How Much of Your Codebook Survives Denoising?

Anonymous authors
Paper under double-blind review

Abstract

Two-stage image generation compresses an image into discrete codewords—a tokenizer with codebook size K—then models the code sequences. How large should K be? Reported optima span two orders of magnitude. We treat the trained second-stage generator as a measurable noisy channel: renoising and redenoising real latents yields its per-token confusion kernel, and a surrogate computed from the codebook alone reproduces the measurement within its validity domain. Three rates must be distinguished—nominal ($\log_2 K$), used ($H(q)$), and transferred through the probe—and three findings follow. *Snapping*: diffusion outputs land exactly on the codeword grid, so generation error means committing to the wrong codeword, invisible to dispersion-based diagnostics. *Saturation*: across CIFAR and ImageNet (64–256px) the nominal rate keeps growing while probe transfer stalls; in the converged CIFAR-10 sweep, about 8 of 15 nominal bits survive our stress probe. The saturation point, computed from the codebook alone, bounds the useful range of K before any stage-2 is trained: the best gFID inside this range matches the global best on both CIFAR variants and trails it by 0.1 on ImageNet. *Allocation*: in our matched-budget experiments, the catastrophic large-K collapse of autoregressive models is driven by vocabulary crowding out core parameters rather than by K itself, while diffusion failures trace to stage-2 commitment capacity or codebook geometry—interventions on each remove them. All numerical predictions were archived before measurement; five falsified predictions are reported.

1 Introduction

Most modern image generation systems are built in two stages (van den Oord et al., 2017; Esser et al., 2021; Rombach et al., 2022). **Stage 1** trains a tokenizer: an encoder maps an image to a grid of latents, each replaced by its nearest neighbour in a learned *codebook* of K codewords, and a decoder reconstructs; an image becomes a sequence of discrete indices ($8\times8{=}64$ in our main sweeps; 16×16 at 256px). **Stage 2** trains a generative model—autoregressive (AR) transformer (Esser et al., 2021), masked model (Chang et al., 2022), or diffusion model—over these index sequences; sampling and decoding produce new images.

The codebook size K is the central capacity knob of this pipeline, and its choice is governed by convention rather than measurement. Enlarging K monotonically improves *reconstruction*; its effect on *generation* is notoriously inconsistent: reported sweet spots span 10^2–10^3 (van den Oord et al., 2017; Esser et al., 2021) to $16{,}384$ (Sun et al., 2024), with degradation beyond $32{,}768$ (Chen et al., 2025) and compute-dependent optima in language (Tao et al., 2024)—two orders of magnitude, with no analysis predicting another setting's optimum. The phenomenon behind this confusion—reconstruction keeps improving while generation stalls or worsens—is often called the *reconstruction–generation trade-off* (Ramanujan et al., 2024).

This paper identifies the missing variable—stage-2 token *recoverability*—and shows that it is *measurable*. Our central move: the trained stage-2 generator induces a **measurable family of denoising channels**, whose single-letter projections provide controlled stress tests of token recoverability. This opens tokenizer design to the machinery of joint source–channel coding, where quantizers are designed for noisy channels (COVQ; Farvardin, 1990; Farvardin & Vaishampayan, 1991), in a form not previously available: a trained *generative model* as the channel, with measured kernels and preregistered codebook-side tests. This also narrows the choice of codebook size before any stage-2

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

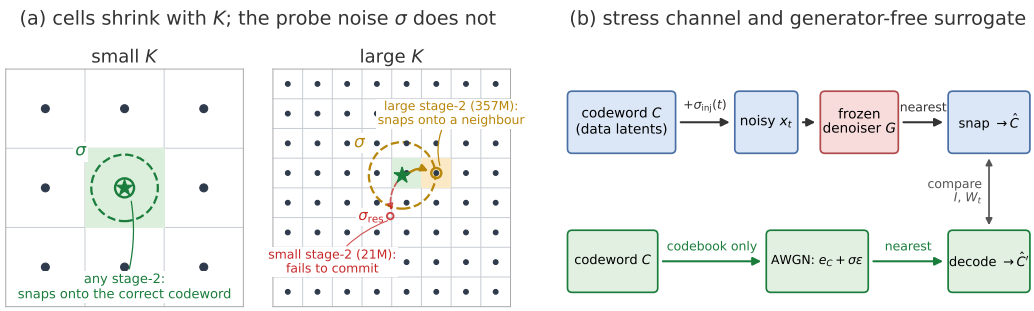


Figure 1: (a) Error geometry: Voronoi cells shrink as K grows while the probe noise scale σ is fixed (schematic: cell size relative to σ, not the exact P_e of Figure 2). At small K any stage-2 snaps back onto the correct codeword. At large K a large stage-2 (357M) still commits, but onto a near neighbour—commitment error, inherent to the geometry—while a small stage-2 (21M) fails to commit at all (dispersion, σ_{res}): not a geometric consequence of K but a symptom of insufficient stage-2 capacity, repaired by scaling (§5.1). (b) The roundtrip stress channel and its generator-free AWGN surrogate, compared at information and kernel level ($\hat{C}$, $\hat{C}'$: the two paths' outputs).

is trained: from the codebook alone, the surrogate shows at which K further increases in nominal rate no longer raise the transferred information—an upper bound on the range worth training—and it identifies codebooks whose transferred rate is abnormally low. Choosing K remains a trade between reconstruction and generation; the generation side of that trade can now be computed from the codebook in seconds.

Contributions.

1. **Operational channel measurements** (§3): terminal dispersion, renoise–denoise confusion, and a generator-free AWGN surrogate, each with calibration, stability criteria, and an explicit validity test.
2. **Commitment vs. precision error** (§3.1): given adequate stage-2 capacity, trained embedding-space diffusion places outputs *on* the codeword grid yet commits a large fraction of tokens to the wrong codeword under moderate perturbation. Dispersion-based diagnostics are blind to this error mode, so codebook-side rates must be evaluated at the injected—not terminal—noise level.
3. **A validated free diagnostic** (§3.3): the surrogate reproduces the measured transfer in its validity domain and, where estimable, the confusion kernel itself; the residual is an interpretable model–surrogate gap. Codebook stress curves cost seconds and no stage-2 training, and bound the range of K worth training a stage-2 on (§5).
4. **Pre-registered predictions and their tests** (§4, §5): generator-measured transfer tracks codebook-side predictions archived before measurement across three datasets and 32–256px; observed diffusion degradations are removed by interventions on stage-2 capacity and codebook geometry, with no evidence of an unavoidable cost of K alone; AR outcomes track vocabulary-vs-core allocation, the collapse causally attributed to core starvation (the vocabulary crowding out core parameters); five falsified predictions are reported.

2 Background and Setup

Stage 1: VQ tokenizer. An encoder E, codebook $\mathcal{C} = \{e_1, \ldots, e_K\} \subset \mathbb{R}^d$, decoder D; an image encodes to a latent grid, each position quantizes to its nearest codeword, giving indices $C(x) \in \{1, \ldots, K\}^L$ ($L = hw$), and $D(e_{C(x)})$ reconstructs. We write $q(k)$ for the empirical usage frequency of codeword k on the training set, $H(q)$ for its entropy (the *used* rate, in bits), and Σ_c for the usage-weighted codebook covariance with eigenvalues $\{\lambda_i\}_{i=1}^d$. Throughout, all geometric and noise quantities live in *normalized coordinates*: embeddings are standardized per channel by the usage-weighted mean and standard deviation, and the codebook is transformed identically.

Stage 2: generator. A generative model G trained on the quantized latents $e_{C(x)}$ produces at sampling time a continuous latent $\hat{z} \in \mathbb{R}^{L \times d}$, quantized to $\hat{C}$ and decoded by D. Ours is a small DiT

Table 1: Stage-1 sweeps (d=8, 8×8 tokens, EMA VQ-VAE with dead-code revival; identical architecture for all K, extra downsampling for 64/128px inputs; 128×128 in §5.1, CIFAR-100 replication in Appendix T). Reconstruction MSE improves monotonically everywhere ($13.8 \to 8.1$, $30.9 \to 21.5$, $42.1 \to 32.3 \times 10^{-3}$); the *used* rate $H(q)$ and utilization saturate on the CIFAR sources but at neither ImageNet resolution (source-associated ceiling; size-sweep rules out dataset size, §5.1).

	K	256	512	1024	2048	4096	8192	16384	32768
	$\log_2 K$ [b]	8	9	10	11	12	13	14	15
CIFAR-10	$H(q)$ [b]	7.80	8.80	9.78	10.75	11.84	12.91	13.80	13.95
	used (%)	100	100	100	100	100	100	92	**53**
ImageNet 64×64	$H(q)$ [b]	7.78	8.76	9.73	10.68	11.83	12.86	13.90	**14.67**
	used (%)	100	100	100	100	100	100	100	**98**
ImageNet 128×128	$H(q)$ [b]	7.78	8.74	9.73	10.72	11.76	12.93	13.90	**14.68**
	used (%)	100	100	100	100	100	100	100	**98**

(21M; Peebles & Xie, 2023) with v-prediction, cosine schedule and DDIM (Salimans & Ho, 2022; Nichol & Dhariwal, 2021; Song et al., 2021), *frozen across all codebook sizes* so the generator, not its tuning, is the object of study. Categorical stage-2 models instead parameterize a K-way softmax per token (§4).

The phenomenon. Table 1 previews the tokenizer sweeps: reconstruction MSE improves monotonically in K everywhere, while the used rate $H(q)$ saturates at ≈14 bits (53% utilization at K=32,768) on both CIFAR variants and does not saturate on ImageNet 64×64 (98%). This is the first of two ceilings the paper quantifies; a pre-registered size-sweep (§5.1) shows the ceiling differs by *source* under the studied tokenizer recipe, with dataset size alone ruled out.

3 The generator as a measurable channel

We formalize "the generator makes mistakes" as a stochastic transition between codewords, in four layers: the full sequence-level denoising channel $\mathbf{C} \to \mathbf{X}_t \to \hat{\mathbf{C}}$, never estimated directly; its measured *single-letter projection* $W_t(\hat{c} \mid c)$ with per-token transfer $R_{\text{token}}(t) = I(C_i; \hat{C}_i)$, a function of the probe level; the *information-equivalent vocabulary* $2^{R_{\text{token}}}$ under the empirical source q—not a Shannon capacity; and the contextual–memoryless gap Δ_{ctx}. Since unconditional generation has no observable "intended codeword", the channels below probe the trained denoiser under data-conditioned corruption; their relevance to generation is tested empirically (§5).

3.1 Terminal dispersion, and the snapping phenomenon

The most obvious noise measurement is the *terminal residual*: sample $\hat{z} \sim G$ unconditionally and record, per position, the distance to the nearest codeword. Because near-cell-boundary noise is attributed to the wrong cell, the raw residual underestimates dispersion; we invert a *calibration curve* $r(\sigma)$ simulated on the same codebook, giving $\hat{\sigma}_{\text{res}}$.

On our K=1024 generator $\hat{\sigma}_{\text{res}}$ is $0.061/0.0056/0.0020$ for $10/50/250$ DDIM steps (monotone; all stability criteria in Table 6 pass). Outputs sit *on* the codeword grid (0.2% of codebook scale), and this is not a sampler artifact: the same DDIM code with the analytic denoiser reaches 10^{-4} (Appendix F).

The central measurement lesson: *errors are wrong commitments—selecting the wrong codeword precisely—not dispersion around the right one* (Figure 1a). Under moderate perturbation 43% of tokens land on the wrong codeword while the terminal residual stays at 0.005: dispersion-based estimates are blind to essentially all of the error (both error modes on decoded images: Figure 2). In particular, the naïve closed-form chain "measure $\sigma_{\text{res}} \to$ capacity $C(\sigma_{\text{res}}) \to$ optimal K^*" is non-predictive for diffusion generators (it yields $K^* \approx 10^{12}$). We demote σ_{res} to a diagnostic and measure the channel directly. Snapping is *capacity-conditioned*: the frozen 21M model stops snapping at 128px for $K \geq 16{,}384$, and a same-step 80M restores it (§5.1).

