Tokenizer–Generator Coupling in Medical Image Generation

Liam Chalcroft
University College London
London, United Kingdom
liam.chalcroft.20@ucl.ac.uk

Abstract

Latent generative models for medical images usually treat the tokenizer as a fixed preprocessing step, with evaluation focused on the generator. This separation was tested in a controlled ChestMNIST study at 64×64 that crosses discrete tokenizers, generator families and sampler settings on a shared latent grid, alongside continuous-latent reference cells. Rankings were found to depend jointly on all three components, with the best quantizer changing with the generator, and the interaction held when the vocabulary-1024 block was retrained at three seeds, where 4 of 9 pairwise quantizer comparisons exceed three seed standard deviations, including a reversal between LFQ and FSQ from MaskGIT to D3PM. Reconstruction quality was not a reliable guide to generation quality. Sampler settings also changed the conclusions, since validation-selected retuning of D3PM and SEDD on LFQ-1024 reduced FID-192 from 0.44 and 0.41 to 0.09 and 0.10 at lower NFE, replicated across seeds, bringing both close to the strongest continuous references. FID-192 is used as an internal ranking metric and agrees with standard FID-2048 (Spearman 0.89) and with a label-free two-sample test (0.86). The results fit a rate-distortion-modelability tradeoff in which modelability depends on the generator, sampler and inference budget. The scope is limited to 54 discrete and 16 continuous cells on low-resolution medical-style images, with unconditional generation, non-clinical FID-based evaluation and single-seed training outside the seeded block.

1 Introduction

Most latent generative pipelines for medical images compress the image with a VQ-GAN [Esser et al., 2021] or KL-VAE, train a diffusion model on the latents and evaluate only the generator. The tokenizer nonetheless defines the sequence, grid or continuous field that the generator has to model, so two tokenizers with similar reconstructions may induce quite different latent distributions, and a generator that appears weak on one tokenizer may be strong on another.

The natural-image literature has moved away from judging tokenizers by reconstruction alone, since better reconstruction does not necessarily give better generation. Dieleman [2025] describes a rate-distortion-modelability tradeoff, where a representation must preserve information whilst remaining easy for the downstream model to learn, building on earlier rate-distortion-usefulness and rate-distortion-perception views [Tschannen et al., 2018, Blau and Michaeli, 2019]. The question is more pressing for medical images, where datasets are smaller, clinically relevant variation can be subtle and most published pipelines occupy a narrow part of the design space [Pinaya et al., 2023, Bluethgen et al., 2025, Khader et al., 2023, Yellapragada et al., 2024]. It remains unclear whether recent tokenizer and generator alternatives help, whether they interact, or whether reconstruction metrics are sufficient to choose between them.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

This work studies the tokenizer–generator–sampler triple directly in a controlled setting, using ChestMNIST at 64×64, a shared encoder-decoder tokenizer backbone, a matched 8×8 token grid, matched vocabulary sizes for the discrete tokenizers and a common evaluation protocol. Such a setting cannot settle questions about clinical image generation, but it can expose interactions that are normally confounded by differences in architecture, resolution, dataset and sampling.

Contributions. (i) A matched factorial design separates quantizer family, categorical rate, generator family and sampler budget, which are usually fixed together within a single pipeline; the ChestMNIST grid contains 54 discrete cells, crossing VQ, LFQ and FSQ at vocabularies of 1K, 2K and 4K with six generators, and 16 continuous VAE/AE reference cells with LDM and rectified flow. (ii) Evidence that quantizer and generator effects are not separable: in the vocabulary-1024 block retrained at three seeds, the best quantizer changes with the generator, and validation-based sampler selection changes the apparent generator ranking. (iii) Evidence that reconstruction quality is insufficient for tokenizer selection in both the discrete and the continuous panels. Supporting analyses include FID bootstrap intervals, a domain-feature FID check, replication of six cells on PneumoniaMNIST and OrganAMNIST, and feature-space and raw-space nearest-neighbour memorization probes.

Rate, distortion and modelability. For a tokenizer T, rate $R(T)$ is the latent capacity, which is $64 \log_2 K$ raw bits for a discrete 8×8 grid and the latent channel count for the continuous references, and distortion $D(T)$ is reconstruction error, measured by PSNR, SSIM and LPIPS [Zhang et al., 2018]. Modelability is treated as conditional on the generator, sampler and budget, with $M(T, G, S, B)$ denoting the quality of decoded samples after training generator G on tokenizer T and sampling with setting S under inference budget B, and FID-192 is used as the primary proxy for M.

2 Related work

Medical latent generation. Latent diffusion and VQ-style compression are now widely used for chest X-ray, MRI and histopathology generation [Bluethgen et al., 2025, Khader et al., 2023, Yellapragada et al., 2024]. The available tooling reflects a narrow set of choices, with MONAI Generative Models providing VQ-VAE and AutoencoderKL [Pinaya et al., 2023], MedVAE focusing on continuous autoencoder latents [Varma et al., 2025], and recent medical image tokenizers generally specialising in a single quantization family. This is a reasonable engineering default, but it leaves most medical studies unable to tell whether a result reflects a strong generator, a tokenizer that produced an easy distribution, or a sampler that happened to be well tuned for that representation.

Tokenizers for natural images. MAGVIT-v2 showed that replacing VQ with LFQ can improve generation with the generator held fixed [Yu et al., 2024], and FSQ avoids learned-codebook collapse by quantizing scalar levels independently [Mentzer et al., 2024]. Tokenizer benchmarks and bit-wise or factorised tokenizer designs extend this picture [Lin et al., 2025, Wu et al., 2025, Weber et al., 2024, Bai et al., 2024], and several works report that tokenizers optimised for reconstruction can be worse substrates for generation [Yao et al., 2025, Xiong et al., 2025, Qiu et al., 2025]. These studies motivate the present question but do not address medical images, and they typically hold either the generator or the sampler fixed. The grid here instead pairs each tokenizer with autoregressive, masked, diffusion, flow and Bayesian samplers and evaluates each generator across matched quantizer families and vocabulary sizes.

3 A controlled tokenizer–generator grid

Tokenizers. All tokenizers map a 64×64 greyscale image to an 8×8 latent grid using the same convolutional encoder-decoder backbone. The discrete panel uses VQ [van den Oord et al., 2017], lookup-free quantization (LFQ) [Yu et al., 2024] and finite scalar quantization (FSQ) [Mentzer et al., 2024] at matched vocabulary sizes. The continuous panel uses VAE tokenizers [Kingma and Welling, 2014, Rombach et al., 2022], sweeping latent channels $c \in \{1, 2, 4, 8, 16\}$ at a fixed KL weight $\lambda = 10^{-6}$ and sweeping the KL weight at $c = 4$, where the $\lambda = 0$ endpoint is a plain autoencoder (AE) trained with the same reconstruction objective. VQ codes are normalised identically at training and decoding.

Generators. Every discrete tokenizer is paired with six generators: an autoregressive (AR) transformer, MaskGIT [Chang et al., 2022], discrete flow matching (DFM) with the categorical probability path of Gat et al. [2024] (Campbell et al. [2024] describe a closely related model), absorbing-state D3PM [Austin et al., 2021], SEDD [Lou et al., 2024] trained with the score-entropy objective and sampled with the D3PM reverse step in place of tau-leaping, and a Bayesian flow network (BFN) [Graves et al., 2023]. The continuous panel uses latent diffusion (LDM) [Rombach et al., 2022] and rectified flow (RF) [Liu et al., 2023] with a matched DiT-style backbone [Peebles and Xie, 2023].

Training controls. All generators use a DiT-style backbone with RoPE [Su et al., 2024] and SiLU activations. AR and MaskGIT have no timestep modulation and 38.9M parameters at vocabulary 1024, the diffusion-style discrete generators use adaLN-Zero timestep conditioning with 57–58M parameters, and the continuous LDM and RF models use a matched 57.1M-parameter DiT. Discrete generators are trained for 100 epochs and continuous generators for 200, with AdamW, cosine decay, 5% warmup, EMA, gradient clipping at 1.0, BFloat16 mixed precision and a batch size of 128. These settings are not tuned per cell and are intended to make the comparison interpretable under a common implementation budget.

