discriminator or GAN loss: adversarial objectives can synthesize plausible detail that the tokenizer did not actually preserve, which would confound a comparison whose subject is the discretization mechanism itself. For the same reason, we evaluate with objective reconstruction and codebook-utilization metrics rather than human preference scores.

For images we report LPIPS (Zhang et al., 2018), PSNR, SSIM (Wang et al., 2004), reconstruction FID (rFID) (Heusel et al., 2017), and reconstruction KID (rKID) (Bińkowski et al., 2018); for audio we report wideband PESQ (Rix et al., 2001) and STOI (Taal et al., 2010). Codebook utilization is measured by the Codebook Valid Usage,

$$\mathrm{CVU} = \exp\Big(-\sum_{c \in \mathcal{Q}} p(c) \log p(c) \Big) \Big/ G,$$

where $p(c)$ is the empirical selection probability of code c. CVU is thus the perplexity of the code-usage distribution normalized by the target capacity G, so that a value of 1 indicates perfectly uniform utilization and smaller values increasing imbalance. Unlike the standard Codebook Usage, which counts only the fraction of codes selected at least once, CVU reflects how uniformly the discrete capacity is actually used: a codebook in which a handful of codes absorb almost all assignments can still report a high Codebook Usage, but its CVU will be low.

B Theoretical Derivations

Throughout, let $\mathcal{S} = \{s_i\}_{i=1}^{N}$ denote the FIFO queue with $N = |\mathcal{S}|$ latent samples drawn from the (unknown) latent density ρ on $\mathbb{R}^d$, let $V_d = \pi^{d/2}/\Gamma(d/2+1)$ be the volume of the unit ball in $\mathbb{R}^d$, and let $D_m(z)$ be the Euclidean distance from z to its m-th nearest neighbor in $\mathcal{S}$.

B.1 Adaptive Radius as a Local Quantization-Scale Estimator

We make precise the claim that, for a target codebook of size G, the radius $R(z) = \eta\, D_M(z)$ matches the local scale of the quantization error induced by an ideal equal-mass quantizer with G cells, where $M = \lceil N/G \rceil$.

Proposition B.1 (Local scale matching)**.** *Consider a quantizer that partitions $\mathbb{R}^d$ into G cells, each carrying probability mass $1/G$ under ρ. Assume ρ is approximately constant on the scale of a single cell (local smoothness), and let $r_{\mathrm{cell}}(z)$ denote the characteristic radius of the cell containing z, defined through its local volume $V_{\mathrm{cell}}(z) = V_d\, r_{\mathrm{cell}}(z)^d$. Then, in the large-sample limit $N \to \infty$ with $M = \lceil N/G \rceil$,*

$$D_M(z) \xrightarrow{\;p\;} r_{\mathrm{cell}}(z), \tag{11}$$

so that $R(z) = \eta\, D_M(z)$ estimates $r_{\mathrm{cell}}(z)$ up to the constant η.

Proof. By the equal-mass assumption, the cell containing z satisfies $\rho(z)\, V_{\text{cell}}(z) \approx 1/G$, hence

$$V_{\text{cell}}(z) \approx \frac{1}{G\,\rho(z)}, \qquad r_{\text{cell}}(z) \approx \left(\frac{1}{G\,\rho(z)\,V_d}\right)^{1/d}. \tag{12}$$

On the empirical side, a ball $\mathcal{B}(z, D_M(z))$ contains exactly M of the N queue samples by definition of the M-th nearest neighbor. Under local smoothness of ρ,

$$\frac{M}{N} \;\approx\; \rho(z)\, V_d\, D_M(z)^d \;\Longrightarrow\; D_M(z) \approx \left(\frac{M}{N\,\rho(z)\,V_d}\right)^{1/d}. \tag{13}$$