3.2 An operational denoising channel and its single-letter projection

Our primary, assumption-free measurement is the *roundtrip* of Figure 1b: renoise a real test latent $x_0 = e_C$ to time t, denoise with the trained generator, quantize to $\hat{C}$. This realizes the single-letter projection W_t; we estimate the marginal kernel directly (pooled over positions), reporting (i)

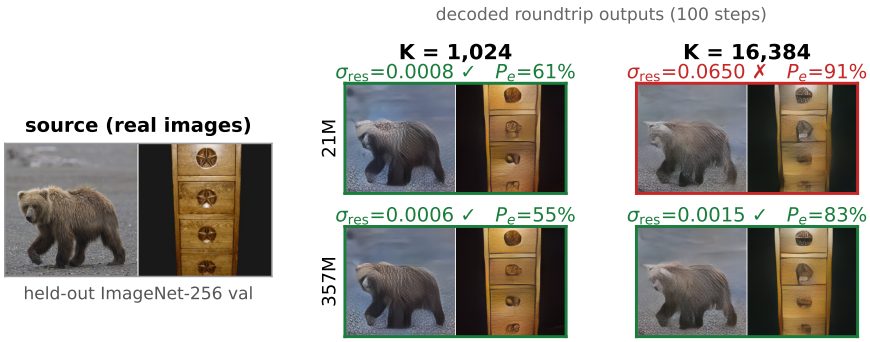


Figure 2: **Roundtrip stress test at t=0.4: dispersion, not wrong commitments, degrades decoded quality.** Held-out ImageNet-256 tokens are noised to t=0.4, denoised (100 steps, the protocol count) and snapped-and-decoded (the gFID pipeline); every panel receives the same source tokens and injected ε (2 of 8 measured sources shown; Appendix P). The panels are the full 2×2 (stage-2 scale × K; E35 perceptual tokenizers, §5.1). Enlarging the codebook triggers the failure (row-wise): the frozen 21M commits at K=1,024 but fails at K=16,384 (σ_{res}=0.065; the unconditional-generation diagnostic of §5.1 reads 0.071—both above the 0.05 bar) and its outputs wash out. Scaling the stage-2 repairs it (column-wise): the 357M commits (σ_{res}=0.0015) and quality is preserved—errors land on near neighbours (A=0.13 of an independent draw), content drifting silently. P_e rises with K at matched noise (55–61% → 83–91%); ✓/× marks commitment, not error rate (t=0.2: Appendix P).

the per-token information transfer $I(C;\hat{C})$ (plug-in entropy with Miller–Madow correction over the full test set; conditioning codewords require ≥ 8 visits, and every estimate carries a half-sample drift diagnostic—Appendix G), (ii) the error rate $P_e = \Pr[\hat{C} \neq C]$, and (iii) the error displacement $A = \mathbb{E}\|e_{\hat{C}} - e_C\|^2/(2\,\mathrm{tr}\,\Sigma_c) \geq 0$ (the pseudo-Gray objective, Zeger & Gersho, 1990, normalized so A=1 is the independent-draw baseline—a reference point, not an upper bound).

3.3 The AWGN surrogate and its validity domain

The surrogate (Figure 1b, green path) replaces the generator by AWGN with nearest-neighbour decoding; its memoryless transfer $R_{\text{eff}}(\sigma)$ is computed by Monte Carlo (a two-pass estimator; Appendix G), and is bounded by the **Gaussian covariance upper bound**—a relaxation, not the Shannon capacity of this finite-constellation, fixed-source channel:

$$R_{\text{eff}}(\sigma) \;\leq\; \min\Big\{H(q),\; C(\sigma)\Big\}, \qquad C(\sigma) = \tfrac{1}{2}\sum_{i=1}^{d} \log_2\!\big(1 + \lambda_i/\sigma^2\big), \tag{1}$$

This bound can be proved by the data-processing inequality and the Gaussian maximal-entropy property (Cover & Thomas, 2006). Both quantities depend only on the codebook and σ—no stage-2 model is needed to evaluate them. A full derivation and its scope are given in Appendix A.

Which σ? When comparing against the roundtrip at level t, the correct noise coordinate is the *injected* noise $\sigma_{\text{inj}}(t) = \sqrt{(1-\bar{\alpha}_t)/\bar{\alpha}_t}$ (rewrite $x_t/\sqrt{\bar{\alpha}_t} = x_0 + \sigma_{\text{inj}}\varepsilon$), *not* the terminal residual: snapping keeps the terminal residual near zero at any error rate, so a residual-based comparison spuriously rejects the surrogate. The validity test on the K=1024 generator is in Table 5; the pre-registered ±15% threshold is generous—observed $t \leq 0.2$ deviations are 0.1–4.9% there and at most 9.1% across all configurations—and exceedances (t=0.4) are reported as failures.

Two readings of that test. First, within $t \leq 0.2$ the surrogate reproduces the roundtrip's information transfer and token error rate to within 0.1–4.9%, and the agreement holds at the kernel level (all twelve statistically-sufficient conditions, $K \leq 2048$): usage-weighted TV 0.05–0.08, JS ≤ 0.025 bits—3–14× below a shuffled-target null—and identical most-likely confusion targets for 100% of codes. Eq. 1 bounds the memoryless R_{eff} only, not the contextual I_{direct} (Appendix A). Second, the deviation is always positive and grows with noise. This is not error but structure: the diffusion model denoises *jointly*, so neighbouring-token correlations let it recover corrupted tokens beyond what memoryless decoding permits—consistent with an error-correcting-code reading of the tokenizer's spatial redundancy. We report Δ_{ctx} empirically (here ≤ 0.8 bits at 8×8; it grows with grid size,

Table 2: Pre-registered predictions (archived before measurement) and outcomes. The predicted R_{eff} is computed *from the codebook alone* at $\sigma_{\text{inj}} = 0.336$ (t=0.2); each measured column comes from an independently trained generator. The generator-measured transfer I_{direct} deviates from the prediction by at most 0.6 bits and saturates near 8 bits while the nominal rate reaches 15; the measured-R_{eff} row is a pipeline reproducibility check ($\leq 3\%$). Non-monotone entries across neighbouring K reflect tokenizer-training stochasticity. †anomalous generator (§5, Table 7). §repeated-corruption estimator (20 rounds, 1.28×10^7 tokens, half-sample drift ≤ 0.16 b).

K	256	512	1024	2048	4096†	8192	16384	32768
predicted R_{eff} [b]	5.82	5.81	6.84	7.44	7.04	7.97	8.10	7.43
measured R_{eff} [b]	5.84	5.78	6.78	7.33	6.84	8.02	8.26	7.54
measured I_{direct} [b]	5.96	6.00	7.06	7.54	7.44	7.86§	8.25§	7.91§
Δ_{ctx} [b]	+0.12	+0.22	+0.28	+0.21	+0.60	−0.16	−0.01	+0.37
$\log_2 K$ [b]	8	9	10	11	12	13	14	15
$H(q)$ [b]	7.80	8.80	9.78	10.75	11.84	12.91	13.80	13.95

§7) and give a position-wise information-theoretic upper decomposition in Appendix B; the pre-registered shuffled-context test causally probes the cross-token contribution where estimable and doubles as a negative control for estimator bias (§5).

3.4 Consequences

With the surrogate validated, classical consequences transfer to tokenizer design. **(i) Rate saturation:** for fixed σ, the Gaussian covariance ceiling in Eq. 1 changes only through the codebook spectrum, whereas R_{eff} depends on the full usage-weighted constellation geometry; empirically R_{eff} saturates (Table 2) while $\log_2 K$ grows linearly—added nominal rate beyond this point is not converted into transferred single-token information. **(ii) Pivot without constants:** on synthetic sources with optimal (k-means) codebooks and a simulated channel, the generation distortion obeys $D_{\text{gen}}(K \to \infty) = D_{\text{rec}}(K^*)$; we verify the predicted log–log slope transition $0 \to 2/d$ of $D_{\text{gen}}/D_{\text{rec}}$ at K^* for three latent dimensions (Appendix I). **(iii) Error displacement: the rise is error-rate-driven.** Unconditional A rises from 0.005 to 0.28 across our sweep, but conditioning on an error occurring (A/P_e) yields 0.03–0.05, essentially flat in K: each individual error lands at $\approx 4\%$ of the independent-draw displacement regardless of codebook size (errors always hit near neighbours in latent space). What grows with K is how *often* errors occur, not how far each one travels.

4 Pre-registered predictions

Quantitative claims about generators are cheap after the fact. We therefore archived the following *before* the corresponding measurements (timestamps and files in Appendix U).

P1′ (numeric). For each codebook in the sweep we computed $R_{\text{eff}}(\sigma_{\text{inj}}{=}0.336)$ from the codebook alone, predicting that the roundtrip I_{direct} of the eight independently trained generators would match these values up to a small positive Δ_{ctx}, and would *saturate* near 7–8.5 bits while $H(q)$ climbs from 7.8 to 14.0 (Table 2).

Diffusion flatness (P2′; the stage-2-formulation claim). Snapping implies the precision error mode is absent ($\sigma_{\text{res}}{=}0.002$ vs. minimum spacing ≈ 0.3), so we predicted, as pre-registered: *gFID of the frozen-specification diffusion stage-2 does not degrade as K grows to* 32,768. Categorical AR stage-2 models must instead resolve a K-way classification per token and are widely reported to degrade (Sun et al., 2024; Shi et al., 2024; Chen et al., 2025); masked models (Chang et al., 2022), the third family in use, are untested here. The outcome (§5, Table 3): confirmed through $K{=}16{,}384$, *partially falsified* at 32,768 on CIFAR and in the 128px capacity-limited condition (§5.1). The outcome refines, post hoc, into *two failure sources, each with its own diagnostic and remedy*. *Stage-2 commitment failure*: the calibrated terminal residual misses the pre-specified snapping bar ($\sigma_{\text{res}} < 0.05$, monotone in steps; mean residual at 250 steps over three seeds; per-K verdicts in Table 7), and scaling the stage-2 restores both snapping and gFID (§5.1). *Codebook geometry*: the codebook-side R_{eff} drops, and retraining the codebook removes the degradation (§5). The CIFAR $K{=}32{,}768$ uptick is of the second kind (it passes the bar at 0.045); the marginal misses at $K{=}8192/16{,}384$ (0.052/0.055) do not degrade gFID, so the bar flags stage-2 bottlenecks within a tokenizer family

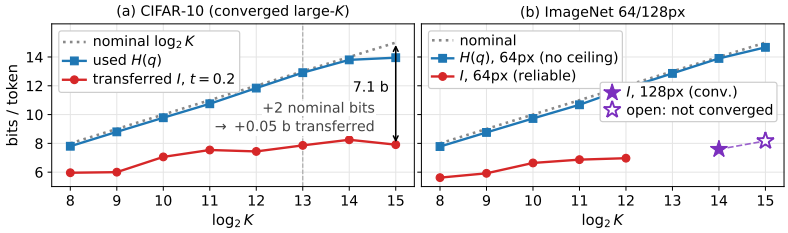


Figure 3: Nominal vs. used vs. transferred rate. (a) CIFAR-10 (converged through K=32,768): $H(q)$ hits a source-associated ceiling; probe transfer saturates near 8 bits, leaving a gap of $\approx$7 b to the 15-bit nominal rate at K=32,768. The dashed line marks K=8,192: enlarging to 32,768 adds $+2$ nominal bits but only $+0.05$ b of I_{direct} at t=0.2 (the surrogate R_{eff} at σ=0.2ρ moves -0.04 b, §5). (b) ImageNet: no ceiling, yet transfer stays far below nominal at every reliably estimated point (open marker: not converged, † in Table 11).

rather than acting as a universal threshold (Appendix Q). Within a family, a configuration with neither signature that still degrades with K would refute the refinement, as would a commitment failure that stage-2 scaling does not rescue. The matched AR comparison is reported with Table 3.