Metrics and seeds. The primary metric is FID-192, a low-dimensional variant of FID [Heusel et al., 2017] computed with `torchmetrics FrechetInceptionDistance(feature=192, normalize=True)` [Detlefsen et al., 2022], which uses globally averaged features from the second max-pooling block of InceptionV3 in place of the standard 2048-dimensional features. It is used as an internal ranking metric for convenience at 64×64, and its rankings agree with standard FID-2048 (Spearman 0.89 over 12 cells), with a label-free classifier two-sample test (0.86) and, more weakly, with a domain-feature FID on six cells ($\rho = 0.71$; Appendix E). Each default cell is evaluated with 10K generated samples. The vocabulary-1024 interaction block is trained at three seeds (42, 43 and 44; median per-cell standard deviation 0.047); the rest of the grid is single-seed, so small differences there are not ranked. The controls are summarised in Table 8 (Appendix A).

4 Results

4.1 Tokenizer and generator rankings are not separable

Table 1 reports generation quality for every cell, with vocabulary size held fixed so that a comparison between quantizers does not also change the categorical rate. Across the single-seed grid LFQ has the lowest FID in 15 of 18 matched-vocabulary comparisons, although several of these gaps are small, and the firmer evidence comes from the vocabulary-1024 block retrained at three seeds (Table 2 and Figure 1). Four of its nine pairwise quantizer comparisons exceed three pooled seed standard deviations. Under D3PM LFQ beats both FSQ and VQ, whereas under MaskGIT FSQ beats both LFQ (by 0.77) and VQ, so the best quantizer depends on the generator. The other five comparisons fall within seed noise and are left unranked: all three under AR, where the quantizers are level, VQ against LFQ under MaskGIT, and VQ against FSQ under D3PM. Standard FID-2048 on the same samples (Table 6) preserves the D3PM ordering and also shows the best quantizer changing with the generator, but it prefers LFQ under MaskGIT and FSQ under AR. The interaction and the D3PM ordering are therefore supported by both metrics, and the AR and MaskGIT winners are not.

A fixed-effect corrected sum-of-squares decomposition of $\log$(FID-192) over the balanced discrete panel (Table 3) attributes 59.5% of the observed variation to generator family, 10.1% to quantizer family and 23.4% to the quantizer–generator and vocabulary–generator interactions combined. These shares are descriptive and come from the single-seed grid; the three-seed block supports the existence of the interaction but not its size.

VQ depends strongly on the generator. Pilot VQ runs with 8K codes performed poorly and are not reported; in the matched sweep VQ with AR reaches 0.52, 0.52 and 0.31 at 1K, 2K and 4K. VQ-4096 + AR is the lowest single default discrete cell, although its margin over LFQ-1024 + AR (0.33) and FSQ-4096 + AR (0.37) is within the seed-to-seed spread of the AR cells in Table 2, so it is not ranked above them. Under the other five generators VQ trails LFQ at every vocabulary, except under MaskGIT at 1K and 2K, where the two are level. At these rates VQ uses all of its codebook with high normalised entropy, so codebook collapse does not explain the LFQ advantage. The choice

Table 1: Generation quality across tokenizer–generator cells (FID-192 ↓, 10K samples). Bold marks the best default generator per tokenizer row. The discrete block is the matched VQ/LFQ/FSQ × 1K/2K/4K grid; the continuous block gives VAE/AE reference cells, where c is the latent channel count and λ the KL weight.

		Discrete generators						Continuous	
Tokenizer	Setting	AR	MaskGIT	DFM	D3PM	SEDD	BFN	LDM	RF
VQ	1K	**0.52**	1.86	4.42	1.30	1.31	5.98	–	–
VQ	2K	**0.52**	3.33	2.75	2.12	2.13	3.94	–	–
VQ	4K	**0.31**	2.03	3.39	3.69	3.54	4.22	–	–
LFQ	1K	**0.33**	1.91	1.77	0.44	0.41	2.27	–	–
LFQ	2K	**0.45**	3.40	1.33	0.74	1.02	1.46	–	–
LFQ	4K	**0.41**	1.27	2.21	1.92	2.08	2.41	–	–
FSQ	1K	**0.39**	1.15	3.02	1.21	1.05	7.60	–	–
FSQ	2K	**0.73**	1.48	1.86	2.50	2.46	4.83	–	–
FSQ	4K	**0.37**	1.46	2.69	7.40	7.34	8.46	–	–
VAE	$c = 1, \lambda = 10^{-6}$	–	–	–	–	–	–	**0.22**	0.26
VAE	$c = 2, \lambda = 10^{-6}$	–	–	–	–	–	–	**0.14**	0.18
VAE	$c = 4, \lambda = 10^{-6}$	–	–	–	–	–	–	0.41	**0.13**
VAE	$c = 8, \lambda = 10^{-6}$	–	–	–	–	–	–	1.61	**0.32**
VAE	$c = 16, \lambda = 10^{-6}$	–	–	–	–	–	–	0.66	**0.11**
VAE	$c = 4, \lambda = 10^{-5}$	–	–	–	–	–	–	0.20	**0.14**
VAE	$c = 4, \lambda = 10^{-7}$	–	–	–	–	–	–	**0.09**	0.10
AE	$c = 4, \lambda = 0$	–	–	–	–	–	–	0.09	**0.07**

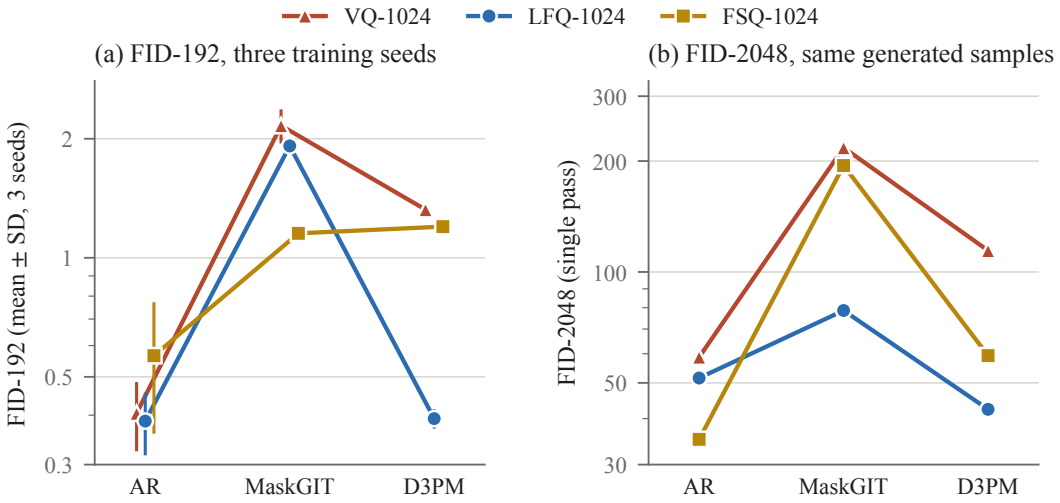


Figure 1: The vocabulary-1024 block as an interaction plot. (a) FID-192, mean and standard deviation over three training seeds; points are offset horizontally for legibility. (b) Standard FID-2048 computed on the same generated samples, one generation per cell. Both axes are logarithmic and lower is better. If quantizer and generator effects were separable the three lines would be parallel.

between learned and lookup-free quantization therefore depends on the generator, the rate and the sampler.

Marginal entropy and neighbour predictability. LFQ has near-uniform token entropy at every vocabulary (0.999 normalised), VQ is also high (0.997–0.998), and FSQ is lower and non-monotonic (0.959, 0.972 and 0.947 at 1K, 2K and 4K). Marginal entropy does not order the tokenizers by generation quality (Spearman -0.20 against the best FID-192 over the nine discrete tokenizers). An exploratory neighbour-conditional predictive gain, measuring how much the left or upper neighbour reduces a token's held-out entropy, separates the three VQ tokenizers from the six LFQ and FSQ ones, but it does not track generation quality either ($\rho = +0.05$ and $+0.17$ for the left and upper neighbour), and the separation itself depends on how the bigram estimates are smoothed. None of the generator-free statistics tested here predicts which tokenizer will generate best.