Substituting $M/N \approx 1/G$ (which holds since $M = \lceil N/G \rceil$) into (13) recovers exactly the right-hand side of (12), giving $D_M(z) \approx r_{\text{cell}}(z)$. The convergence in (11) is the usual consistency of nearest-neighbor distance statistics: as $N \to \infty$ the empirical mass of $\mathcal{B}(z, r)$ converges to its population mass, which pins $D_M(z)$ to the radius enclosing mass $M/N \to 1/G$. □

Remark 1 (Role of η and the choice of its value)**.** *The quantity controlling reconstruction is the* expected error magnitude*, not the cell radius itself. For a uniform proposal on $\mathcal{B}(0, R(z))$, the expected displacement is $\mathbb{E}\|u\|_2 = \frac{d}{d+1} R(z)$, whereas the expected quantization error within a cell depends on its shape and dimension d. The scalar η absorbs this geometry-dependent constant, which is precisely why we fix it by matching the expected perturbation magnitude to the K-Means quantization error measured on a standard normal reference (Appendix A.2) instead of tuning it. This alignment is verified empirically in Figure 1a.*

B.2 Stationarity of the MH Perturbation

We show that the perturbation operator of Eq. (5) defines a Markov kernel whose invariant distribution is the target density π, ensuring that accepted perturbations concentrate on high-density regions of the latent manifold.

Proposition B.2 (Detailed balance)**.** *Let $g(\cdot \mid z)$ be the uniform proposal density of Eq.* (3) *and let the acceptance probability be $\alpha(z, z') = \min\big(1,\, \pi(z')g(z \mid z')/[\pi(z)g(z' \mid z)]\big)$. Then the transition kernel*

$$\mathcal{K}(z, z') = g(z' \mid z)\, \alpha(z, z') + \delta_z(z') \Big[1 - \int g(y \mid z)\, \alpha(z, y)\, dy\Big] \tag{14}$$

satisfies detailed balance with respect to π, i.e. $\pi(z)\, \mathcal{K}(z, z') = \pi(z')\, \mathcal{K}(z', z)$ for all $z \neq z'$, and hence π is invariant.

Proof. For $z \neq z'$ the off-diagonal kernel is $g(z' \mid z)\, \alpha(z, z')$, so

$$\begin{aligned} \pi(z)\, g(z' \mid z)\, \alpha(z, z') &= \pi(z) g(z' \mid z) \min\Big(1, \tfrac{\pi(z')g(z|z')}{\pi(z)g(z'|z)}\Big) \\ &= \min\big(\pi(z)g(z' \mid z),\, \pi(z')g(z \mid z')\big). \end{aligned} \tag{15}$$

The right-hand side is symmetric under $z \leftrightarrow z'$, hence equals $\pi(z')g(z \mid z')\alpha(z', z)$, establishing detailed balance. Integrating over z and noting the Dirac part trivially preserves π gives invariance. □

Remark 2. *We do not iterate the chain to equilibrium; a single MH step per latent is applied at each training iteration. Proposition B.2 guarantees that this step is a π-reversible move, which is exactly the property we exploit: it biases each accepted perturbation toward the support of the empirical latent density and rejects moves into low-density or unsupported regions, satisfying the* distribution-consistency *criterion.*

B.3 kNN Density Estimation and the Acceptance Ratio

We justify the target density of Eq. (2) and derive the acceptance ratio of Eq. (4).

kNN density estimate The classical k-nearest-neighbor density estimator is $\hat{\rho}(z) = k/[N\, V_d\, D_k(z)^d]$, obtained by noting that the smallest ball centered at z containing k of the N samples has radius $D_k(z)$ and empirical mass k/N, so $\hat{\rho}(z)\, V_d\, D_k(z)^d \approx k/N$. Since k, N, and V_d are constants independent of z, this yields the unnormalized target $\pi(z) \propto 1/D_k(z)^d$, i.e. Eq. (2).