5 Experiments

Setup. CIFAR-10; tokenizers as in Table 1; one frozen-spec 21M-parameter DiT per codebook, trained for 100k steps on the quantized latents; all channels measured per §3. Full hyper-parameters in Appendix E. Models are deliberately small (21M frozen spec; 80M/357M controls)—the mechanism study needs controlled geometry, not SOTA samples. The datasets play distinct roles: CIFAR-10 is the *dense controlled regime* where large-K channel quantities estimate to convergence (this section); ImageNet at 64–256px tests *transfer* beyond it; CIFAR-100 replicates (§5.1).

Protocol criteria. Every pre-fixed stability criterion passed on every generator but the flagged K=4096 one (seed and position CV, step-monotonicity, calibration invertibility, surrogate error; Appendix D).

Rate saturation on real codebooks. At fixed source-scale noise (σ=0.2ρ, milder than the σ_{inj}=0.336 reporting point), K=8192 $\rightarrow$ 32,768 gains $+2$ nominal bits while R_{eff} changes -0.04 b (9.02 $\rightarrow$ 9.05 $\rightarrow$ 8.98): quadrupling the codebook buys nothing the channel can carry.

Saturation (P1′ verification). Table 2 reports the outcome on all eight independently trained generators. Three readings.

(i) Pipeline reproducibility (3%, 8/8). Archived and recomputed R_{eff} agree on every codebook; the generator-dependent tests are (ii), (iii), and gFID.

(ii) Transfer saturates, with converged large-K estimates. I_{direct} climbs 5.96$\rightarrow$7.54, then stays at 7.9–8.3 bits through K=32,768 against a 15-bit nominal rate (Figure 3). Convergence: repeated corruption (1.28×10^7 tokens, drift $\leq$ 0.16 b), bootstrap CIs $\pm$0.017 b, position-pooling check (Appendix G).

(iii) The cross-token component of the model–surrogate gap is real, causal, modest, K-independent. Shuffling reduces the raw gap Δ_{ctx} from $+0.28$ to $+0.05$ b, attributing $+0.23$ b to cross-token context (causal); the probe doubles as a bias control: a $+1.5$-b "gain" at K=32,768 survives shuffling, exposing it as bias. Bias-cancelled: $+0.23$/$+0.19$ b at K=1024/32,768.

Robustness to the probe level t. Across $t \in \{0.05, 0.1, 0.2, 0.4\}$ the codebook ranking is stable wherever estimates are reliable (pairwise Spearman 0.90–1.00; all-K 0.95–1.00 with the single-round large-K bias at $t\neq$0.2). As $t\rightarrow$0 the roundtrip approaches identity and I tracks $H(q)$ (Appendix H); we report t=0.2 (mid-range error rate), and no conclusion changes at any other measured level. In particular the headline claim—added nominal capacity is not converted—holds at *every* probe level, for complementary reasons: at low t the cap is the usage ceiling (13.95 of 15 b), at stressed t transfer saturation (gap $\geq$6.9 b); only the gap's size is t-dependent.

Is the probe the sampler's own channel? No—measurably (Appendix M). Along unconditional trajectories the model revises far fewer provisional commitments than the probe confuses at matched

noise (0.10 vs. 0.43 at t=0.2, K=1024): at moderate probe levels the probe is empirically *harsher* than the trajectory in every evaluated condition; at t=0.05 the ordering can reverse.

A generator-derived operating point. The sampler's commitment schedule is itself measurable: t_{50} orders exactly with capacity (0.39 at K=1024 vs. 0.17 at 16,384; 0.19 for 80M vs. 0.15 for 21M at 128px)—a generator-derived operating point σ_G with no external choice. $R_{\text{eff}}(\sigma_G)$ is a *generator-indexed reference point* on the codebook stress curve: earlier commitment is noisier commitment, so stronger models can show lower $R_{\text{eff}}(\sigma_G)$ (7.6 vs. 8.8 b); cross-model comparisons need full curves.

A dispersion transition in K. σ_{res} sits at the 0.002 calibration floor for $K \leq 1024$ (0.0026 at K=2048) and rises to 0.052/0.055/0.045 at K=8192/16,384/32,768 (Table 7): two codebooks cross the 0.05 bar, though the median residual there (0.028–0.039) stays an order of magnitude below the median codeword spacing, so the P2′ precondition holds for the bulk of tokens while denser codebooks make clean commitment measurably harder. One generator (K=4096, σ_{res}=0.134, non-monotone) violates a protocol criterion; a pre-registered seed-retrain recovers normal behaviour (0.009, monotone)—training stochasticity; the original stays flagged in all tables.

Diffusion flatness (P2′ verification). Table 3 reports gFID (10k samples, DDIM 250, unconditional, specification frozen across K). The prediction holds over the entire range where the received AR view expects degradation: gFID *improves* monotonically from 80.3 (K=256) to a plateau of ≈70 (K=4096–16,384), where AR pipelines are widely reported to turn over (Sun et al., 2024; Shi et al., 2024). At the extreme K=32,768 the prediction *fails* (+3.8 FID, 7.7σ over three seeds); we count P2′ as partially falsified.

Anatomy of the failure. The uptick is *not* precision-driven (σ_{res} actually decreases); it coincides with the usage ceiling ($H(q)$=13.95 bits, 53% utilization). Two facts are consistent with codebook geometry rather than hard utilization alone. The archived *pre-training* flag: this codebook's predicted R_{eff} uniquely dropped, and it later produced the worst uptick (on CIFAR-10; not on CIFAR-100). And an intervention: retraining the K=32,768 tokenizer with an entropy regularizer cannot move the ceiling (52% utilization under direct pressure) yet removes the uptick outright (gFID 67.9, below the original plateau) while the codebook-side stress rate rises (8.44 vs. 7.43 b)—dissociating the uptick from hard utilization (the intervention changes geometry and reconstruction jointly; Appendix Q).

A worked decision stage-1 statistics cannot make. At 128px, K=16,384, $H(q)$ and utilization are healthy and *identical* for both stage-2 models, so they cannot distinguish "codebook too large" from "generator too small"; the roundtrip attributes the failure (σ_{res}=0.053, t_{50}=0.15) and the same-step 80M control confirms *scale stage-2, keep K*; a pre-registered replication at 256px/357M repeats the decision (§5.1). The gFID-level findings stand independently of the channel machinery.

Bounding the search over K from the codebook alone. The codebook-side curve turns saturation into a pre-training rule: train stage-2 models only up to the smallest K whose predicted R_{eff} comes within 0.25 b of the sweep maximum (any tolerance in 0.15–0.5 b gives the same bounds). Applied retrospectively, the rule stops at K=8192 on both CIFAR variants (7.97 vs. 8.10 b; 7.77 vs. 7.82 b) and at 16,384 on ImageNet 64×64, excluding the CIFAR-10 K=32,768 codebook whose predicted R_{eff} had dropped (8.10 → 7.43 b) and which later upticks. The retained range contains the gFID optimum on CIFAR-10 (70.0 at K=4096) and CIFAR-100 (71.4 at 4096; Appendix T), and on ImageNet its best (136.8 at 8192) is within 0.1 of the out-of-range 136.7 (Table 4). The sweep thus shrinks from eight codebooks to five (CIFAR) or seven (ImageNet), at the cost of a codebook-side computation of seconds per K. The rule bounds K from above; the optimum inside the range, and its lower end, remain set by the reconstruction side of the trade.

The AR comparison (pre-registered): allocation, not K. A matched AR transformer on the same latents, frozen 6.3M core, degrades from K=2048 (88.7→102.9)—exactly where diffusion plateaus; both pre-registered criteria hold (Table 3). Vocabulary grows from 2% to 73% of parameters along this row, a capacity confound we flag and then dissect with four controls.

Parameter-matched control: the capacity confound is real. Re-allocating the K=32,768 budget to a wide core at K=1024 reaches gFID 47.5 ± 0.2, beating frozen-core AR (88.9) and diffusion (74.0).

Table 3: P2′ and the AR controls: gFID across K (10k samples; unconditional). Diffusion improves then plateaus with a mild extreme-K uptick (dissected below); frozen-core AR degrades from K=2048; fixed-budget AR (≈23M total) *improves* through 16,384, collapsing only when vocabulary (0.3→16.8M) crowds the core (22.9→6.3M): allocation, not K.

K	256	512	1024	2048	4096	8192	16384	32768
gFID diffusion (frozen spec) ↓	80.3	79.6	74.0	73.2	**70.0**	70.5	70.4	74.3
gFID AR (frozen 6.3M core) ↓	89.1	90.1	88.9	**88.7**	97.1	98.6	104.0	102.9
gFID AR (fixed 23M budget) ↓	59.2	—	47.5±0.2	—	44.0	—	**33.8**±0.1	103.4±0.5

Table 4: ImageNet 64×64 replication (CIFAR-100: Appendix T; codebook-side pipeline reproduces on **24/24** codebooks across all three datasets, max. dev. 3.3%). The extreme-K uptick appears on both CIFAR variants (53% utilization) and *vanishes* on ImageNet (> 90%).

	K	256	512	1024	2048	4096	8192	16384	32768
ImageNet 64×64	pred. R_{eff}	5.47	5.81	6.37	6.64	6.67	7.04	8.16	7.99
	meas. R_{eff}	5.50	5.78	6.50	6.69	6.65	6.99	8.19	7.97
	gFID ↓	148.9	152.7	145.2	141.3	144.5	136.8	140.0	**136.7**
	$H(q)$ [b]	7.78	8.76	9.73	10.68	11.83	12.86	13.90	**14.67**

Fixed-core control. Under a *regularized* recipe (label smoothing, dropout; held-out CE drops from ∼50 to 13–22 b/token), K=32,768 loses by 6.9 FID despite 2.4× parameters, under our pre-registered 10-FID bar (Table 12).

Fixed-budget sweep: our monotonicity prediction fails. gFID 33.8 ± 0.1 over three seeds—the best stage-2 result in this paper—collapsing only at 32,768 (103.4 ± 0.5), where vocabulary leaves a 6.3M core; gain and collapse both exceed 25× the seed variation.

Allocation-response sweep. At fixed K=32,768 and fixed vocabulary, core 6.3→12M takes gFID 102.9→34.1 (further core slowly degrades: 37.5, 39.8)—in our matched setup the collapse is *caused* by core starvation, with an interior optimum.

Together: the categorical large-K "burden" is an *allocation* response (cf. compute-optimal vocabulary scaling, Tao et al., 2024), while diffusion remains comparatively stable through K=16,384 without allocation tuning. The failed CE-gap prediction is reported in Appendix U, its diagnosis in Appendix S.

Memorization control. Across all eight diffusion generators and every AR model, zero of 4096 samples copies a training sequence, and nearest-sequence statistics match held-out baselines: gFID comparisons are unconfounded.

5.1 Cross-dataset replication and the ceiling mechanism

We replicated the entire pipeline on CIFAR-100 and ImageNet-1k at 64×64 (1.27M images (van den Oord et al., 2016; Karras et al., 2022); Table 4) and, below, at 128×128; extra downsampling keeps the grid 8×8, so stage-2 is byte-identical throughout. Predictions were archived per dataset before measurement (Appendix U).

(i) Estimator reproducibility transfers (24/24). Archived vs. recomputed R_{eff} agree within 3.3% on every codebook; I_{direct} stays within 0.6 b of R_{eff} wherever reliably estimated (Appendix J).

(ii) The ceiling is source-associated—our own size hypothesis was falsified. CIFAR-100 (same token budget as CIFAR-10) reproduces the ceiling within 0.2%; ImageNet 64×64 shows none (98% utilization). Our pre-registered data-*quantity* prediction failed: ImageNet tokenizers retrained on 50k/200k images (the CIFAR budget) keep 96–97% utilization. A same-resolution control sharpens this: ImageNet retokenized at 32×32, at CIFAR's resolution and grid, still shows no ceiling (97.6% at K=32,768). With size and resolution both ruled out, the ceiling tracks the *source* under the studied recipe (Zheng et al., 2026's budget mechanism stays orthogonal).

(iii) The uptick co-occurs with the ceiling—and dissociates from it under intervention. The extreme-K uptick (+3.8/+4.5 FID) occurs exactly where the ceiling occurs (both CIFAR variants, 5/5) and is absent in all three ImageNet size conditions (uptick $-5.4/-2.1/-7.8$ FID). The five settings are not independent draws, and ceiling and dead-code pathology are entangled by design; the entropy-regularization intervention separates them (§5, Appendix Q): it removes the uptick without moving the ceiling, pointing to codebook geometry rather than hard utilization.