Table 2: Training-seed variance on the vocabulary-1024 block (default sampling, FID-192 ↓, 10K test samples). Seed 42 is the Table 1 run, recomputed; seeds 43 and 44 retrain the generator from scratch on the same frozen tokenizer. The last row repeats the validation-based sampler selection separately for each trained LFQ-1024 model.

Quantizer	Generator	Seed 42	Seed 43	Seed 44	Mean	SD
VQ-1024	AR	0.52	0.35	0.35	0.40	0.080
VQ-1024	MaskGIT	1.86	2.36	2.26	2.16	0.214
VQ-1024	D3PM	1.30	1.38	1.29	1.32	0.043
LFQ-1024	AR	0.34	0.34	0.49	0.39	0.070
LFQ-1024	MaskGIT	1.90	1.86	1.99	1.92	0.052
LFQ-1024	D3PM	0.42	0.40	0.36	0.39	0.022
LFQ-1024	SEDD	0.41	0.42	0.48	0.44	0.032
FSQ-1024	AR	0.40	0.86	0.44	0.57	0.207
FSQ-1024	MaskGIT	1.15	1.14	1.17	1.15	0.012
FSQ-1024	D3PM	1.19	1.15	1.25	1.20	0.042
LFQ-1024	D3PM, tuned	0.09	0.15	0.10	0.11	–
LFQ-1024	SEDD, tuned	0.10	0.11	0.13	0.11	–

Table 3: Descriptive fixed-effect decomposition of $\log$(FID-192) over the balanced $3 \times 3 \times 6$ discrete panel.

Term	Generator	Vocab. × gen.	Quantizer	Quantizer × gen.	Vocabulary	Other
Share of variation	59.5%	15.0%	10.1%	8.4%	4.4%	2.5%

4.2 Sampler settings change the generator ranking

With default sampling AR appears to be the strongest discrete generator, winning all nine discrete rows, but this is partly a consequence of the sampler defaults. Sampling hyperparameters were swept at vocabulary 1K over temperature, step count, top-k and mask schedule where applicable, with each setting screened on a held-out validation split at 2K samples and the selected setting evaluated once on the test split at 10K samples (Table 4, with grids in Appendix D). On LFQ-1024, D3PM improves from 0.44 to 0.09 and SEDD from 0.41 to 0.10 whilst the step budget falls from 1000 to 100–500, and the same retuning brings both to 0.10–0.15 on FSQ-1024 and VQ-1024, so the three tokenizers that D3PM separates widely at default settings end up within 0.06 of each other. Repeating the selection for the seed-43 and seed-44 models gives 0.15 and 0.10 for D3PM and 0.11 and 0.13 for SEDD (Table 2), so the gain is reproducible across training runs. The selected setting itself does not transfer, with the seed-42 D3PM configuration giving 0.34 when applied unchanged to the seed-43 model. Tuned LFQ-1024 + D3PM reaches the same range as the strongest LDM and RF reference cells, although the continuous references were not given an equivalent sampler sweep.

Table 4: Sampling hyperparameter sensitivity at vocabulary 1K. Values are default → tuned FID-192, both at 10K test samples; settings are selected on the validation split (Appendix D). Training hyperparameters are not retuned in this sweep. Bold marks the LFQ-1024 D3PM and SEDD operating points discussed above.

Generator	LFQ-1024	FSQ-1024	VQ-1024
AR	$0.33 \to 0.33$	$0.39 \to 0.29$	$0.52 \to 0.46$
MaskGIT	$1.91 \to 1.89$	$1.15 \to 0.88$	$1.86 \to 1.42$
DFM	$1.77 \to 1.58$	$3.02 \to 1.47$	$4.42 \to 3.13$
D3PM	$0.44 \to$ **0.09**	$1.21 \to 0.13$	$1.30 \to 0.15$
SEDD	$0.41 \to$ **0.10**	$1.05 \to 0.10$	$1.31 \to 0.13$
BFN	$2.27 \to 2.13$	$7.60 \to 5.80$	$5.98 \to 4.10$

On LFQ-1024, D3PM and SEDD have U-shaped cost-quality curves (Figure 2), with too few steps under-denoising and the canonical 1000-step setting giving worse FID than 100–500 steps, whereas MaskGIT, DFM and BFN improve by 11% or less under retuning. At this grid size AR is also cheap, requiring 64 sequential positions against 100 function evaluations for the best D3PM configuration and 1000 for the default diffusion samplers.

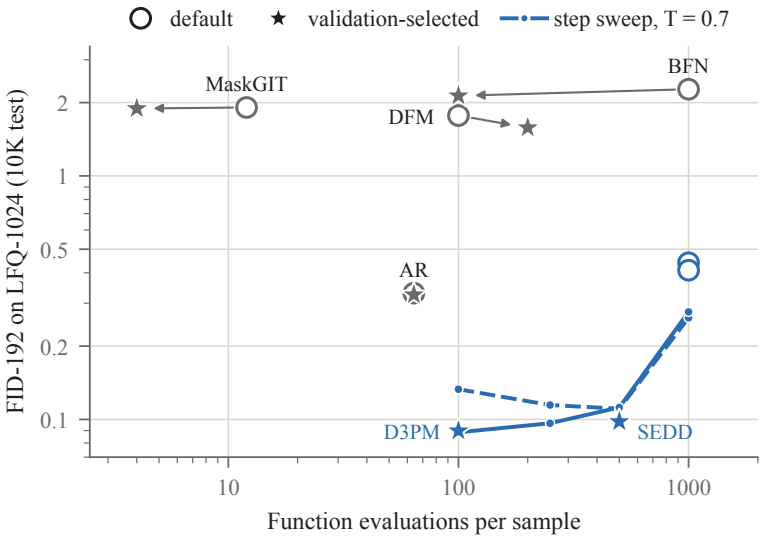


Figure 2: Sampling budget against FID-192 on LFQ-1024, all at 10K test samples. Open circles are the default configurations and stars the validation-selected ones; arrows join the two where both the budget and the score change. Blue lines are D3PM (solid) and SEDD (dashed) evaluated at 100, 250, 500 and 1000 steps with temperature 0.7. The SEDD line and the selected SEDD star are separate generations of the same 500-step setting (0.111 and 0.098).

Throughput. Sampling cost follows the number of forward passes (Appendix B). MaskGIT and AR are the fastest at about 330 and 150 samples/s, the 100-step DFM and RF reach about 37 samples/s, and the 1000-step LDM, D3PM, SEDD and BFN run at 3.5–3.7 samples/s, with RF about 10 times faster than LDM. Retuning LFQ-1024 + D3PM to 100 steps raises its throughput from 3.7 to 33 samples/s whilst improving FID, so the same retuning that changes the quality ranking also changes the cost ranking, although tuned D3PM remains about 4.4 times slower than AR.

4.3 Reconstruction quality does not predict generation quality

Figure 3 plots test-set PSNR against the best FID-192 for each tokenizer, separately for the discrete grid, the VAE channel sweep and the KL sweep. Spearman correlations are close to zero for the discrete and channel panels ($\rho = -0.03$ and $+0.10$), and the four-point KL panel is descriptive only. FSQ-4096 reconstructs best among the discrete tokenizers (29.7 dB) but is not meaningfully better for generation than LFQ-1024 (28.0 dB), and VQ-4096, with a mid-range 29.4 dB, has the lowest FID-192 (0.31). In the channel sweep VAE-c8 is the best reconstructor and the worst generation point, and AE + RF gives the lowest FID in the study (Table 5).

Table 5: Tokenizers from Figure 3 whose reconstruction and generation rankings disagree, with the best FID-192 over generators. Bold marks the highest PSNR and lowest FID-192 within the discrete and the continuous rows.