Acceptance ratio With the uniform proposal of Eq. (3), g is nonzero only within the respective balls, which gives the support condition stated in Section 3.2.2: $\alpha(z, z') = 0$ whenever $\|z - z'\|_2 > R(z')$. When both $\|z' - z\|_2 \le R(z)$ and $\|z - z'\|_2 \le R(z')$ hold, the proposal ratio is

$$\frac{g(z \mid z')}{g(z' \mid z)} = \frac{V_d R(z)^d}{V_d R(z')^d} = \left(\frac{R(z)}{R(z')}\right)^d = \left(\frac{D_M(z)}{D_M(z')}\right)^d, \tag{16}$$

using $R(z) = \eta D_M(z)$ and the cancellation of η. The density ratio is $\pi(z')/\pi(z) = \big(D_k(z)/D_k(z')\big)^d$. Multiplying the two gives

$$\frac{\pi(z')g(z \mid z')}{\pi(z)g(z' \mid z)} = \left(\frac{D_k(z)\, D_M(z)}{D_k(z')\, D_M(z')}\right)^d, \tag{17}$$

which, capped at 1, is exactly Eq. (4).

B.4 Optimality of Centroid Decoding for FSP

We justify the centroid decoder of FSP (Section 3.4, Eq. (10)). Consider a single latent dimension modeled, under the approximately-uniform prior, as $z \sim \mathcal{U}[0, 1]$, encoded with L reconstruction levels under squared-error distortion.

Proposition B.3 (Optimal scalar quantizer for a uniform source)**.** *Let $\{b_\ell\}_{\ell=0}^{L}$ with $0 = b_0 < \cdots < b_L = 1$ be decision boundaries and $\{c_\ell\}_{\ell=0}^{L-1}$ reconstruction levels, and define $\mathcal{D} = \int_0^1 (z - q(z))^2\, dz$ with $q(z) = c_\ell$ for $z \in [b_\ell, b_{\ell+1})$. Then $\mathcal{D}$ is minimized by $b_\ell = \ell/L$ and $c_\ell = (\ell + 1/2)/L$, with minimum distortion $\mathcal{D}^\star = 1/(12L^2)$.*

Proof. The Lloyd–Max necessary conditions are (i) the centroid condition $c_\ell = \mathbb{E}[z \mid z \in [b_\ell, b_{\ell+1})]$ and (ii) the nearest-neighbor condition $b_\ell = (c_{\ell-1} + c_\ell)/2$. For a uniform source the conditional mean is the interval midpoint, so (i) gives $c_\ell = (b_\ell + b_{\ell+1})/2$. Substituting into (ii) yields the recursion $b_\ell = (b_{\ell-1} + b_{\ell+1})/2$, i.e. equally spaced boundaries. With $b_0 = 0$, $b_L = 1$ this forces $b_\ell = \ell/L$ and $c_\ell = (\ell + 1/2)/L$. As the uniform density is log-concave, these stationary conditions are also sufficient for global optimality. The distortion is then

$$\mathcal{D}^\star = \sum_{\ell=0}^{L-1} \int_{\ell/L}^{(\ell+1)/L} \Big(z - \tfrac{\ell+1/2}{L}\Big)^2 dz = L \cdot \frac{(1/L)^3}{12} = \frac{1}{12L^2}. \tag{18}$$

□

Corollary B.4 (Uniform output distribution)**.** *Under $z \sim \mathcal{U}[0, 1]$, the FSP quantizer of Eq.* (10) *maps z to each centroid $c_\ell = (\ell + 1/2)/L$ with probability $\Pr[z \in [\ell/L, (\ell + 1)/L)] = 1/L$. The induced distribution over the L codes is therefore exactly uniform, attaining the maximum per-dimension entropy $\log L$ (equivalently $\mathrm{CVU} = 1$ per dimension).*

Remark 3 (Contrast with FSQ)**.** *FSQ places its L levels on an equally spaced grid that includes both ends of the range—after the affine rescaling to $[0, 1]$, the levels $\{\ell/(L-1)\}_{\ell=0}^{L-1}$—and rounds to the nearest level, so the two extreme levels own half-width decision cells while all interior levels own full-width cells. Under a uniform source the extreme codes therefore receive only half the probability mass of an interior code, which makes the code distribution non-uniform and the expected distortion strictly larger than that of the centroid quantizer of Proposition B.3. This is the mechanism visualized in Figure 1c and underlies the consistent reconstruction improvement of FSP over FSQ reported in Section 4.2.*