Resolution and generator capacity (128×128). At ImageNet 128×128 (grid 8×8; Appendix L) snapping replicates at $K \leq 4096$; at $K \geq 16{,}384$ the *frozen* 21M DiT stops snapping (σ_{res}=0.05–0.10) and a same-data, same-step 80M DiT restores it (0.007)—capacity, not resolution. Saturation replicates (7.6 b vs. 14 b nominal at K=16,384, inside the CIFAR-10 band; K=32,768 low-confidence), and the surrogate's domain carries over.

A modern-configuration capacity control (256px, perceptual tokenizer, 357M). On ImageNet-256 (adversarial + VGG-perceptual tokenizer, 16×16 grid; Appendix O) the surrogate's domain survives at both K and both scales ($\leq 9.1\%$ through t=0.2), and transfer stays far below nominal (10.8 vs. 14 b). Snapping splits along the commitment-bottleneck line—perfect at K=1024, failing under the frozen 21M at K=16,384 (0.071), *restored* by a 357M model (0.002; Figure 2, controls in Appendix P)—and the diagnosis is predictive: with gFID measured only afterwards, the 21M penalty (153.0→164.6) shrinks 80% under 357M (125.8→128.1; K=16,384 trails at both scales), replicating the 128px attribution: a stage-2 bottleneck, not a codebook one.

6 Related work

COVQ designs quantizers for known noisy channels (Farvardin, 1990; Farvardin & Vaishampayan, 1991; Meng et al., 2025); pseudo-Gray coding makes likely errors cheap (Zeger & Gersho, 1990); treating a trained generator as a channel with a measured kernel is, to our knowledge, new. RobusTok and related work perturb tokens in training: perturbed FID predicts generation better than reconstruction FID (Qiu et al., 2025a;b); we add the kernel- and information-level treatment (continuous-latent analogue: Zhong et al., 2026). CRT quantifies the reconstruction–generation trade-off via scaling laws (Ramanujan et al., 2024); the Entropy Cliff analyses a data-side budget (Zheng et al., 2026); vocabulary scaling laws locate compute-optimal LLM vocabularies (Tao et al., 2024). These mechanisms are orthogonal to ours. Lookup-free/finite-scalar quantization (Yu et al., 2024; Mentzer et al., 2024) imposes $O(1)$ per-dimension resolution, and utilization training sustains very large codebooks (Shi et al., 2024; Chang et al., 2025). Categorical diffusion places literal token transition kernels inside the model (Austin et al., 2021; Gu et al., 2022); information-theoretic diffusion (Kong et al., 2023) lives along the trajectory, ours between codewords.

7 Scope and limitations

(1) The σ-route needs continuous latents (categorical models: the kernel route, §3.2); AR measurements are teacher-forced (generated-start control: no difference above the floor, Appendix R). (2) Distortion theory is MSE-based (Blau & Michaeli, 2019); adversarial, FSQ and 256px-perceptual controls reproduce the surrogate domain (Appendix O). (3) The ceiling is source-associated, not universal; a pre-registered size-sweep rules out dataset size. (4) Models are 21M at frozen specification, with 80M/357M capacity interventions (SOTA scale open); masked-model stage-2 and data-abundant AR (ours: 50k sequences) are untested, and AR-vs-diffusion contrasts are dependence, not causal attribution (objective, factorization and optimization differ at once); kernels are single-letter, and memoryless conclusions are *grid-indexed*: the corrected contextual gain grows with grid size (+0.2/+0.4/+0.9 b at 64/256/1024 tokens, the last above its band—Appendix N); snapping and the surrogate domain persist at all three grids; sequence-level quantities remain unestimated (neighbourhood MI: Appendix M). (5) R_{token} measures distinguishability, not perceptual quality; evaluation is MSE and 10k-sample gFID (rFID and 50k-sample gFID not run).

8 Conclusion

Treating a trained generator as a measurable denoising channel gives codebook design a validated diagnostic: the stress curve $R_{\text{eff}}(\sigma)$ comes from the codebook alone, in seconds, and shows that only about eight of CIFAR-10's fifteen offered bits survive our stress probe. Where that curve saturates bounds the range of K worth training a stage-2 on, before any stage-2 exists. The reconstruction–generation trade-off is not intrinsic to K: degradations trace to bottlenecks that targeted interventions remove (scaling stage-2; retraining the codebook), and categorical outcomes track allocation.

Reproducibility statement. All code, protocols, and timestamped pre-registrations will be released; predictions were archived before measurement.

AI USE STATEMENT

In this work, we used generative AI tools (a large language model) for translation of author-written text. We have not used generative AI tools for generating synthetic data sets, cleaning or reformatting datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing ingredients for or writing proofs, proposing or refining hypotheses, designing or providing feedback on the research methodology or experiments, implementing the methods, or interpreting results; qualitative and thematic data analysis is not applicable to this work.

Additionally, we used generative AI tools for writing and refactoring visualization and figure-plotting scripts, for mechanically cross-checking reported numbers against archived measurement files, for drafting and editing prose (including figure captions) from author-determined results and conclusions, and for generating critiques limited to the clarity, organization, and presentation of earlier drafts. The AI-assisted visualization code was not used to implement the proposed methods or to design or conduct the experiments.

We have reviewed all AI-assisted work: LLM-generated scripts were verified by re-running them against the archived outputs, every number in the text was traced to its source file by the authors, translated text was checked by the authors against the original, and all AI-assisted prose was read and approved by the authors. We take responsibility for the final content of this work, including text, claims, data, or artifacts produced with the aid of generative AI.

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.

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

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

Yifan Chang, Jie Qin, Limeng Qiao, Xiaofeng Wang, Zheng Zhu, Lin Ma, and Xingang Wang. Scalable training for vector-quantized networks with 100% codebook utilization. *arXiv preprint arXiv:2509.10140*, 2025.

Yanzhe Chen, Huasong Zhong, Yan Li, and Zhenheng Yang. UniCode^2: Cascaded large-scale codebooks for unified multimodal understanding and generation. *arXiv preprint arXiv:2506.20214*, 2025.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

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

Nariman Farvardin. A study of vector quantization for noisy channels. *IEEE Transactions on Information Theory*, 36(4):799–809, 1990.

Nariman Farvardin and Vinay Vaishampayan. On the performance and complexity of channel-optimized vector quantizers. *IEEE Transactions on Information Theory*, 37(1):155–160, 1991.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022.

Xianghao Kong, Rob Brekelmans, and Greg Ver Steeg. Information-theoretic diffusion. In *ICLR*, 2023.

Zian Meng, Qiang Li, Wenqian Tang, Mingdie Yan, and Xiaohu Ge. Channel-aware vector quantization for robust semantic communication on discrete channels. *arXiv preprint arXiv:2510.18604*, 2025.

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

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 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 preprint arXiv:2509.12474*, 2025a.

Kai Qiu, Xiang Li, Jason Kuen, Hao Chen, Xiaohao Xu, Jiuxiang Gu, Yinyi Luo, Bhiksha Raj, Zhe Lin, and Marios Savvides. Robust latent matters: Boosting image generation with sampling error synthesis. *arXiv preprint arXiv:2503.08354*, 2025b.

Vivek Ramanujan, Kushal Tirumala, Armen Aghajanyan, Luke Zettlemoyer, and Ali Farhadi. When worse is better: Navigating the compression-generation tradeoff in visual tokenization. *arXiv preprint arXiv:2412.16326*, 2024.

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

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022.

Fengyuan Shi, Zhuoyan Luo, Yixiao Ge, Yujiu Yang, Ying Shan, and Limin Wang. Scalable image tokenization with index backpropagation quantization. *arXiv preprint arXiv:2412.02692*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

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

Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. Scaling laws with vocabulary: Larger models deserve larger vocabularies. In *NeurIPS*, 2024.

Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *ICML*, 2016.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 2017.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A Ross, and Lu Jiang. Language model beats diffusion: Tokenizer is key to visual generation. In *ICLR*, 2024.

Kenneth Zeger and Allen Gersho. Pseudo-Gray coding. *IEEE Transactions on Communications*, 38 (12):2147–2158, 1990.

Bowen Zheng, Weijian Luo, Guang Yang, Colin Zhang, and Tianyang Hu. Taming the entropy cliff: Variable codebook size quantization for autoregressive visual generation. *arXiv preprint arXiv:2605.06207*, 2026.

Tianxiong Zhong, Xingye Tian, Xuebo Wang, Xin Tao, and Pengfei Wan. Diffusing in the right space: A systematic study of latent diffusability. *arXiv preprint arXiv:2606.03578*, 2026.

A Derivation of the Gaussian covariance upper bound

We give the derivation and scope of Eq. 1 for completeness. Let $C \sim q$ denote a codeword index, let $X = e_C \in \mathbb{R}^d$ be its embedding, let

$$Y = X + Z, \qquad Z \sim \mathcal{N}(0, \sigma^2 I), \tag{2}$$

and let $\hat{C}$ be the nearest-neighbour decoding of Y onto the same codebook. These variables form the Markov chain

$$C \;\longrightarrow\; X \;\longrightarrow\; Y \;\longrightarrow\; \hat{C}. \tag{3}$$

The source entropy gives the first bound directly,

$$I(C; \hat{C}) \le H(C) = H(q). \tag{4}$$

For the learned codebooks used here, distinct indices correspond to distinct embeddings, so the lookup $C \mapsto X$ is injective and $I(C; \hat{C}) = I(X; \hat{C})$. Applying data processing to $X \to Y \to \hat{C}$ therefore gives

$$I(C; \hat{C}) = I(X; \hat{C}) \le I(X; Y). \tag{5}$$

Write $\Sigma_c = \mathrm{Cov}(X)$ for the usage-weighted codebook covariance. Since $Y = X + Z$ with Z independent of X,

$$\mathrm{Cov}(Y) = \Sigma_c + \sigma^2 I, \tag{6}$$

and

$$I(X; Y) = h(Y) - h(Y \mid X) = h(Y) - h(Z). \tag{7}$$

Among all continuous distributions with a fixed covariance, the Gaussian has maximal differential entropy. Therefore

$$h(Y) \le \frac{1}{2} \log_2\!\big((2\pi e)^d \det(\Sigma_c + \sigma^2 I)\big)\,, \tag{8}$$

$$h(Z) = \frac{1}{2} \log_2\!\big((2\pi e)^d \sigma^{2d}\big)\,, \tag{9}$$

which yields

$$I(X; Y) \le \frac{1}{2} \log_2 \det\!\left(I + \frac{\Sigma_c}{\sigma^2}\right) \tag{10}$$

$$= \frac{1}{2} \sum_{i=1}^{d} \log_2\!\left(1 + \frac{\lambda_i}{\sigma^2}\right), \tag{11}$$

where $\{\lambda_i\}_{i=1}^d$ are the eigenvalues of Σ_c. Combining the source-entropy bound with this Gaussian covariance bound gives

$$\boxed{R_{\text{eff}}(\sigma) = I(C; \hat{C}) \le \min\left\{ H(q),\ \frac{1}{2} \sum_{i=1}^{d} \log_2\!\left(1 + \frac{\lambda_i}{\sigma^2}\right) \right\}.} \tag{12}$$

The second term is obtained by relaxing the actual discrete, fixed-source input distribution to the maximum-entropy distribution with the same covariance; it is therefore a Gaussian covariance upper bound, not the Shannon capacity of the finite-codebook channel used in our experiments. Importantly, this bound applies to the memoryless nearest-neighbour surrogate R_{eff}. It is not a theoretical upper bound on the contextual roundtrip I_{direct}, because the latter may decode token i from the full noisy grid $\mathbf{Y}$ rather than from Y_i alone; Appendix B formalizes that distinction.