Tokenizer	PSNR	Best generator	Best FID-192
LFQ-1024	28.0	AR	0.33
VQ-4096	29.4	AR	**0.31**
FSQ-4096	**29.7**	AR	0.37
VAE-c8	**34.4**	RF	0.32
VAE-c16	32.3	RF	0.11
AE	32.3	RF	**0.07**

The continuous panel shows the same decoupling. VAE-c16 has a reduced latent scale (standard deviation 0.62) that hurts LDM, whose fixed-variance noise schedule assumes approximately unit-scale latents, whilst RF remains strong, possibly because its straight-line ODE is less sensitive to latent scale. AE removes the KL constraint altogether and gives less Gaussian latents, yet produces

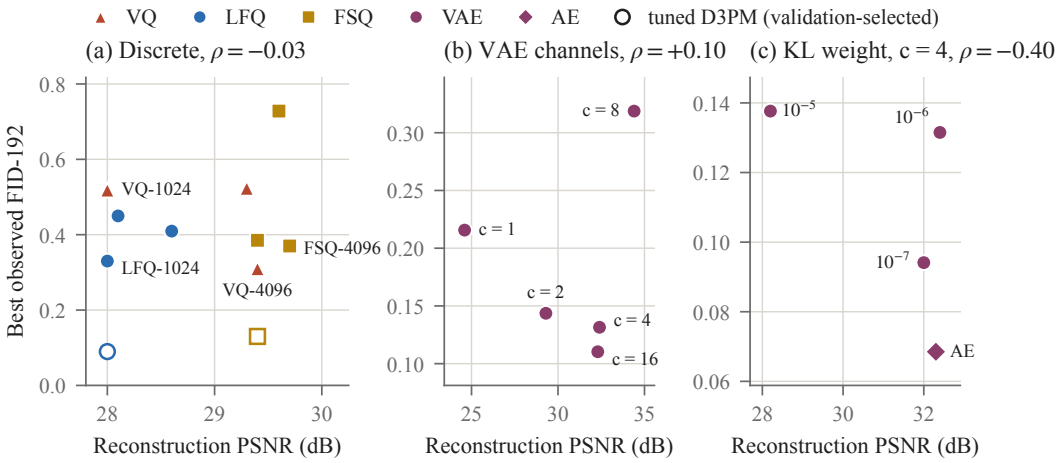


Figure 3: Reconstruction PSNR against the lowest default-configuration FID-192 over all generators paired with each tokenizer, with Spearman ρ in each panel title. (a) The nine discrete tokenizers; hollow markers are the validation-selected D3PM results for LFQ-1024 and FSQ-1024. (b) VAE channel count c at $\lambda_{\mathrm{KL}} = 10^{-6}$. (c) KL weight at $c = 4$, where AE is the $\lambda_{\mathrm{KL}} = 0$ endpoint.

the lowest FID when paired with RF, so even among continuous latents the best representation for one generator is not necessarily the best for another.

4.4 Metric checks and cross-dataset scope

Because FID-192 is not a standard metric, its rankings were checked against three alternatives. Standard FID-2048, computed from the same 10K generated images on 12 cells spanning FID-192 0.07 to 8.44, has a real-vs-real floor of 2.31 (SD 0.016, no negative values over 20 resamples) and rank-agrees with FID-192 at Spearman 0.89. Both metrics place AE + RF, tuned LFQ-1024 + D3PM and the VAE references at the top, and they disagree on the order of some weaker cells and on the LFQ-versus-FSQ pair under AR. Table 6 gives the vocabulary-1024 block under both metrics. A classifier two-sample test (real-vs-generated discriminator AUC, three seeds, 12 cells), which uses neither Inception features nor a Gaussian approximation, agrees with FID-192 at Spearman 0.86; AE + RF reaches AUC 0.54 and the weakest cells approach 1.0. A domain-feature FID from a ChestMNIST ResNet-18 classifier gives Spearman $\rho = 0.71$ over six cells (exact permutation $p = 0.14$; Appendix E), which is consistent with FID-192 but too small a check to confirm it on its own. Paired-bootstrap intervals are narrow relative to the gaps discussed here, for example LFQ-1024 + AR 0.337 ± 0.017 and FSQ-4096 + BFN 8.441 ± 0.158 (Appendix F), but these capture FID sampling noise only, and training-seed variance is covered by Table 2.

Table 6: The vocabulary-1024 block under FID-192 and standard FID-2048, computed on the same 10K generated images per cell. FID-192 values come from this generation pass and so differ from Table 1 by up to 0.07. Bold marks the lowest value per generator and metric.

	VQ-1024		LFQ-1024		FSQ-1024	
Generator	FID-192	FID-2048	FID-192	FID-2048	FID-192	FID-2048
AR	0.53	59	**0.34**	52	0.41	**35**
MaskGIT	1.83	217	1.94	**79**	**1.16**	194
D3PM	1.30	114	**0.42**	**42**	1.14	59

For external validity, six representative cells were retrained on two further MedMNIST datasets. LFQ-1024 + AR remains the best default discrete cell on OrganAMNIST, whereas on PneumoniaMNIST, which has about 17 times fewer training images, LFQ-1024 + D3PM is the best discrete cell and VQ-1024 + AR beats LFQ-1024 + AR (2.09 vs 2.73), the reverse of their order on the two larger datasets. Training-set size does not explain the reversal. When ChestMNIST is subsampled to 4.7K, 20K and the full 78K images and AR is retrained, LFQ-1024 beats VQ-1024 at every size (0.95 vs 1.11, 0.37

vs 0.49 and 0.34 vs 0.52), by a roughly constant margin. The cause of the PneumoniaMNIST result remains unidentified; neither reconstruction nor dataset size is a safe rule for choosing the tokenizer. AE + RF is the most consistent configuration, with the lowest FID on all three datasets.

Table 7: Cross-dataset replication for six representative cells (FID-192 ↓, default sampling). Real-vs-real floors are 0.053 on PneumoniaMNIST and 0.0089 on OrganAMNIST at the largest disjoint split available. Bold marks the best discrete and the best continuous cell in each column.

Cell	ChestMNIST	PneumoniaMNIST	OrganAMNIST
LFQ-1024 + AR	**0.33**	2.73	**2.12**
FSQ-1024 + AR	0.39	5.52	4.87
VQ-1024 + AR	0.52	2.09	3.84
LFQ-1024 + D3PM	0.44	**1.93**	4.89
VAE-c2 + LDM	0.14	4.14	9.88
AE + RF	**0.07**	**1.05**	**0.94**

4.5 Failure modes by generator family

MaskGIT is fast, parallel and supports inpainting without retraining (Appendix G), but as an unconditional generator it is uneven. Under FID-192 it does best on FSQ at vocabulary 1K (FID-2048 prefers LFQ; Table 6). It degrades at vocabulary 2K on both VQ and LFQ, and the sampler sweep does little for it on LFQ-1024. A plausible cause is that confidence-based parallel decoding depends on the calibration of the token predictions, which neither marginal entropy nor reconstruction fidelity measures.

DFM uses 100 steps, an order of magnitude fewer than the 1000-step diffusion samplers, yet its FID remains behind the best AR, diffusion and continuous-flow cells. Retuning helps it on every tokenizer (to 1.58, 1.47 and 3.13 on LFQ-1024, FSQ-1024 and VQ-1024, two of them by moving to 200 Euler steps), but it stays behind tuned D3PM and SEDD. This result is specific to DFM at this training budget and token resolution.

BFN is the most consistent failure in the discrete panel, with elevated FID on every tokenizer family and particularly poor results on FSQ at high vocabulary. Its confidence-state dynamics may be poorly matched to large categorical vocabularies at the capacity and training budget shared with the other generators.

The continuous reference cells also show that discrete tokenizers are not always preferable, since AE + RF is the best default cell in the study and tuned LFQ-1024 diffusion only reaches the same range after sampler retuning. The continuous panel is a reference sweep under the same resolution and evaluation protocol and is not rate-matched to the discrete tokenizers, so the results argue for evaluating the representation and the generative dynamics together, and do not establish a general preference for LFQ or RF.