C Additional Experiments

C.1 Out-of-Distribution Benchmark Results

Table 2 reports the complete out-of-distribution comparison analyzed in Section 4.2: every tokenizer is trained on COCO (image) or LibriSpeech (audio) and then evaluated without any adaptation on the ImageNet validation set and the Common Voice English test set, respectively. The evaluation protocol, metrics, and per-method configurations are identical to those of the in-domain benchmark.

Table 2: Out-of-distribution generalization results. Models trained on COCO/LibriSpeech are evaluated on unseen datasets: ImageNet (image) and Common Voice (audio). rKID is reported in units of $\times 10^3$. "–" indicates training failure due to severe codebook collapse.

Method	Codebook Size	Image (ImageNet)						Audio (Common Voice)		
		CVU	LPIPS ↓	PSNR ↑	SSIM ↑	rFID ↓	rKID ↓	CVU	PESQ ↑	STOI ↑
VQ-VAE	256	0.3617	0.2121	23.5942	0.6704	1.0034	4.4967	-	-	-
SimVQ	256	0.9363	**0.1927**	23.6599	0.6827	**0.7314**	**2.2244**	0.0041	1.2103	0.3781
TokenBridge	256	0.9887	0.6079	11.6525	0.2024	29.473	276.72	0.9887	1.0967	0.2688
FSQ	240	0.8574	0.2182	23.4555	0.6645	1.0543	5.8357	0.1690	1.2064	0.7630
FSP	240	0.8821	0.1968	**24.0268**	**0.6863**	0.8204	2.9822	0.2860	1.8423	0.8528
VP-VAE	256	0.8431	0.1972	23.8125	0.6855	0.8164	2.8182	0.2055	**1.8601**	**0.8645**
VQ-VAE	1024	0.3604	0.1911	23.8520	0.6851	0.7115	2.2676	0.0010	1.2103	0.3371
SimVQ	1024	0.8736	**0.1719**	**24.3343**	0.7070	**0.6025**	1.7755	0.0010	1.1568	0.3627
TokenBridge	1024	0.9887	0.6108	11.4561	0.1985	34.199	318.71	0.9887	1.1059	0.2856
FSQ	1000	0.8000	0.1931	24.0973	0.6878	0.7949	3.4781	0.1012	1.7776	0.8498
FSP	1000	0.8488	0.1799	24.2206	0.7023	0.6682	2.015	0.1776	**1.9529**	**0.8715**
VP-VAE	1024	0.8376	0.1755	24.2487	**0.7182**	0.6754	**1.7439**	0.1441	1.9337	0.8683
VQ-VAE	4096	0.2371	0.1770	24.3765	0.7113	0.6462	1.9483	-	-	-
SimVQ	4096	0.8504	**0.1578**	24.8154	0.7339	**0.5325**	1.5875	0.0002	1.1440	0.3713
TokenBridge	4096	0.9887	0.6071	11.5554	0.1920	113.96	374.57	0.9888	1.1017	0.2787
FSQ	4375	0.7759	0.1749	24.4162	0.7065	0.6719	2.6462	0.0670	1.9712	0.8707
FSP	4375	0.8159	0.1580	**24.9989**	0.7323	0.5405	1.2128	0.0936	**2.0686**	**0.8802**
VP-VAE	4096	0.8382	0.1640	24.8861	**0.7400**	0.5663	**1.0074**	0.1151	2.0335	0.8779
VQ-VAE	16384	0.2621	0.1592	24.8349	0.7359	0.5304	1.2741	-	-	-
SimVQ	16384	0.8164	**0.1458**	25.3304	0.7496	**0.4606**	1.1636	0.0001	1.1807	0.4002
TokenBridge	16384	0.9888	0.6166	11.4447	0.1969	65.669	259.07	0.9887	1.1202	0.2722
FSQ	15360	0.7756	0.1627	24.8198	0.7237	0.6092	2.1768	0.0499	1.9566	0.8721
FSP	15360	0.7258	0.1561	25.2168	0.7396	0.5146	1.2216	0.0541	2.1029	0.8884
VP-VAE	16384	0.8604	0.1524	**25.4548**	**0.7521**	0.5073	**1.1215**	0.1218	**2.1565**	**0.8990**