B Information-theoretic scope of the model–surrogate gap

The memoryless surrogate and the trained roundtrip use different observations. For a token position i, let $Y_i = e_{C_i} + Z_i$ denote its local AWGN observation and let $\mathbf{Y} = (Y_i, Y_{-i})$ denote the full noisy grid. Writing Q for nearest-neighbour quantization onto the codebook, define the position-specific quantities

$$R_{\text{eff}}^{(i)} = I\big(C_i; Q(Y_i)\big), \qquad I_{\text{direct}}^{(i)} = I\big(C_i; \hat{C}_i^{G}\big), \qquad \Delta_{\text{ctx}}^{(i)} = I_{\text{direct}}^{(i)} - R_{\text{eff}}^{(i)}, \tag{13}$$

where $\hat{C}_i^G = g_i(\mathbf{Y})$ is the roundtrip decoder for token i, deterministic (as with DDIM, η=0) or randomized. The decomposition below holds position-wise. The estimates reported in §3.2 pool positions, so we use it as a structural scope statement rather than as a numerical bound on the pooled estimator.

The surrogate decoder sees only Y_i, so by data processing

$$R_{\text{eff}}^{(i)} \leq I(C_i; Y_i). \tag{14}$$

By contrast, the roundtrip decoder can depend on the entire $\mathbf{Y}$, hence

$$I_{\text{direct}}^{(i)} \leq I(C_i; \mathbf{Y}). \tag{15}$$

Therefore the position-wise model–surrogate gap obeys

$$\Delta_{\text{ctx}}^{(i)} \leq I(C_i; \mathbf{Y}) - R_{\text{eff}}^{(i)} \tag{16}$$

$$= \underbrace{I(C_i; Y_i) - R_{\text{eff}}^{(i)}}_{\text{local hard-decision information loss}} + \underbrace{I(C_i; Y_{-i} \mid Y_i)}_{\text{available contextual information}}. \tag{17}$$

The first term is non-negative by Eq. 14 and depends only on the local codebook channel. The second term quantifies information about C_i available outside the local noisy observation. Because the AWGN noise is independent across positions, conditioned on (C_{-i}, Y_i) the remote observations Y_{-i} are independent of C_i. Conditional data processing therefore gives

$$I(C_i; Y_{-i} \mid Y_i) \leq I(C_i; C_{-i} \mid Y_i). \tag{18}$$

Moreover $C_{-i} \to C_i \to Y_i$ is a Markov chain, hence

$$I(C_i; C_{-i} \mid Y_i) = I(C_i; C_{-i}) - I(Y_i; C_{-i}) \tag{19}$$

$$\leq I(C_i; C_{-i}). \tag{20}$$

Combining these relations yields the classical upper decomposition

$$\boxed{\Delta_{\text{ctx}}^{(i)} \leq \left[I(C_i; Y_i) - R_{\text{eff}}^{(i)}\right] + I(C_i; C_{-i}).} \tag{21}$$

Thus excess transfer over the memoryless hard-decision surrogate can arise from two sources: information discarded by local nearest-neighbour quantization, and spatial redundancy in the source that a contextual denoiser can exploit. The first term is computable from the codebook and q alone ($I(C_i; Y_i)$ is the mutual information of a K-component Gaussian mixture, estimable by Monte Carlo); the second requires the source's sequence statistics. We do not estimate the two terms in Eq. 21 separately at large K; the bound is a scope statement rather than a new estimator.

Two empirical controls in the main paper probe the second term. First, shuffling destroys cross-token consistency and reduces the measured gap from $+0.28$ to $+0.05$ bits in the well-estimated $K = 1024$ condition (§5), causally implicating cross-token structure in most of the observed excess. We do not identify the residual shuffled gap exactly with the local hard-decision term in Eq. 21, because the trained denoiser under shuffling need not realize the optimal local decoder. Second, the shuffle-corrected gap increases from approximately $+0.2/+0.4/+0.9$ bits across the $64/256/1024$-token grid controls (Appendix N). This growth is consistent with a larger contextual contribution as more spatial structure becomes available, but it is not implied by Eq. 21: the three tokenizers define different source representations rather than nested observations of one fixed random system. These controls are why we treat the memoryless approximation as *grid-indexed* rather than universally context-free.

C Surrogate validity table

D Protocol criteria

E Generator specification

DiT: 64 tokens, width 256, depth 8, heads 8, adaLN-zero conditioning; v-prediction; cosine $\bar{\alpha}$, T=1000; DDIM (η=0); unconditional; batch 512; AdamW lr 2×10^{-4} cosine; 100k steps; EMA

Table 5: Surrogate validity on the $K{=}1024$ generator. At matched injected noise, the memoryless AWGN channel reproduces the measured $I(C;\hat{C})$ within 5% for $t \leq 0.2$; the deviation is positive and grows with noise—the *model–surrogate transfer gap* $\Delta_{\text{ctx}} = I_{\text{direct}} - R_{\text{eff}}$. Measured in an independent corruption run from Table 2: the 7.11 here vs. 7.06 there is ± 0.05 b Monte-Carlo run-to-run variation.

t	σ_{inj}	$R_{\text{eff}}(\sigma_{\text{inj}})$ [b]	I_{direct} [b]	Δ_{ctx} [b]	P_e (AWGN / direct)	verdict
0.05	0.090	9.77	9.78	+0.01	0.001 / 0.001	pass
0.10	0.169	9.31	9.39	+0.08	0.059 / 0.052	pass
0.20	0.336	6.78	7.11	+0.33	0.479 / 0.434	pass
0.40	0.738	3.67	4.44	+0.77	0.901 / 0.846	fail (+17%)

Table 6: Measurement protocol criteria (thresholds fixed in advance).

criterion	threshold	measured	verdict
$\hat{\sigma}$ seed CV (3 seeds)	$< 5\%$	$\leq 0.9\%$	pass
$\hat{\sigma}$ position CV (64 positions)	$< 30\%$	$\leq 9.5\%$	pass
$\hat{\sigma}$ monotone in DDIM steps	—	0.061/0.0056/0.0020	pass
calibration invertibility (slope)	> 0.2	0.97	pass
surrogate relative error ($t \leq 0.2$)	$< 15\%$	$\leq 4.9\%$	pass

0.9995; fp32. Frozen across all K. Tokenizer: EMA vector quantizer (decay 0.99), commitment $\beta{=}0.25$, dead-code revival every 100 steps. Encoder: conv stem, two stride-2 convolutions ($32 \to 8$ spatial), two residual blocks (GroupNorm/SiLU), $1{\times}1$ projection to $d{=}8$; decoder mirrors with transposed convolutions; width 128 throughout (≈4M parameters). The quantizer runs in fp32 under autocast: bf16 nearest-neighbour assignment corrupts the very error rates under study (measured, not hypothetical).

F Analytic-denoiser control

For a 3-component mixture the posterior $\mathbb{E}[x_0|x_t]$ is available in closed form; running our DDIM implementation with this analytic denoiser yields terminal residual 10^{-4} at every sampling budget, while a deliberately under-parameterized toy model plateaus near 0.25 regardless of steps (Figure 4): terminal precision is a property of the model, not of the sampler or the step count.

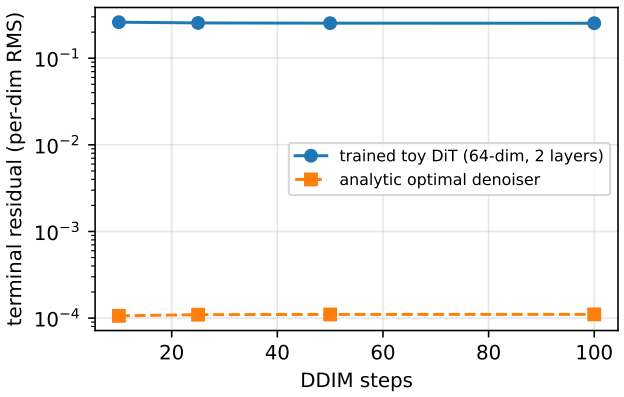


Figure 4: Analytic-denoiser control on a 3-component mixture. The same DDIM code run with the closed-form optimal denoiser reaches 10^{-4} residual; a small trained model plateaus at its own capacity limit. Snapping precision is a model property.

G Two-pass estimator for R_{eff}

Single-pass Monte Carlo estimation of $I(C;\hat{C})$ fails in two opposite ways: the marginal entropy is clipped at $\log_2(\#\text{source samples})$ (underestimating R_{eff} at low noise), and undersampled condi-

Table 7: CIFAR-10 terminal residual per codebook (frozen 21M DiT; E5 protocol run). Rows 1–4 are the calibrated *mean* residual $\hat{\sigma}_{\mathrm{res}}$ at 250 DDIM steps in normalized coordinates—the aggregate on which the $\sigma_{\mathrm{res}} < 0.05$ commitment bar of §4 is adjudicated (row 5 is the across-seed CV). Bold: above the bar. The median aggregate is uniformly smaller ($0.028/0.039/0.029$ at K=8192/16,384/32,768) because the residual is heavy-tailed at large K; it is not the reported convention and the two are not interchangeable. 0.0020 is the calibration-grid floor. For scale, the *median* nearest-codeword spacing over used codewords, in the same normalized coordinates, is $0.61/0.52/0.48$ at those three K while the *minimum* spacing is $0.13/0.058/0.054$: the mean residual is lifted above the bar by a tail living at the scale of the tightest cells, not by the typical token. The step ladder (last two rows, seed-averaged) is the protocol's monotonicity check; this unconditional ladder runs $10/50/250$ steps (the 256px roundtrip's 100 steps is $\lfloor 250t \rfloor$ at t=0.4, Appendix P). †anomalous generator, non-monotone in steps (Table 2); the pre-registered seed-retrain gives 0.0087, monotone (E14).

K	256	512	1024	2048	4096†	8192	16384	32768
seed 0	0.0020	0.0020	0.0020	0.0026	0.1344	0.0519	0.0550	0.0449
seed 1	0.0020	0.0020	0.0020	0.0026	0.1343	0.0518	0.0548	0.0448
seed 2	0.0020	0.0020	0.0020	0.0026	0.1341	0.0515	0.0549	0.0449
mean	0.0020	0.0020	0.0020	0.0026	**0.1343**	**0.0517**	**0.0549**	0.0449
seed CV	0.0%	0.0%	0.0%	1.1%	0.1%	0.4%	0.2%	0.0%
mean, 50 steps	0.0020	0.0036	0.0056	0.0108	0.1298	0.0792	0.0858	0.0738
mean, 10 steps	0.0394	0.0496	0.0610	0.0755	0.1108	0.1146	0.1233	0.1106

tionals underestimate $H(\hat{C}|C)$ (overestimating R_{eff} at high noise, even past the analytic bound). Concretely, write $I(C;\hat{C}) = H(\hat{C}) - H(\hat{C}\,|\,C)$ and estimate each term by plug-in entropies with Miller–Madow correction, $\hat{H} = -\sum_k \hat{p}_k \log_2 \hat{p}_k + \frac{m-1}{2N \ln 2}$, where m is the observed support and N the sample count (the correction removes the leading $O(m/N)$ downward bias of the plug-in estimator). The two terms have opposite failure modes. (i) *Marginal term*: $H(\hat{C})$ can approach $\log_2 K$, so its support must not be truncated by the sampling design; drawing source codewords *with* replacement from q with $N_{\mathrm{marg}} = \max(20K,\, 2 \times 10^5)$ keeps $m \ll N$ and the correction small. A single-pass design that reuses a small set of source codewords clips $H(\hat{C})$ at $\log_2 N_{\mathrm{src}}$ and under-estimates R_{eff} precisely in the low-noise regime (we measured a $12 \to 7.9$ bit error at K=4096 before the fix). (ii) *Conditional term*: $H(\hat{C}\,|\,C{=}c)$ must be estimated from replicates of the *same* source codeword; with R replicates the plug-in support is capped at R, so whenever the true conditional support is comparable to R the term is under-estimated and R_{eff} is inflated—in the extreme, past the analytic bound of Eq. 1. We use R=1024 replicates over 256 codewords and flag any estimate whose observed conditional support exceeds $R/8$. For the data-conditioned roundtrip (where replicates of a fixed codeword are not available), the same failure appears as visit-count starvation at large K; there we require ≥ 8 visits per conditioning codeword, use the full test set, and accompany every estimate with a half-sample drift diagnostic ($|\hat{I}_N - \hat{I}_{N/2}| > 0.2$ bits $\Rightarrow$ flagged unreliable), preferring a reported gap over a silently biased number.