4.6 Nearest-neighbour memorization probes

A low FID at low resolution could in principle arise from copying training images, so a feature-space nearest-neighbour probe was run over all 70 ChestMNIST cells in the same InceptionV3 192-dimensional space as FID-192 (Figure 5). Across the grid the minimum distance ratio R is 0.975 and the maximum AuthPct is 0.107. A raw pixel and LPIPS check on seven representative cells gives the same broad family ordering but places continuous cells closer to the training set, with LFQ-1024 + AR at $R_{\text{pixel}} = 1.116$ and $R_{\text{LPIPS}} = 1.289$, VQ-1024 + MaskGIT further from the training set (1.22 and 2.43), and AE + LDM closest (0.873 and 0.821) but with a non-zero minimum LPIPS distance (0.028, against 0.032 for held-out real images). Neither probe indicates exact copying, although neither amounts to a membership-inference audit.

5 Discussion and limitations

Across this controlled grid, tokenizer, generator and sampler had to be evaluated together to obtain a reliable ranking. LFQ-1024 was the strongest default discrete tokenizer and AE + RF the strongest

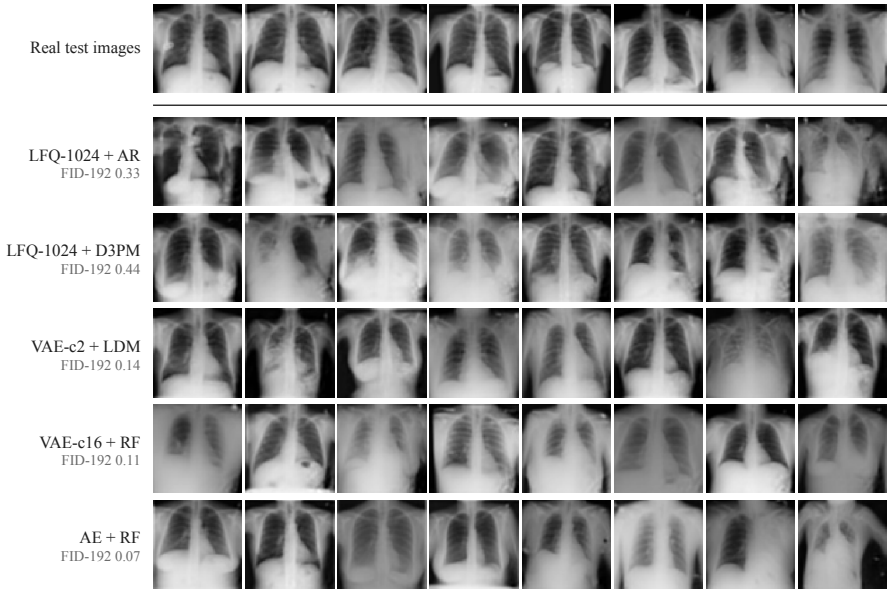


Figure 4: Uncurated samples from five of the lowest-FID default cells, with real test images in the top row. All samples use default sampling and a fixed seed.

continuous reference, and tuned D3PM and SEDD on LFQ-1024 showed that default samplers can understate iterative discrete generators by a factor of four to five in FID-192. These results fit a rate-distortion-modelability view of tokenizer quality, in which reconstruction metrics are useful diagnostics but are not adequate criteria for selecting a latent representation.

Implications for medical image generation studies. Studies should report the tokenizer as part of the model, including the quantizer family, grid shape, vocabulary or channel count, reconstruction metrics, token entropy or latent-scale statistics, generator family and the sampler budget used for the reported samples, since two systems with the same named generator may otherwise be solving quite different latent modelling problems.

Tokenizer selection should also include at least one downstream generation check, because in this grid reconstruction alone would have selected the wrong tokenizer in both the discrete and the continuous panels. A practical workflow is to use reconstruction as a floor, discard tokenizers that visibly fail to preserve the anatomy of interest, and choose among the remainder with a small matched sweep over generators and sampler settings.

For 64×64 ChestMNIST-like benchmarks, LFQ-1024 is a reasonable discrete starting point and AE + RF the strongest continuous one. VQ is worth testing with AR, where it matched the best LFQ and FSQ cells, and FSQ is worth including because its ranking against LFQ changed with both the generator and the metric. These suggestions apply to benchmark comparisons, not to clinical use.

The study is low-resolution, greyscale and mostly on ChestMNIST, and the two-dataset replication is small relative to the design space. Only the vocabulary-1024 interaction block was trained at three seeds, so per-cell variance and rank stability are not estimated for the rest of the grid. FID-192 is an internal ranking metric (§3) whose absolute values are not comparable to FID-2048 and which is not a clinical measure, and the specific AR and MaskGIT winners in the seeded block differ between FID-192 and FID-2048. The models are unconditional and do not learn pathology-conditional distributions, and there is no radiologist evaluation, synth-then-classify utility study, membership-inference audit or patient-level duplication audit. Finally, the sampler sweep uses a finite validation grid and evaluates each selected configuration once on the test split. The results are therefore a controlled case study of tokenizer–generator–sampler interaction in this grid and do not rank medical image generators in general.

Reproducibility. All experiments use public MedMNIST datasets [Yang et al., 2023]. The tokenizers are implemented in MEDTOKENIZERS (https://github.com/liamchalcroft/medtokenizers) and

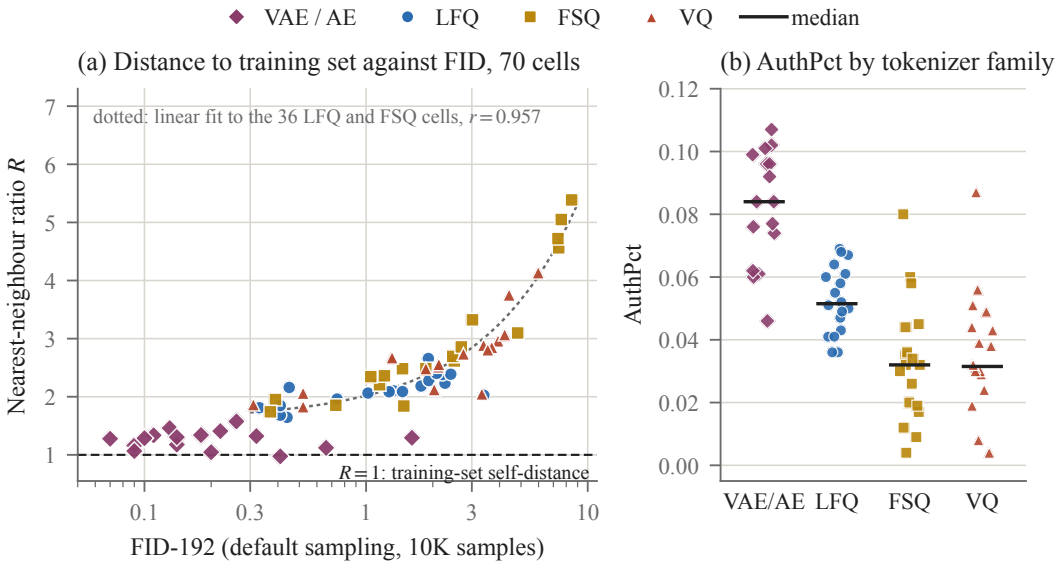


Figure 5: Nearest-neighbour memorization probe in the InceptionV3 192-dimensional feature space used by FID-192, over all 70 cells. (a) R is the median distance from a generated sample to its nearest training image divided by the median nearest-neighbour distance within the training set, so $R < 1$ would indicate samples closer to training images than training images are to each other. (b) AuthPct is the fraction of generated samples closer to a training image than that image is to its own nearest neighbour; horizontal bars are family medians.

the generators, samplers and evaluation code in MEDLATENTS (https://github.com/liamchalcroft/medlatents), both MIT-licensed and installable from PyPI; the experiment scripts, result files and figure scripts are in a separate repository (https://github.com/liamchalcroft/tokenizer-generator-coupling), and an extended version of this paper is available as arXiv:2608.07713. The main grid was run on a single NVIDIA A6000 48GB GPU with a budget of about 556 GPU-hours (80.6 for generation and evaluation, 53.4 for sampler sweeps and an estimated 422.1 for training). The LFQ and FSQ seed runs, the LFQ data-scale runs and the FID-2048 evaluation of the LFQ and FSQ D3PM cells used a single NVIDIA L40S with the same locked package versions; on that GPU LFQ-1024 + D3PM gives a default FID-192 of 0.43 (0.44 on the A6000) and a tuned FID-192 of 0.0885 (0.0898). The VQ seed and data-scale runs, the remaining FID-2048 evaluations and the two-sample tests used the A6000; re-evaluating six cells on both GPUs gives a mean difference of -0.004 in FID-192.