Table 3: Ablation study on the MH-based VP-VAE perturbation design. We evaluate the contribution of various perturbation designs and regularization on COCO image reconstruction (G=4096). Both design requirements—scale alignment and distribution consistency—as well as the latent normalization regularizer contribute to reconstruction quality and codebook balance.

Setting	CVU	LPIPS ↓	PSNR ↑	SSIM ↑
VP-VAE (full)	0.8493	0.1540	24.6418	0.7481
no perturbations	0.8768	0.2579	22.2179	0.6831
fixed Gaussian perturbations	0.8755	0.1548	24.5875	0.7387
w/o MH (always accept)	0.8031	0.1540	24.5910	0.7407
w/o $\mathcal{L}_{\text{norm}}$	0.8772	0.1672	24.3944	0.7371

C.2 Component-wise Ablation of the MH-based Perturbation

Table 3 quantifies the contribution of each design choice of MH-based VP-VAE on COCO image reconstruction (G=4096); Section 4.3 summarizes the outcome. Removing perturbations entirely ('no perturbations') yields a continuous decoder that is unprepared for discrete quantization errors: although post-training K-Means still partitions the latent space into G balanced cells (yielding a seemingly high CVU of 0.8768), the train–inference mismatch degrades reconstruction sharply (PSNR drops to 22.2179, LPIPS jumps to 0.2579). As with TokenBridge in Section 4.2, high utilization alone is not informative once the decoder cannot robustly decode the quantized latents. Replacing adaptive perturbations with fixed Gaussian perturbations (using a tuned, globally constant variance) yields a slightly inflated CVU (0.8755) but degrades reconstruction fidelity (PSNR drops from 24.6418 to 24.5875, SSIM from 0.7481 to 0.7387); this occurs because isotropic noise lacks scale alignment, applying excessive noise in dense regions (forcing the model to waste capacity on artificial variance) and insufficient noise in sparse regions. Disabling the MH step ('w/o MH', i.e., always accepting proposals) instead breaks distribution consistency: perturbed latents drift into low-density regions, so the post-training codebook centroids are poorly matched to the actual data manifold and the effective codebook utilization (CVU) drops from 0.8493 to 0.8031. Finally, the latent normalization regularizer ($\mathcal{L}_{\text{norm}}$) is crucial for the stability of the Euclidean distances used in our kNN scale estimation: without it, individual dimensions with naturally larger variances dominate the distance computation and heavily skew the radius estimation, which degrades all reconstruction metrics, particularly LPIPS (worsening to 0.1672) and PSNR (dropping to 24.3944).

Table 4: Effect of the uniformizing activation in FSP. Image reconstruction on COCO (G=4096) under different CDF-like activations g and their corresponding target pre-activation variances σ_g^2. The last row intentionally mismatches the variance constraint to the activation.

Activation	Target σ_g^2	CVU	LPIPS ↓	PSNR ↑	SSIM ↑
tanh-based	0.82	0.8341	0.1506	24.8914	0.7473
normal-CDF	1.0	0.8288	0.1481	24.9337	0.7504
Laplace-CDF	2.0	0.7952	0.1532	24.8283	0.7441
sigmoid	3.29	0.8015	0.1533	24.5900	0.7423
sigmoid	1.0	0.4960	0.1538	24.5947	0.7402

Table 5: Absolute computational time (ms) of the quantization/perturbation module. We report the average wall-clock time required to process a batch of 1,024 latent tokens ($1024 \times d$) during training and inference.