H $R_{\mathrm{token}}(t)$ ACROSS THE PROBE RANGE

Full grid of the corrected roundtrip transfer $I_t(C;\hat{C})$ (bits) on CIFAR-10; ‡ = flagged by the half-sample drift diagnostic (upward-biased upper estimates). Rankings are stable across the range (§5); t=0.2 is a reporting point, not an intrinsic operating point (Table 8).

I SYNTHETIC-SOURCE VALIDATION

Gaussian sources (4×10^5 samples), k-means codebooks (K = 256–32,768), simulated AWGN channel with nearest-neighbour decoding. The prediction: the log–log slope of $D_{\mathrm{gen}}/D_{\mathrm{rec}}$ vs. K is 0 for $K \ll K^*$ and $+2/d$ for $K \gg K^*$. Mid-range slopes (columns "mid") are measured around the centre of the sweep, late-range ("late") at its end; when K^* exceeds the swept range the slope stays near 0, when K^* lies inside it approaches $+2/d$ (slightly above, as D_{rec} has not fully entered its asymptotic regime); Table 9 reports the test.

Table 8: $I_t(C;\hat{C})$ [bits] vs. probe level.

K	t=0.05	t=0.1	t=0.2	t=0.4
256	7.80	7.59	5.96	3.61
512	8.78	8.13	6.00	3.69
1024	9.77	9.37	7.06	4.20
2048	10.74	10.18	7.54	4.54
4096	11.69	10.16	7.44	5.14‡
8192	12.73	11.18	8.35‡	6.48‡
16384	13.48	11.88	9.34‡	7.94‡
32768	13.42	11.64	9.25‡	8.10‡

Table 9: Slope test across 12 conditions.

d	σ	K^*	predicted $2/d$	mid slope	late slope
4	0.15	2.03e+03	0.500	+0.387	+0.443
4	0.25	285	0.500	+0.388	+0.486
4	0.40	51.8	0.500	+0.445	+0.546
4	0.60	14.1	0.500	+0.490	+0.575
8	0.15	10^6	0.250	+0.000	+0.008
8	0.25	6.15e+04	0.250	+0.080	+0.203
8	0.40	2.09e+03	0.250	+0.243	+0.302
8	0.60	161	0.250	+0.240	+0.313
16	0.15	10^{12}	0.125	+0.000	+0.000
16	0.25	10^9	0.125	+0.000	+0.000
16	0.40	10^6	0.125	+0.016	+0.035
16	0.60	7.81e+03	0.125	+0.118	+0.175

J THE 18 RELIABLY-ESTIMATED GENERATOR-SIDE CONDITIONS

Table 10 lists all 18 conditions whose transfer estimate passes the drift diagnostic (CIFAR-10: all eight K via the E16 repeated-corruption estimator; CIFAR-100 and ImageNet 64×64: $K \leq 4096$ under the single-round estimator), with the codebook-side memoryless value, the measured generator-side transfer, and their deviation.

Table 10: Generator-side agreement, all reliably-estimated conditions.

dataset	K	R_{eff} [b]	I_{direct} [b]	Δ [b]
CIFAR-10	256	5.84	5.96	+0.12
CIFAR-10	512	5.78	6.00	+0.22
CIFAR-10	1024	6.78	7.06	+0.28
CIFAR-10	2048	7.33	7.54	+0.21
CIFAR-10	4096	6.84	7.44	+0.60
CIFAR-10	8192	8.02	7.86	-0.16
CIFAR-10	16384	8.26	8.25	-0.01
CIFAR-10	32768	7.54	7.91	+0.37
CIFAR-100	256	5.86	6.01	+0.15
CIFAR-100	512	5.99	6.18	+0.19
CIFAR-100	1024	6.76	6.96	+0.20
CIFAR-100	2048	6.87	7.12	+0.25
CIFAR-100	4096	6.79	7.15	+0.36
ImageNet 64×64	256	5.50	5.62	+0.12
ImageNet 64×64	512	5.78	5.92	+0.14
ImageNet 64×64	1024	6.50	6.64	+0.14
ImageNet 64×64	2048	6.69	6.87	+0.18
ImageNet 64×64	4096	6.65	6.97	+0.32

K EXPERIMENT INDEX

Internal experiment IDs referenced in the text, in reading order.

E5/E5c	CIFAR-10 main sweep (8 codebooks); memorization audit
E8	contextual-gain causal ablation (shuffled probe)
E9	diffusion gFID seed control at $K \in \{4096, 16384, 32768\}$
E10/E16/E17	large-K MI: full-set re-estimate; repeated corruption; bootstrap
E11	kernel-level surrogate comparison (three datasets)
E12	ceiling size-sweep (falsified our size hypothesis)
E13/15/18/19/29	AR allocation: budget, core, sweeps, seeds, fixed-K response
E14	anomalous-run seed-retrain control (K=4096)
E31/E33	grid-size controls (16×16, 32×32)
E32	128px stage-1 completion
E20/E22	ImageNet 128px replication; its converged large-K MI
E23–E25	AR CE diagnosis; generated-start probe; adversarial tokenizer
E26/E26b/E27/E28	trajectory commitment (t_{50}); operating point; neighbourhood MI; FSQ
E34/E35	same-resolution ceiling control; 256px perceptual configuration
E36/E37	utilization intervention; 357M generator control

L 128×128 REPLICATION (E20)

Same protocol as the main experiments with one more downsampling stage (ImageNet-1k, 1.27M images, latent grid 8×8, d=8; frozen 21M DiT, 100k steps as everywhere in this paper). Criteria were pre-registered before measurement (Appendix U). The 80M control (w=512, depth 12) is trained with the *identical* data, latents, schedule, and 100k-step budget; capacity is the only varied factor. Table 11 collects the outcomes.

Table 11: ImageNet 128×128. σ_{res} at 250 denoising steps (source scale); I_{direct} at t=0.2. K=16,384 uses the converged 20-round repeated-corruption estimate (E22: 12.8M tokens, drift 0.17<0.2; single-round 9.04 revised down). Daggered entries do not converge at this sample size (K=32,768: drift 0.28) or are single-round (80M); low confidence. Surrogate column: ±15% test at t=0.05/0.1/0.2.

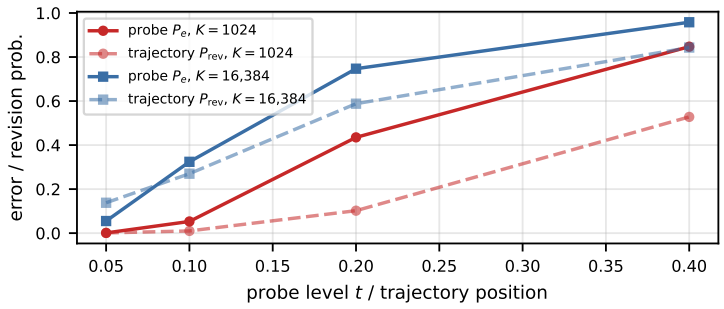
	K=1024	4096	16384	32768	16384 (80M)
$H(q)$ [b] (nominal)	9.73 (10)	11.76 (12)	13.90 (14)	14.68 (15)	13.90 (14)
$\sigma_{\mathrm{res}}(250)$	0.0020	0.0087	0.053	0.102	**0.0070**
monotone in steps	✓	✓	✓	✓	✓
$I_{\mathrm{direct}}(t{=}0.2)$ [b]	6.54	6.99	**7.59**	8.17†	8.93†
surrogate ±15%	✓✓✓	✓✓✓	×✓×	✓✓×	×✓×

Pre-registered readings: (b) saturation replicates with a *converged* estimate at K=16,384: 7.59 b against a 14-b nominal rate—inside the CIFAR-10 saturation band—while K=32,768 does not converge at 12.8M tokens (drift 0.28) and we claim only codebook-side quantities there; (a) snapping replicates fully at $K \leq 4096$ and fails the 0.05 bar at $K \geq 16{,}384$ under the frozen 21M model; (c) at same-data, same-step training, the 80M control restores σ_{res} to 0.007, locating the failure in stage-2 scale (commitment precision) rather than resolution. The 80M surrogate test at t=0.05 fails marginally (15.2% vs. the 15% bar); we report it as a failure.

M THE SAMPLER'S OWN COMMITMENT CHANNEL (E26)

Generation from pure noise has no intended codeword; does the data-conditioned probe describe anything the sampler does? Along unconditional DDIM trajectories (250 steps, 10k samples) we track the *provisional commitment* C_t (nearest codeword of the model's $\hat{x}_0$ prediction at trajectory position t) against the final commitment, giving a revision probability $P_{\mathrm{rev}}(t)$ and a trajectory kernel $W(c_{\mathrm{final}} \mid c_t)$ comparable to the probe kernel at the same t (split-half null floors as in Appendix R).

P_{rev} / P_e (TV excess)	t=0.05	0.1	0.2	t_{50} [σ_G]
CIFAR K=1024	.001/.001 (.00)	.01/.05 (.02)	.10/.43 (.23)	0.387 [0.71]
CIFAR K=16,384	.14/.06 (.09)	.27/.32 (.12)	.59/.75 (.14)	0.172 [0.29]
128px 21M	.17/.10 (.11)	.35/.42 (.14)	.64/.80 (.12)	0.152 [0.25]
128px 80M	.05/.08 (.05)	.19/.36 (.15)	.54/.77 (.15)	0.188 [0.32]



Pre-registered readings. (a) The pointwise-match criterion *fails*: at moderate probe levels ($t \geq 0.1$) the trajectory revises less than the probe confuses in every evaluated condition (ratios 0.2–0.9), while at $t=0.05$ the ordering can reverse (e.g. 0.14 vs. 0.06 at $K=16{,}384$); kernel TV excess is above the 0.05 bar throughout. The probe is therefore a fresh-noise *stress* channel, empirically harsher at moderate levels but not a uniform upper bound, and all transfer numbers in this paper are properties of that stress channel. (b) The commitment schedule delivers what an external probe cannot: t_{50} orders exactly with capacity—perfect snapping commits earliest (0.387), the capacity-limited 21M model at 128px commits latest (0.152), the 80M control recovers (0.188)—so a generator-derived operating point $\sigma_G = \sigma_{\text{inj}}(t_{50})$ exists and requires no external choice. Cross-model I comparisons at these points inherit the large-K estimator caveat of §5. Evaluating the codebook-side rate at each generator's own operating point (E26b) gives $R_{\text{eff}}(\sigma_G) = 3.8$ b ($K=1024$, which commits in deep noise), 9.0/8.8/7.6 b for CIFAR-16,384 and the 21M/80M 128px pair—the operating-point-indexed rate the reviewer program asks for; we report it descriptively.

Neighbourhood-MI control (E27). Does information about C_i survive in *other* positions of $\hat{C}$? With 20-round repeated corruption (12.6M position pairs, $t=0.2$) and a shuffled-source null for bias control, adding the right neighbour to the conditioning alphabet changes the estimate by $+0.09$ b raw (-0.7 b after conservative null subtraction) at $K=256$, where the K^2 joint alphabet is well sampled; at $K=1024$ the null bias (5.0 b) dominates and we report the condition as not estimable.