Acknowledgments and Disclosure of Funding

This work received no specific funding. The author declares no competing interests.

References

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *NeurIPS*, 2021.

Zechen Bai, Jianxiong Gao, Ziteng Gao, Pichao Wang, Zheng Zhang, Tong He, and Mike Zheng Shou. Factorized visual tokenization and generation. *arXiv preprint arXiv:2411.16681*, 2024.

Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *ICML*, 2019.

Christian Bluethgen, Pierre Chambon, Jean-Benoit Delbrouck, et al. A vision-language foundation model for the generation of realistic chest x-ray images. *Nature Biomedical Engineering*, 9: 494–506, 2025. Published online 2024.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In *ICML*, 2024.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. MaskGIT: Masked generative image transformer. In *CVPR*, 2022.

Nicki Skafte Detlefsen, Jiri Borovec, Justus Schock, Ananya Harsh Jha, Teddy Koker, Luca Di Liello, Daniel Stancl, Changsheng Quan, Maxim Grechkin, and William Falcon. TorchMetrics – measuring reproducibility in PyTorch. *Journal of Open Source Software*, 7(70):4101, 2022. doi: 10.21105/joss.04101.

Sander Dieleman. Generative modelling in latent space. *Blog post*, 2025. URL https://sander.ai/2025/04/15/latents.html. Accessed 2026-06-18.

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

Itai Gat, Tal Remez, Felix Kreuk, et al. Discrete flow matching. In *NeurIPS*, 2024.

Alex Graves, Rupesh Kumar Srivastava, Timothy Atkinson, and Faustino Gomez. Bayesian flow networks. *arXiv:2308.07037*, 2023.

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.

Firas Khader et al. Denoising diffusion probabilistic models for 3d medical image generation. *Scientific Reports*, 2023.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Huawei Lin, Tong Geng, Zhaozhuo Xu, and Weijie Zhao. VTBench: Evaluating visual tokenizers for autoregressive image generation. *arXiv preprint arXiv:2505.13439*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *ICML*, 2024.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Walter HL Pinaya et al. Generative AI for medical imaging: extending the MONAI framework. *arXiv:2307.15208*, 2023.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Liu, and Yunfeng Wen. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024.

Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. *arXiv:1812.05069*, 2018.

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

Maya Varma, Ashwin Kumar, Rogier van der Sluijs, Sophie Ostmeier, Louis Blankemeier, Pierre Chambon, Christian Bluethgen, Jip Prince, Curtis Langlotz, and Akshay Chaudhari. MedVAE: Efficient automated interpretation of medical images with large-scale generalizable autoencoders. *arXiv:2502.14753*, 2025.

Mark Weber, Lijun Yu, Qihang Yu, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. MaskBit: Embedding-free image generation via bit tokens. *Transactions on Machine Learning Research*, 2024.

Junfeng Wu, Dongliang Luo, Weizhi Zhao, Zhihao Xie, Yuanhao Wang, Junyi Li, Xudong Xie, Yuliang Liu, and Xiang Bai. TokBench: Evaluating your visual tokenizer before visual generation. *arXiv preprint arXiv:2505.18142*, 2025.

Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. GigaTok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation. In *ICCV*, 2025.

Jiancheng Yang et al. MedMNIST v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 2023.

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

Srikar Yellapragada et al. PathLDM: Text conditioned latent diffusion model for histopathology. In *WACV*, 2024.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Lisa Dunlap, et al. Language model beats diffusion – tokenizer is key to visual generation. In *ICLR*, 2024.

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.

A Tokenizer and latent details

Table 8: Factorial controls used in the study.

Component	Protocol
Primary dataset	ChestMNIST, 64×64, 78K train / 22.4K test images
Replication	PneumoniaMNIST and OrganAMNIST at 64×64 for six representative cells
Discrete tokenizers	VQ/LFQ/FSQ $\times$ vocabularies 1K/2K/4K, shared encoder-decoder
Discrete generators	AR, MaskGIT, DFM, D3PM, SEDD, and BFN on 8×8 token grids
Continuous references	Eight VAE/AE settings spanning channel and KL sweeps, including the AE no-KL endpoint, each with LDM and RF
Evaluation	10K samples per default cell; FID-192 primary; domain-FID, bootstrap CIs, throughput, and nearest-neighbour probes as checks

Discrete tokenizers use 64 latent positions. VQ, LFQ and FSQ are trained at vocabularies 1024, 2048 and 4096, corresponding to 640, 704 and 768 raw token bits per image before entropy coding. Reconstruction PSNR spans a narrow range for the discrete grid: VQ 28.0–29.4, LFQ 28.0–28.6, and FSQ 29.4–29.7 (Table 9). Continuous VAE reconstruction improves with channel count up to c8 (34.4 dB) and then regresses at c16 (32.3 dB). The KL-strength sweep holds $c = 4$ fixed and varies λ from AE ($\lambda = 0$) through weak and strong VAE regularisation.

Table 9: Tokenizer reconstruction quality on ChestMNIST test images. PSNR and SSIM are higher-is-better; LPIPS is lower-is-better.

Tokenizer	Type	Vocab/rate	PSNR	SSIM	LPIPS
VQ-1024	Discrete	1K	28.0	0.891	0.026
VQ-2048	Discrete	2K	29.3	0.924	0.020
VQ-4096	Discrete	4K	29.4	0.925	0.018
LFQ-1024	Discrete	1K	28.0	0.893	0.024
LFQ-2048	Discrete	2K	28.1	0.894	0.025
LFQ-4096	Discrete	4K	28.6	0.904	0.021
FSQ-1024	Discrete	1K	29.4	0.912	0.020
FSQ-2048	Discrete	2K	29.6	0.915	0.018
FSQ-4096	Discrete	4K	29.7	0.916	0.019
VAE-c1	Continuous	$c = 1$	24.6	0.855	0.130
VAE-c2	Continuous	$c = 2$	29.3	0.896	0.027
VAE-c4	Continuous	$c = 4$	32.4	0.940	0.013
VAE-c8	Continuous	$c = 8$	34.4	0.963	0.011
VAE-c16	Continuous	$c = 16$	32.3	0.949	0.014
VAE-10^{-5}	Continuous	$c = 4, \lambda = 10^{-5}$	28.2	0.881	0.028
VAE-10^{-7}	Continuous	$c = 4, \lambda = 10^{-7}$	32.0	0.937	0.014
AE	Continuous	$c = 4, \lambda = 0$	32.3	0.940	0.012

Table 10: Tokenizer settings used in the controlled grid. All rows use the same convolutional encoder-decoder family at 8×8 latent resolution. AE is the no-KL endpoint of the continuous family.

Family	Setting varied in study	Training/loss detail
VQ	$K \in \{1024, 2048, 4096\}$	learned codebook; shared decoder/discriminator recipe
LFQ	$K \in \{1024, 2048, 4096\}$	lookup-free binary code; shared decoder/discriminator recipe
FSQ	effective $K \in \{1024, 2048, 4096\}$	scalar-level quantization; shared decoder/discriminator recipe
VAE channel sweep	$c \in \{1, 2, 4, 8, 16\}$	KL weight $\lambda = 10^{-6}$
VAE KL sweep	$\lambda \in \{0, 10^{-7}, 10^{-6}, 10^{-5}\}$	latent channels fixed at $c = 4$; $\lambda = 0$ is AE

Figure 6 shows the spatial entropy pattern at vocabulary 1024.