Method	Training Time (ms/batch)				Inference Time (ms/batch)			
	Target Codebook Size (G)				Target Codebook Size (G)			
	256	1024	4096	16384	256	1024	4096	16384
VQ-VAE	2.11	2.08	1.90	2.29	0.90	0.89	0.90	1.83
SimVQ	1.04	1.27	0.98	0.97	1.06	2.04	0.95	0.92
TokenBridge	0.63	0.63	0.62	0.60	0.74	0.71	0.73	0.69
FSQ	0.97	1.84	0.82	0.89	0.71	0.82	0.67	0.68
FSP	1.46	1.36	1.45	1.48	0.55	0.50	0.54	0.53
VP-VAE	6.42	10.62	40.12	79.83	0.57	0.54	0.55	0.67

C.3 Choice of Uniformizing Activation in FSP

FSP maps pre-activations through a smooth, monotone, CDF-like activation g to obtain bounded latents in $[0, 1]^d$ (Section 3.3). Table 4 compares several natural choices for g: a tanh-based activation $g(a) = (\tanh(a) + 1)/2$, the normal CDF, the Laplace CDF, and the sigmoid. These activations differ in their tail shape and gradient profile, and we observe a mild trade-off among them. For instance, the normal-CDF achieves the best reconstruction metrics (PSNR 24.9337, LPIPS 0.1481), while the tanh-based activation yields a slightly higher codebook utilization (CVU 0.8341). In the main experiments we nevertheless adopt the tanh-based activation, so that FSP matches the latent range of FSQ and the comparison between the two isolates the perturbation and decoding rules alone.

The specific activation must be paired with its correct theoretical pre-activation variance σ_g^2 (Eq. (7)). The last row of Table 4 tests this geometric assumption directly. When we intentionally impose a mismatched target (σ_g^2=1.0 on a sigmoid activation, whose correct value is ≈ 3.29), the reconstruction fidelity remains largely intact, but the CVU falls from 0.8015 to 0.4960. This is consistent with our uniform-prior premise: FSP decodes via equal-width bins (Eq. (10)); if the variance mismatch prevents the pre-activations from forming an approximately uniform distribution, equal-width discretization naturally results in imbalanced bin assignments.

C.4 Computational Overhead

To provide a transparent and objective assessment of the cost summarized in Section 4.4, we conduct a micro-benchmark isolating the absolute time spent inside the quantization or perturbation module. As shown in Table 5, we report the wall-clock time (in milliseconds) required to process a typical batch of $1{,}024$ latent tokens across different target codebook sizes (G). For context, the forward–backward pass of the image encoder–decoder backbone alone requires approximately 90.62 ms under the same.

From the algorithmic structure, the main extra training cost of VP-VAE comes from repeated kNN search and MH accept–reject decisions, and therefore grows with G; FSP instead stays close to FSQ, because its perturbation rule is fully factorized and requires neither density estimation nor iterative sampling. The general instantiation thus pays for its generality, whereas FSP shows that the perturbation-first paradigm itself need not be computationally heavy once a stronger latent prior is available. We emphasize that the current MH-based timings reflect only a straightforward CUDA-compiled implementation of the kNN search; this cost is an engineering artifact rather than a