N GRID-SIZE CONTROLS (E31, E33)

Is the small model–surrogate gap an artifact of the 8×8 grid? Retokenizing CIFAR-10 to 16×16 (256 tokens) and 32×32 (1024 tokens) at $K=1024$ and rerunning the full pipeline:

tokens	64 (8×8)	256 (E31)	1024 (E33)
Δ_{ctx} (shuffle-corrected)	$+0.2$ b	$+0.40$ b	$+0.89$ b
pre-registered band	—	≤ 0.5 ✓	≤ 0.8 ×
$\sigma_{\text{res}}(250)$	0.002	0.002	0.006
surrogate domain	$t \leq 0.2$	$t \leq 0.2$	$t \leq 0.2$
$I(t=0.2)$	7.1 b	6.5 b	5.9 b

Snapping and the surrogate's validity domain persist at all three grids, and the 16×16 gain fell inside its pre-registered band; the 32×32 gain ($+0.89$ b) exceeds its pre-registered ≤ 0.8 b band and we report it as such. The corrected gain roughly doubles per $4\times$ tokens—so the memoryless approximation is *grid-indexed*: tight at 8×8, good at 256 tokens, and quantitatively degrading by 1024 tokens. Quantitative memoryless claims in this paper are therefore per-grid statements; extrapolation to larger grids (e.g. modern 16×16–32×32 latents at higher resolution) should assume a growing contextual term. No sequence-level growth is detectable at $K=256$ (E27); the large-K regime ($K \geq 8192$) remains unresolved, consistent with limitation (5).

O MODERN-TOKENIZER CONTROLS (E25, E28)

Lookup-free FSQ (E28). Finite scalar quantization (Mentzer et al., 2024) replaces the learned codebook by a fixed grid (no collapse, no lookup); the codebook of the analysis is the enumerated grid. With levels $(4,4,4,4,2,2)/(8,8,8,8,4)/(8,8,8,8,4,4)$, i.e. $K=1024/16{,}384/65{,}536$, and the standard frozen DiT: $\sigma_{\text{res}}(250)=0.002$ at all three sizes (monotone in steps)—snapping is not a

VQ-EMA artifact and persists to K=65,536. The surrogate's validity domain shrinks with K as in VQ—all probe levels pass at K=1024, $t\leq$0.2 at 16,384, $t\leq$0.1 at 65,536 (t=0.2 fails marginally, 16.1%)—but the probe retains resolution inside the domain even at the largest K (5.1% error at t=0.1). Transferred information remains far below nominal ($5.2/7.8/9.8^{\dagger}$ b vs. $10/14/16$; the large-K entry is single-round, low-confidence), and FSQ shows its own usage ceiling (12–44% utilization). Perceptual-loss variants remain untested. Mainstream tokenizers train with adversarial (and perceptual) losses. We retrained the K=1024 CIFAR-10 tokenizer with a PatchGAN hinge loss (weight 0.1 after a 2-epoch warm-up; no perceptual term) and re-ran the full pipeline with the standard frozen DiT. Reconstruction trades as expected (MSE 0.0226 vs. 0.0099; full usage, $H(q)$=9.79 b). Snapping is unchanged: $\sigma_{\text{res}}(250)$=0.002, monotone in steps—identical to the MSE tokenizer. The surrogate's validity domain is also identical (pass at t=0.05/0.1/0.2 with relative error 0.0/0.6/3.8%; fail at t=0.4), with $I(t$=0.2)=7.07 b vs. 7.06 for the MSE tokenizer. Snapping and the surrogate are insensitive to the decoder's training objective in this control.

Perceptual configuration at 256px (E35). A VQGAN-style tokenizer (VGG-feature perceptual loss plus PatchGAN adversarial loss; trained on a 400k random subset) on ImageNet-256 with a 16×16 grid, $K \in \{1024, 16{,}384\}$, standard frozen DiT: full codebook usage at both K ($H(q)$=9.58/13.92 b); surrogate within 0.1/0.3/3.0% (K=1024) and 0.1/1.5/3.6% (K=16,384) at t=0.05/0.1/0.2, failing at t=0.4 as elsewhere; $I(t$=0.2)=8.41/10.17 b against 10/14 nominal; shuffle-corrected contextual gain +0.41/+0.31 b. $\sigma_{\text{res}}(250)$=0.002 at K=1024 and 0.071 (monotone) at K=16,384—the same frozen-21M commitment bottleneck seen at 128px.

357M control (E37). A w768×16 DiT (bf16 training, otherwise identical protocol) on the same latents restores snapping at both K ($\sigma_{\text{res}}(250)$=0.002, monotone), keeps the surrogate domain (0.1/0.4/4.5% and 0.3/1.3/9.1% at t=0.05/0.1/0.2; t=0.4 fails as elsewhere), lifts transfer moderately ($I(t$=0.2)=8.53/10.79 b, still far below the 14-b nominal), and t_{50}=0.400/0.335 orders with K as on CIFAR. gFID (10k samples; predictions archived before measurement): 153.0/164.6 (21M) vs. 125.8/128.1 (357M) at K=1024/16,384—the commitment-diagnosed 21M penalty at K=16,384 (+11.6) shrinks to +2.3, closing the second worked decision. Single seed at both scales; the CIFAR seed deviation was 0.5 FID at 10k samples, and ImageNet-256 seed variance was not measured. Whether the extra K is worth keeping depends on the reconstruction–generation trade the practitioner wants; the diagnosis settles only *which* stage is the bottleneck.

P DECODED ROUNDTRIPS AT 256PX (E35/E37)

Controls for Figure 2. Setup shared by all panels: the E35 perceptual/adversarial tokenizers (ImageNet-256, $K\in\{1{,}024, 16{,}384\}$ in Figure 2, K=16,384 throughout this appendix; d=8, 16×16 grid) and the 21M (w256×8) and 357M (w768×16) DiT checkpoints of E35/E37, reused as trained; roundtrip protocol as in §3.2 (held-out tokens → noise to t → DDIM denoise → nearest-codeword snap → decode). σ_{res} and P_e are measured on the eight-source batch (subsets of which are shown). At K=1,024 the batch residuals (0.0008/0.0006, 21M/357M) match the n=2,000 roundtrip protocol (0.0012/0.0008); at K=16,384 the residual is heavy-tailed and RMS- vs. mean-aggregated values differ, so batch (mean) values are quoted throughout, while the *unconditional-generation* diagnostic of §5.1—a different probe—reads calibrated $\sigma_{\text{res}}(250\ \text{steps})$=0.071/0.002 (21M/357M, n=1024). Protocol P_e: 62/57/89/83% (batch: 61/55/91/83%). The commitment bar is the pre-registered $\sigma_{\text{res}} < 0.05$. Figures 5 and 6 vary only the DDIM step count at t=0.4 and t=0.2; Figure 7 shows the committed high-P_e regime; Figure 8 maps the benignity boundary under controlled on-lattice injection.

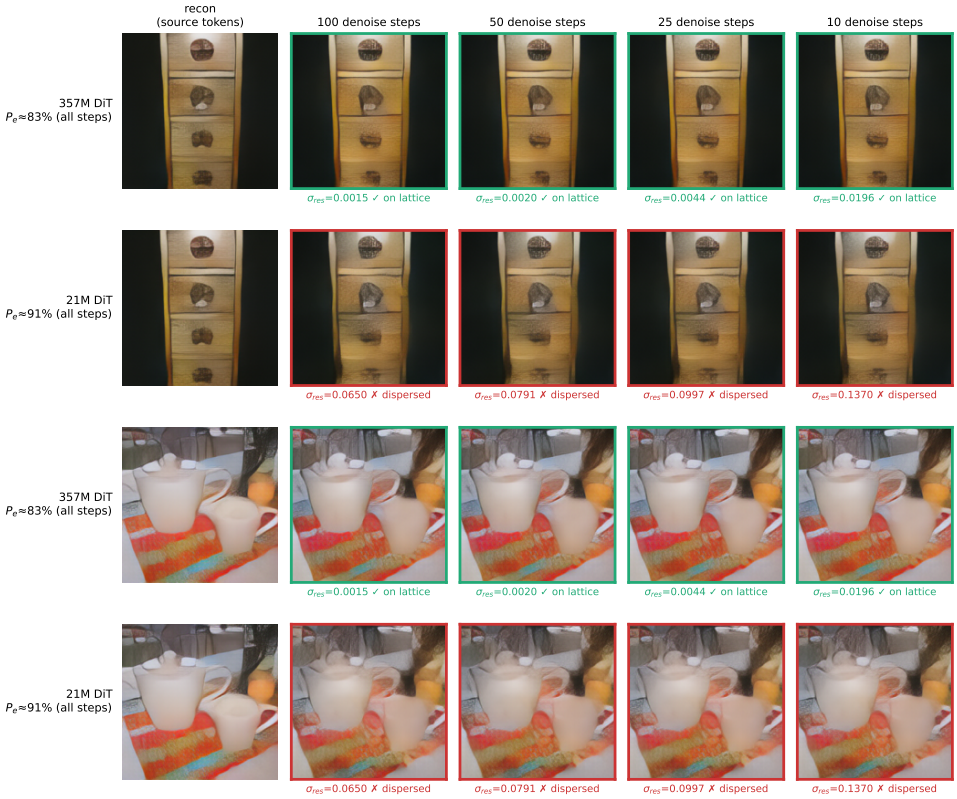


Figure 5: **Dose–response of terminal residual (roundtrip t=0.4).** Same source tokens and injected ε; only the number of DDIM denoise steps varies (100/50/25/10, left to right). Within each row the token error rate P_e is constant, ruling out wrong commitments as the cause. The 357M model (green) commits onto the lattice at every step count ($\sigma_{\text{res}} \leq 0.02$) and its outputs remain intact, ruling out step-count discretisation as the cause. The 21M model (red) never commits (σ_{res} rising $0.065 \rightarrow 0.137$) and degrades monotonically with the residual. Quality is predicted by σ_{res}, not by model identity or step count alone.

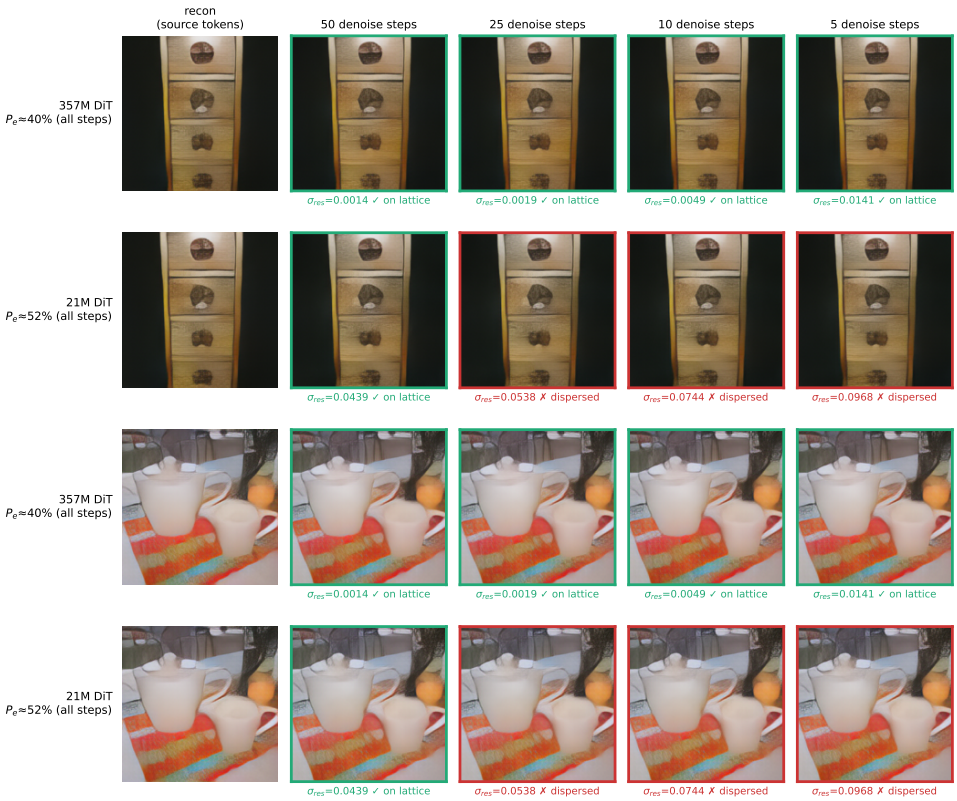


Figure 6: **Dose–response at the milder stress level (t=0.2).** Same protocol as the t=0.4 sweep with steps 50/25/10/5. At the protocol step count (50) the 21M model sits just under the commitment bar (σ_{res}=0.044) and decodes cleanly; reducing steps pushes it over the bar ($0.054/0.074/0.097$) and quality deteriorates accordingly, while P_e stays flat at 51–52%. The 357M control remains committed throughout ($\sigma_{\text{res}} \leq 0.015$). The residual–quality relationship is not specific to one noise level.