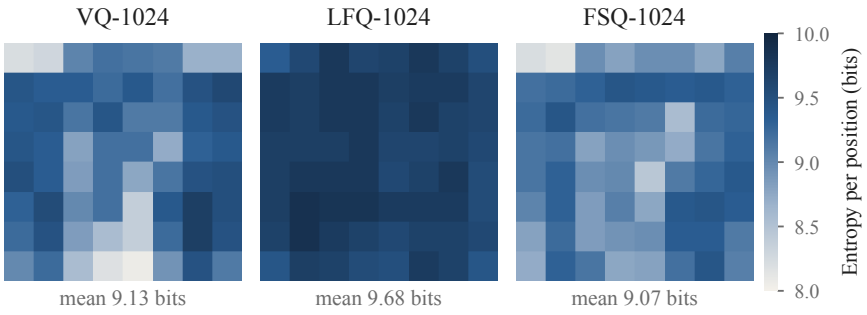


Figure 6: Shannon entropy of the token at each of the 8×8 latent positions over the ChestMNIST test set, at vocabulary 1024 (maximum 10 bits). Positions are laid out as in the image, so the top row corresponds to the top of the radiograph.

Continuous latent statistics. Across the channel and KL sweeps, test-split latents have means consistently $|\mu| < 0.1$, but their standard deviations differ in ways that may account for generator-specific behaviour. VAE-c16 has $\mathrm{std} = 0.62$ and $|z|_{\mathrm{mean}} = 0.35$; strong KL ($10^{-5}$) reduces std to 0.54; AE ($\lambda = 0$) has $\mathrm{std} = 2.03$ and $|z|_{\mathrm{mean}} = 1.60$. LDM is sensitive to these deviations because

its noise schedule assumes approximately unit-scale latents, whereas RF appears less sensitive to latent magnitude.

B Inference throughput

Table 11: Generation throughput for 10K samples at 8×8 token resolution and vocabulary 1,024 on one NVIDIA RTX A6000, as used for the main grid. Timings cover sampling and decoding but not model loading or metric computation. Samples/s is the median of three timed passes after a warm-up batch (one pass for the 1,000-step generators), in float32 at batch size 128 (64 for the tuned rows), with randomly initialised weights, since the cost of a step does not depend on them. Default rows use the canonical settings; tuned rows are the validation-selected LFQ-1024 settings. "Parallel decoding" indicates whether several tokens are predicted per forward pass.

Generator	Steps	Samples/s	Time for 10K	Parallel decoding
MaskGIT	12	333	30 s	yes
AR transformer	64	146	68 s	no
DFM	100	37.0	270 s	yes
D3PM	1000	3.67	2726 s	yes
D3PM, tuned (LFQ-1024)	100	33.2	301 s	yes
SEDD	1000	3.67	2724 s	yes
SEDD, tuned (LFQ-1024)	500	6.60	1515 s	yes
BFN	1000	3.48	2878 s	yes
RF	100	37.7	265 s	yes
LDM	1000	3.63	2755 s	yes

C Implementation notes

Matrix-free absorbing D3PM. Standard D3PM precomputes transition matrices $\mathbf{Q}_t, \bar{\mathbf{Q}}_t \in \mathbb{R}^{(K+1)\times(K+1)}$ for all T timesteps, which requires $O(TK^2)$ memory. For absorbing transitions, the posterior $q(x_{t-1} \mid x_t, x_0)$ has closed form: unmasked tokens persist, while masked tokens are drawn from a two-point distribution over $\{x_0, \mathrm{MASK}\}$ with probabilities proportional to $\beta_t \bar{\alpha}_{t-1}$ and $1 - \bar{\alpha}_{t-1}$. The implementation reduces storage to $O(T)$ and per-step computation to $O(BLK)$ for batch size B, sequence length L and vocabulary size K.

FID-192 choice. FID-192 is defined in Section 3. Standard FID-2048 is also usable at 10K samples (Section 4); FID-192 is preferred because its lower real-vs-real floor ($\sim$0.002) gives better separation among the strongest cells, and its absolute values are not comparable to FID-2048.

D Sampling sweep grids

Each candidate setting in Table 4 was screened with 2K generated samples against the held-out validation split ($n = 11{,}219$), and the selected setting was then evaluated once on the test split with 10K samples. The grids, with the number of configurations per tokenizer in brackets, were: MaskGIT temperature $\{0.5, 0.7, 0.9, 1.0, 1.2\}$, refinement steps $\{4, 8, 12, 16, 24, 32\}$ and mask schedule $\{\mathrm{cosine}, \mathrm{linear}\}$ (60); BFN temperature $\{0.3, 0.5, 0.7, 0.9, 1.0, 1.2\}$ and steps $\{10, 25, 50, 100, 200\}$ (30); DFM temperature $\{0.5, 0.7, 0.9, 1.0, 1.2\}$ and Euler steps $\{10, 25, 50, 100, 200\}$ (25); D3PM and SEDD temperature $\{0.5, 0.7, 0.9, 1.0, 1.2\}$ and timesteps $\{100, 250, 500, 1000\}$ (20 each); AR temperature $\{0.5, 0.7, 0.8, 0.9, 1.0, 1.2\}$ and top-$k \in \{0, 64, 256\}$ (18). D3PM and SEDD therefore had two of the smaller grids, and their large gains are not a product of a larger search. The continuous LDM and RF references were not swept.

E FID-192 domain-feature sanity check

FID was recomputed on six representative ChestMNIST cells using 512-dimensional features from a ChestMNIST ResNet-18 classifier trained on real images (mean test AUC 0.761). The rank correlation

between FID-192 and domain-FID is 0.71 (exact permutation $p = 0.14$). The disagreements are among the three AR cells, whose domain-FIDs lie within 0.6 of each other, and between VQ-1024 + AR and LFQ-1024 + D3PM.

Table 12: Domain-feature sanity check. Lower is better; Spearman $\rho = 0.71$.

Cell	FID-192	Domain-FID
LFQ-1024 + AR	0.33	17.08
FSQ-1024 + AR	0.39	16.53
VQ-1024 + AR	0.52	16.93
LFQ-1024 + D3PM	0.44	24.70
VAE-c2 + LDM	0.14	7.67
AE + RF	0.07	0.74

F Bootstrap confidence intervals

Table 13: Paired-bootstrap 95% confidence intervals for representative FID-192 cells ($B = 200$, 10K generated samples).

Cell	Mean	Std.	2.5%	97.5%
AE + LDM	0.093	0.004	0.086	0.101
LFQ-1024 + AR	0.337	0.009	0.320	0.354
LFQ-1024 + MaskGIT	1.938	0.029	1.876	1.988
FSQ-4096 + BFN	8.441	0.077	8.281	8.597

G Qualitative probes

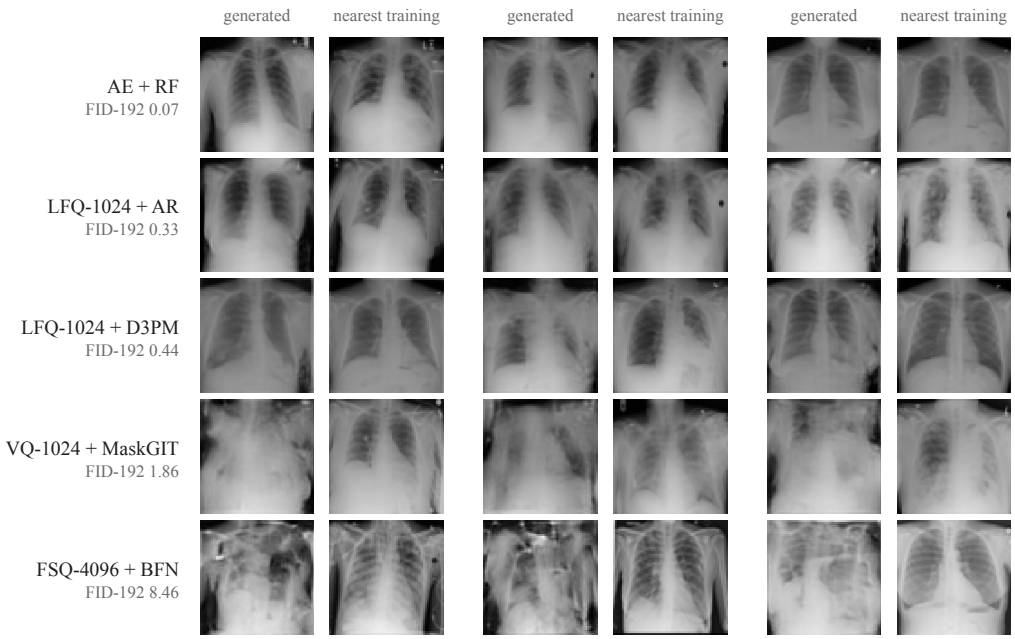


Figure 7: The three closest generated-to-training pairs for five cells under the feature-space probe, with the generated sample on the left of each pair.