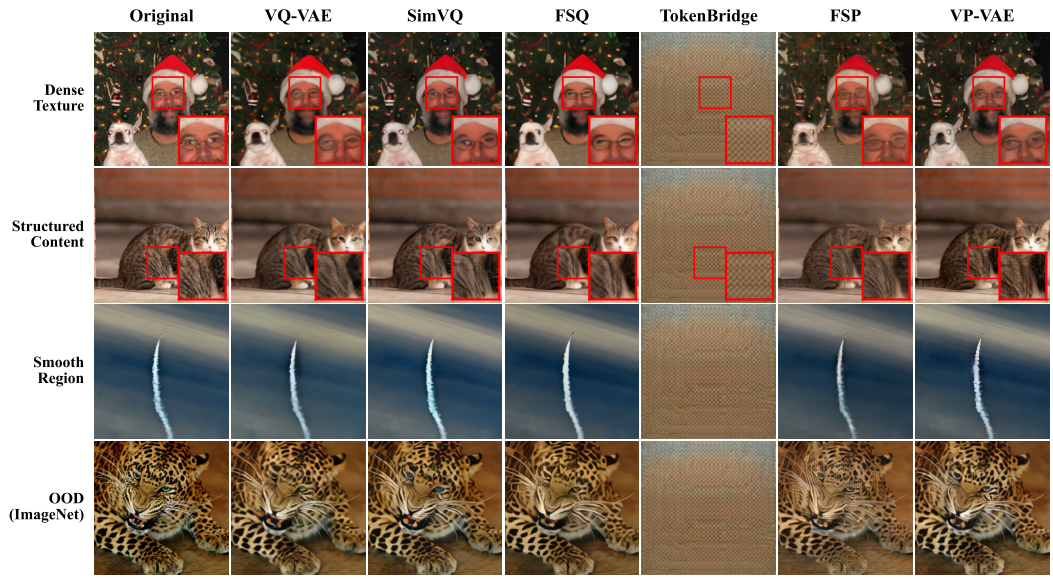


Figure 2: **Qualitative reconstruction comparison** (G=4096)**.** Columns correspond to methods and rows to representative samples, spanning dense texture, structured content, smooth regions, and an OOD ImageNet sample (bottom row). Red boxes mark the zoomed-in crops shown beneath each image. VP-VAE and FSP preserve detail and structural coherence.

fundamental limitation, and could be substantially reduced via approximate nearest-neighbor methods (e.g., hashing- or graph-based ANN) and asynchronous queue maintenance, which we leave to future work. The overhead is confined entirely to training: at inference, MH-based VP-VAE reduces to a single nearest-centroid lookup (0.54–0.67 ms) and FSP to per-dimension scalar discretization (0.50–0.55 ms), both on par with the lightest baselines.

C.5 Qualitative Reconstruction

Figure 2 presents qualitative reconstructions at G=4096 across all methods. We select representative samples that span varying image characteristics: dense high-frequency textures (human face with glasses), structured spatial content (animal fur), smooth gradient regions (sky), and an out-of-distribution (OOD) sample from the ImageNet validation set (leopard). The visual results closely align with our quantitative findings:

Baseline methods. Consistent with the severe degradation in quantitative metrics (e.g., rKID $>$ 290 on COCO), TokenBridge fails to reconstruct meaningful content, producing only a uniform, noisy artifact. This supports our reading in Section 4.2: forcing a pre-trained continuous latent space onto a rigid, predefined Gaussian grid disrupts the decoder, rendering its near-perfect codebook utilization uninformative. Among the coupled quantization baselines, VQ-VAE exhibits severe blurring and structural distortion; it notably fails to reconstruct the thin frames of the glasses (first row, zoomed-in crop) and heavily smudges the intricate patterns of the cat's fur (second row). SimVQ partly mitigates this but still struggles with localized high-frequency details, showing noticeable deformations in the glasses and over-sharpening elsewhere. FSQ performs reasonably well thanks to the stability of its fixed grid, but still yields slightly over-smoothed textures compared to the original images.

Our methods (FSP and VP-VAE). In contrast, both instantiations of our perturbation-first paradigm consistently produce the highest fidelity reconstructions across all scenarios. As highlighted by the red zoomed-in boxes, FSP and VP-VAE accurately preserve fine geometries (such as the distinct rims of the glasses) and crisp textural details (such as the individual strands of animal fur).

C.6 Downstream Autoregressive Generation

Discrete tokenizers are most often used as the vocabulary for a downstream generative model. To check that the tokens produced by our perturbation-first paradigm are readily learnable and suitable for such tasks, we conduct a downstream image generation experiment.