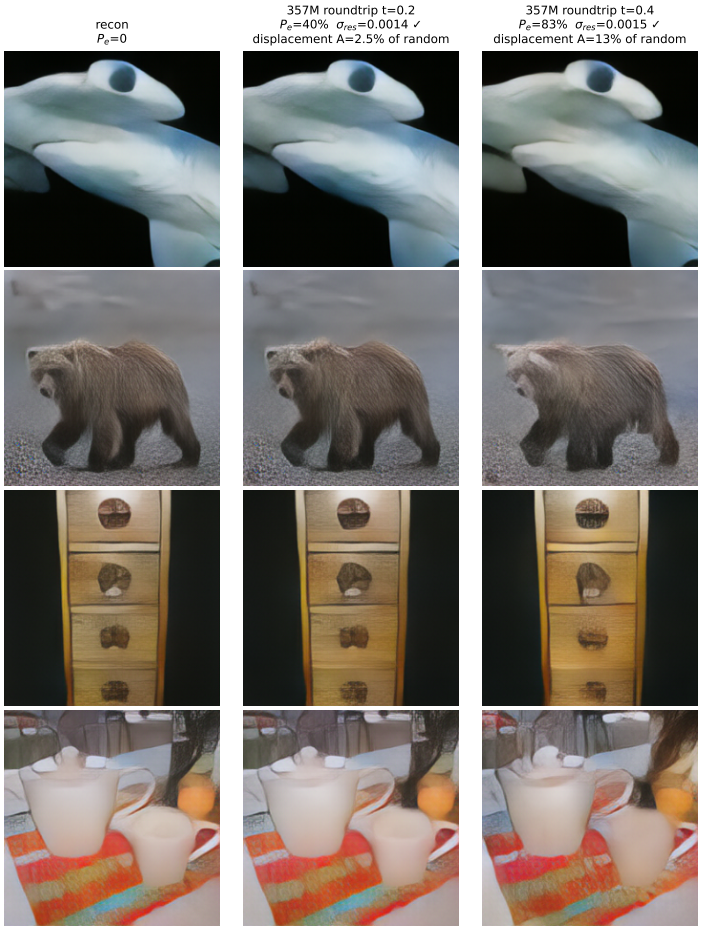


Figure 7: **Perfectly snapped, mostly wrong, still fine.** Roundtrip outputs of the 357M model decode cleanly even when 40% (t=0.2) or 83% (t=0.4) of tokens are assigned to the wrong codeword, because the terminal state commits exactly onto the lattice ($\sigma_{\text{res}} \approx 0.0015$) and the errors have small displacement: the mean embedding-space error is only 2.5% / 13% of that of an independent random draw (A-factor, measured on 2,000 images). Commitment errors trade content fidelity, not perceptual quality.

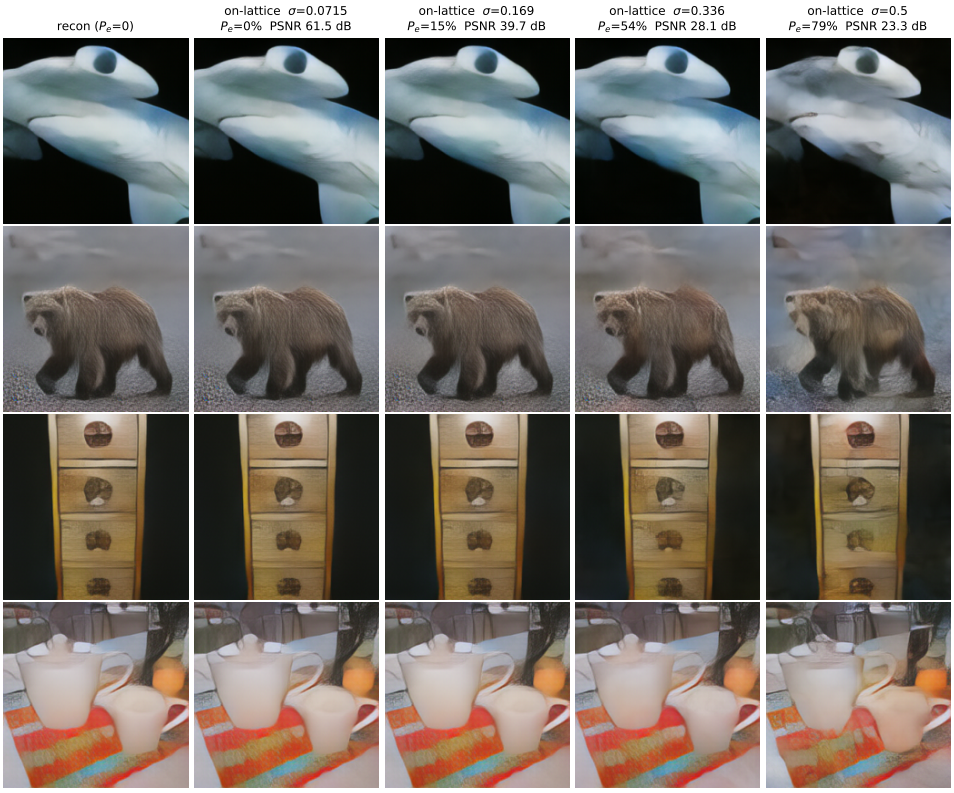


Figure 8: **How wrong can the picks get?** Controlled probe: held-out tokens are perturbed in the normalised latent space (σ_{inj} increasing left to right) and re-snapped to the nearest codeword before decoding, so all error mass is on-lattice. With 15% and even 54% of tokens wrong (at the σ_{inj} of the real channel at t=0.2), images stay sharp and plausible; visible structural damage only appears at 79% wrong tokens (σ_{inj}=0.5, beyond what the measured generator channel produces inside the surrogate's validity domain, $t \leq 0.2$), where the error displacement has grown to A=0.11. Injection here is *independent* per token; the generator's own errors stay benign at comparable P_e and A (Figure 7), so on-lattice error rate and displacement alone do not determine damage. Benignity is a property of metric-local errors, with a measurable boundary.

Q UTILIZATION INTERVENTION (E36)

Pre-registered probe of the usage ceiling at K=32,768 (CIFAR): retrain the tokenizer with a batch soft-assignment entropy penalty (KL to uniform, λ=0.25, temperature = batch-mean nearest distance), then the standard frozen DiT. The regularizer drives the *soft* assignment near uniform (14.9 of 15 b) yet hard utilization does not move (52.0% vs. 52.9%; $H(q)$=13.96 vs. 13.95 b): the pre-registered $\geq$ 90% target was not reached, and we report the outcome as the ceiling resisting direct entropy pressure—the strongest source-association evidence (a third training objective reproducing the same ceiling). Not pre-registered, reported as an observation: at unchanged usage the uptick vanishes (gFID 67.9 vs. 74.3; original plateau 70.0) while the codebook-side stress rate rises ($R_{\text{eff}}(t{=}0.2)$: 8.44 vs. 7.43 b) and reconstruction worsens (0.0099 vs. 0.0081 MSE)—implicating codebook geometry rather than hard utilization alone; the intervention changes geometry, R_{eff} and reconstruction jointly, so no single factor is isolated. Caveats: a single intervention with one stage-2 seed (the 6.4-FID margin is $\sim$13$\times$ the seed deviation); the new generator's σ_{res} (0.14, against used-code spacing comparable to the original family's) exceeds the 0.05 bar while producing the best frozen-spec gFID, so the absolute commitment bar is family-internal, as noted in §4.

R GENERATED-START PROBE CONTROL (E24)

Generation has no observable intended codeword, so our roundtrip probe is data-conditioned. As a control, we re-ran the probe on sequences sampled by the generator itself (pure-noise DDIM, snapped to nearest codewords; 10k sequences, t=0.2). Kernel level: with a same-probe split-half null (the TV between two independent estimates of the *same* kernel), cross-probe TV minus null is -0.007 at K=1024 (0.026 vs. 0.033) and -0.070 at K=16,384 (0.311 vs. 0.376/0.388): no detectable difference above the estimation floor at either K—absolute TV values at large K are floor-dominated and not quotable as "small" on their own. The pre-registered TV bar (0.10), set without the null control, is failed by the raw large-K number and passed by the floor-adjusted one; we report both. Information level: $I_{\text{gen}} - I_{\text{data}} = -0.16$ b at K=1024 (within the 0.3 bar) and -0.35 b at K=16,384 (marginally over, with the large-K single-round bias caveat of §5).

S AR OVERCONFIDENCE DIAGNOSIS (E23)

Teacher-forced CE on training vs. held-out sequences for the fixed-budget AR models (bits/token): K=1024 (w=480): 0.24 vs. 43.8; K=16,384 (w=352): 0.25 vs. 56.0; K=32,768 (w=256): 0.63 vs. 50.1. Training CE near zero identifies sequence-level overfitting (50k sequences, $\approx$1000 epochs at 100k steps): held-out teacher-forced CE measures overconfidence on contexts the model never visits when sampling from its own distribution, which is why gFID and the (clean) memorization audit are unaffected. Under the regularized recipe (E15) held-out CE drops to 13–22 b and the K-ordering persists (Table 12).

Table 12: Regularized AR control (E15): label smoothing 0.1, dropout 0.1, fixed 22.2M core, 50k steps, matched budget. The recipe *reduces* the CE pathology (held-out CE remains well above $H(q)$); the K-ordering persists in both CE and gFID—direction preserved but below the pre-registered 10-FID bar, so we state it directionally only.

	K=1024 (23.2M total)	K=32,768 (55.7M total)
held-out teacher-forced CE [b/token]	12.85	22.09
gFID ↓	89.8	96.7

T CIFAR-100 REPLICATION

Same protocol as CIFAR-10; reconstruction MSE improves monotonically ($14.2 \to 8.5 \times 10^{-3}$), the ceiling (53% utilization, within 0.2% of CIFAR-10's $H(q)$) and the extreme-K uptick both replicate.

K	256	512	1024	2048	4096	8192	16384	32768
$H(q)$ [b]	7.81	8.79	9.77	10.77	11.85	12.92	13.80	13.97
used (%)	100	100	100	100	100	100	92	53
pred. R_{eff}	5.84	6.02	6.90	6.79	6.75	7.77	7.56	7.82
meas. R_{eff}	5.86	5.99	6.76	6.87	6.79	7.65	7.31	7.86
gFID ↓	80.6	80.1	77.5	76.3	**71.4**	72.8	72.3	75.9

U PRE-REGISTRATION RECORDS

The five falsified predictions counted in the abstract: (1) P2′ at extreme K (partial); (2) the data-quantity explanation of the usage ceiling (E12); (3) the fixed-budget AR monotonicity prediction (E18); (4) the CE-gap monotonicity prediction (E6-P4); (5) the bounded-contextual-gain prediction at 1024 tokens (E33; marginal, $+0.89$ b vs. a $\leq$ 0.8 b band). The prediction file was archived before any roundtrip measurement of the sweep generators and before any gFID computation, at $\sigma_{\text{inj}} = 0.3357$; its contents are reproduced verbatim in Table 13 (timestamped files accompany the code release):

Table 13: Archived numerical predictions, reproduced verbatim.

K	R_{eff} pred. [b]	$C(\sigma)$ [b]	$H(q)$ [b]	P_e pred.
256	5.82	7.59	7.8	0.347
512	5.81	7.09	8.8	0.535
1024	6.84	8.23	9.78	0.478
2048	7.44	8.6	10.75	0.541
4096	7.04	7.83	11.84	0.755
8192	7.97	8.89	12.91	0.719
16384	8.1	9.34	13.8	0.742
32768	7.43	8.82	13.95	0.794

The directional predictions (P2′: no diffusion degradation with K; E6-P3: AR uptick > 3.8 FID with optimum at $K \leq 4096$) and the failed secondary prediction (E6-P4, cross-entropy gap monotonicity) are quoted in the main text with their outcomes; the full experiment log with per-experiment timestamps will be released with the code.