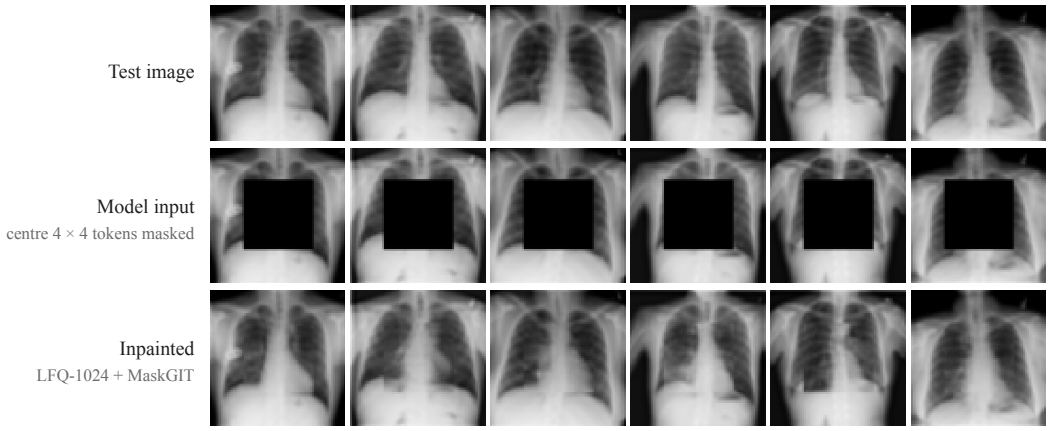


Figure 8: Centre inpainting with LFQ-1024 + MaskGIT and no retraining. The central 4×4 block of the 8×8 token grid is masked and filled over 12 refinement steps.

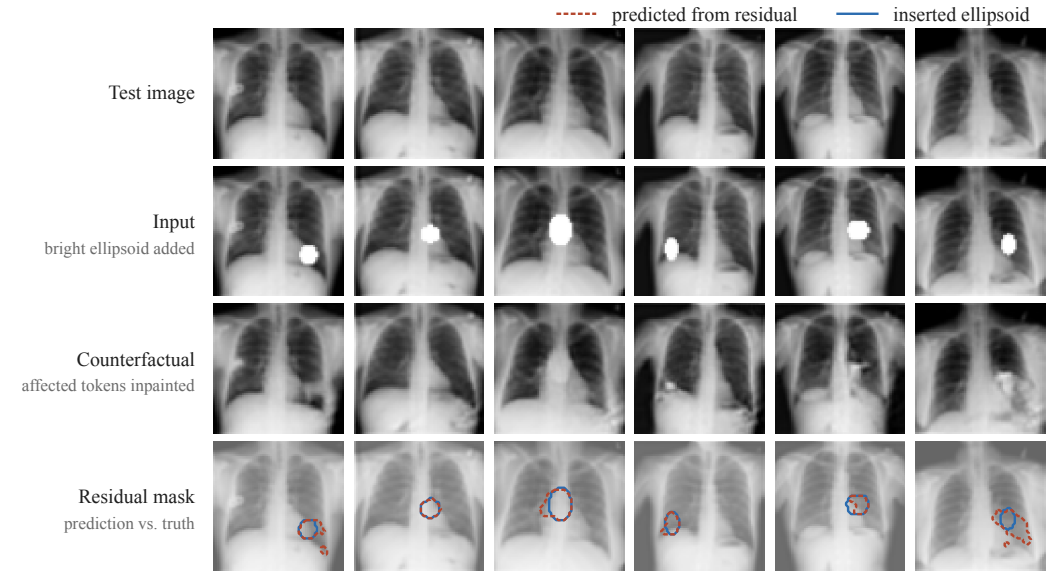


Figure 9: Counterfactual inpainting probe with LFQ-1024 + MaskGIT. A bright ellipsoid is added to each test image, the tokens it touches are inpainted, and thresholding the input-counterfactual residual gives the predicted mask, drawn against the inserted ellipsoid in the bottom row. The probe illustrates a use of the token model and is not evaluated as an anomaly-detection or clinical method.

NeurIPS Paper Checklist

1. **Claims.** Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The abstract and introduction state the controlled low-resolution scope, the factorial design, and the limitations around generalization.
2. **Limitations.** Does the paper discuss the limitations of the work performed by the authors? [Yes] Section 5 lists dataset, resolution, single-seed wider grid (the interaction block is three-seed), metric, conditioning, clinical-evaluation, and sampler-selection limitations.
3. **Theory, assumptions and proofs.** If you are including theoretical results, did you state the full set of assumptions of all theoretical results, and did you include complete proofs? [NA] The paper is empirical and does not present new theoretical theorems.
4. **Experimental result reproducibility.** If the contribution is empirical, what steps did you take to make results reproducible or verifiable? [Yes] We specify the design, datasets, metrics, sample counts, compute, and result summaries needed to audit the reported tables.
5. **Open access to data and code.** If you ran experiments, did you include the code, data, and instructions needed to reproduce the main experimental results? [Yes] The datasets are public MedMNIST datasets, and the code is released as the MIT-licensed MEDTOKENIZERS and MEDLATENTS packages (GitHub and PyPI), with the experiment scripts, result files and figure scripts in a separate public repository.
6. **Experimental setting/details.** If you ran experiments, did you specify all the training details? [Yes] The main paper specifies the grid, tokenizer/generator families, data splits, sample counts, metric choice, and sampling-hyperparameter sweep protocol; further implementation details are in the appendix.
7. **Experiment statistical significance.** Does the paper report error bars suitably and correctly defined or other appropriate information about statistical significance? [Yes] We report bootstrap confidence intervals for representative FID estimates, and we estimate training-seed variance on the vocabulary-1024 interaction block (three seeds; median per-cell std 0.047) and use three pooled seed standard deviations to decide which quantizer comparisons to claim; the wider grid remains single-seed.
8. **Experiments compute resources.** Does the paper provide sufficient information on compute resources? [Yes] Section 5 reports the GPUs used and the compute budget in GPU-hours for generation, evaluation, sampler sweeps and training.
9. **Code of ethics.** Have you read the NeurIPS Code of Ethics and ensured that your research conforms to it? [Yes] The work uses public benchmark data, reports limitations, and does not claim clinical deployment readiness.
10. **Broader impacts.** If appropriate, did you discuss potential negative societal impacts? [Yes] The discussion states that the models are unconditional and not clinically evaluated and that the results are not recommendations for clinical use, and it reports nearest-neighbour memorization probes while noting that they are not a membership-inference audit.
11. **Safeguards.** Do you have safeguards in place for responsible release of models with high risk for misuse? [NA] The work does not release a high-risk foundation model; it studies low-resolution public benchmark images and includes explicit non-clinical-use limitations.
12. **Licenses.** If you are using existing assets, did you cite the creators and respect licenses and terms of use? [Yes] The paper cites MedMNIST and the modeling components used; both released packages are MIT-licensed.
13. **Assets.** If you are releasing new assets, did you document them and provide relevant details? [Yes] The MEDTOKENIZERS and MEDLATENTS packages are released with documentation, licence files and citation metadata. No new dataset or model checkpoint is released.
14. **Crowdsourcing and research with human subjects.** If you used crowdsourcing or conducted research with human subjects, did you include relevant details? [NA] No crowdsourcing or new human-subject data collection was performed.
15. **IRB approvals.** Did you describe participant risks and obtain IRB approvals where applicable? [NA] The work uses already-public benchmark datasets and does not involve new human-subject recruitment or intervention.

16. **Declaration of LLM usage.** Does the paper describe LLM usage if it is an important, original, or non-standard component of the core methods? [NA] LLMs are not part of the scientific method, model, data generation, evaluation, or experiments.