Table 6: **Downstream generation on COCO (G=4096).** All tokenizers are evaluated using an identical 18-layer autoregressive Transformer. The metrics reveal a trade-off: methods with slightly more concentrated token distributions (e.g., VQ-VAE) yield lower NLL and higher Precision, while methods with more uniformly utilized codebooks (e.g., FSP, VP-VAE) achieve higher Recall and competitive overall fidelity (FID/KID).

Method	FID↓	KID↓	Precision↑	Recall↑	NLL↓
VQ-VAE	17.514	25.2114	**0.9139**	0.5697	**8.9396**
SimVQ	**13.2357**	24.6292	<u>0.8942</u>	<u>0.7484</u>	<u>9.0650</u>
FSQ	14.5561	30.8593	0.8848	0.7011	9.4724
FSP	<u>13.7160</u>	<u>20.9637</u>	0.8716	**0.7542**	9.5423
VP-VAE	13.9034	**20.6388**	0.8802	0.7404	9.3708

Experimental setup. We freeze the trained tokenizers at a target capacity of $G = 4096$ and extract discrete token sequences for the COCO dataset. We then train an identical 18-layer, purely autoregressive (AR) Transformer on these sequences. TokenBridge is left out of this comparison, since its reconstructions are already too degraded (Section 4.2) for downstream generation to be meaningful. To evaluate the generation quality, we report Fréchet Inception Distance (FID) and Kernel Inception Distance (KID). To disentangle fidelity from diversity, we use the Precision and Recall metrics (Kynkäänniemi et al., 2019). Additionally, we report the validation Negative Log-Likelihood (NLL), which reflects the predictability of the discrete token sequences. All metrics are computed over $5{,}000$ generated samples under identical sampling hyperparameters (temperature $= 1.0$, top-$k = 16$).

Results and analysis. As shown in Table 6, our decoupled training paradigm is competitive with well-tuned coupled quantization methods in downstream generation, with each approach exhibiting different trade-offs.

VQ-VAE achieves the lowest validation NLL (8.9396) and the highest Precision (0.9139). This suggests that its generated sequences are highly predictable and that the generated images safely reside within high-fidelity regions. However, this comes at the cost of generation diversity, as indicated by its relatively low Recall (0.5697). The trade-off is intuitive: a slightly more concentrated token usage simplifies the next-token prediction task for the Transformer, but provides a less diverse vocabulary for covering the full underlying data distribution.

Conversely, our proposed methods (FSP and VP-VAE) operate at the other end of this trade-off. By maintaining a far better utilized codebook than VQ-VAE (CVU 0.83–0.85 versus 0.24 at this capacity), our methods provide a structurally richer vocabulary. The higher entropy of the token distribution makes the AR modeling inherently slightly more challenging (resulting in moderately higher NLLs of 9.5423 and 9.3708). However, this balanced utilization pays off in generation diversity and fine-grained distribution matching: FSP achieves the best Recall (0.7542), while VP-VAE attains the best KID (20.6388).

SimVQ also performs strongly and in a balanced way, achieving the best FID (13.2357) and the second-best scores on several metrics, as expected from a carefully tuned coupled quantizer. Yet FSP and VP-VAE reach comparable downstream quality *without* coupling representation learning to a jointly-optimized codebook: VP-VAE attains the best KID (20.6388) and FSP the best Recall (0.7542), both within a narrow margin of SimVQ on FID. The trade-off they occupy is also more desirable for generation: by keeping codebook utilization balanced and high-entropy, they trade a small amount of next-token predictability (slightly higher NLL) for markedly better diversity (Recall) and distributional fidelity (KID)—the properties governing how faithfully a generator covers the data distribution.

In summary, the perturbation-first paradigm delivers downstream generation quality on par with well-tuned coupled quantizers, without their training-time instability: SimVQ generates competitive images but collapses on audio (cf. Table 1), whereas FSP and VP-VAE retain a usable vocabulary on both modalities. This suggests that decoupling discretization from representation learning buys diversity and robustness without costing generation quality.