RAE-VAR: COARSE-TO-FINE GENERATION AS CONSTRAINED ALLOCATION IN REPRESENTATION SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation autoencoders provide semantic features for image generation, but how should these features be organized for coarse-to-fine prediction? Much of their decodable content can be retained in a low-rank channel subspace, yet placing that information in the earliest tokens can make those tokens difficult to predict. We introduce RAE-VAR, guided by *constrained allocation*: a design principle, rather than an explicitly optimized objective, for making early scales informative without asking them to predict too much from a short context. Its staircase tokenizer grows spatial resolution and channel rank together, using a nested PCA basis and Fourier resampling while leaving the VAR transformer unchanged. On ImageNet 256×256, the 310M-parameter, 16-layer model reaches gFID $\mathbf{2.98}$ in 43 epochs, surpassing released VAR's 3.15 after 200 epochs with $4.65\times$ fewer generator epochs. At 200 epochs, RAE-VAR improves to $\mathbf{2.58}$ with nine next-scale sampling steps and 655 tokens. Larger 20- and 24-layer models approach their released VAR counterparts with roughly one fifth of the generator training epochs. Controlled ablations support both the structured basis and Fourier resampling, beyond what final reconstruction quality alone explains. Linear probes further show that early prefixes contain class information before their decoded images become recognizable.

1 INTRODUCTION

Visual autoregressive modeling (VAR) (Tian et al., 2024) generates images by predicting entire token scales in parallel, reducing hundreds of raster-order predictions to ten coarse-to-fine steps. Its token hierarchy, however, is built in a pixel-reconstruction latent space. Representation alignment (Yu et al., 2025) and representation autoencoders (RAEs) (Zheng et al., 2026) show that pretrained visual features can improve generative learning. This raises a structural question for next-scale prediction: *what should coarse-to-fine mean in representation space?*

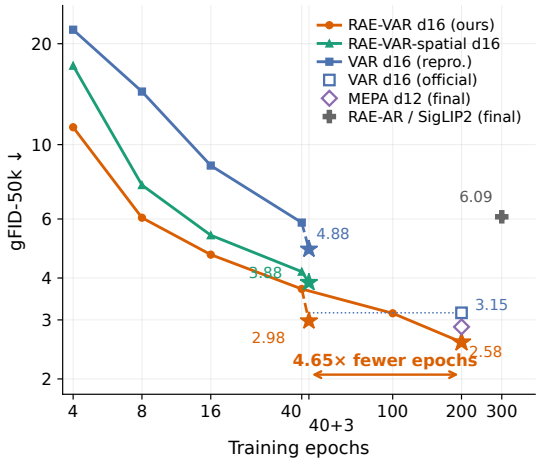

Figure 1: **Faster learning on ImageNet-256.** RAE-VAR beats the official VAR-d16 final with $4.65\times$ fewer generator epochs. Stars mark final checkpoints and 40+3 cooldowns; MEPA and RAE-AR provide reported context.

The distinction matters because the tokenizer determines both what a scale must predict and what its successors can condition on. A hierarchy chosen only for pixel reconstruction may distribute information differently from one suited to semantic features. Reusing a pretrained encoder therefore does not by itself specify a useful code. Our focus is this interface: how to organize a representation into a short sequence of discrete predictions that remains compatible with an existing next-scale model.

Our argument has three steps. First, representation features retain much of their decodable content in a low-rank channel subspace, even when spatial downsampling removes too much information.

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.

Second, exposing that content too early can make autoregressive prediction difficult. Third, spatial resolution and channel rank should therefore grow together, so that early scales are informative but still predictable from a short prefix.

Replacing the VQGAN (Esser et al., 2021) tokenizer with a representation-based tokenizer of comparable trainable size already improves generation, while keeping the VAR-d16 transformer and spatial schedule unchanged. We call this baseline **RAE-spatial-VAR**. Our truncation experiments suggest room for further improvement: a low-rank channel subspace preserves decoded content better than spatial downsampling at the same scalar budget, motivating a hierarchy that also exploits channel structure.

To build on this gain, we consider not only what information each scale preserves, but also the context available to predict it. Our rank-only experiments support this distinction: providing the ground-truth first rank band substantially improves subsequent generation (Appendix C.4). We call this design principle **constrained allocation**, rather than an explicitly optimized objective. Our **staircase tokenizer** grows spatial resolution and channel rank together, introducing additional channel information as the conditioning context grows and further improving generation with the same transformer backbone.

Our headline result is generation with a much shorter training schedule: the 310M-parameter, 16-layer RAE-VAR reaches gFID $\mathbf{2.98}$ after 43 epochs, surpassing released VAR-d16's 3.15 after 200 under the same evaluation pipeline (Fig. 1). Longer training and larger generators provide complementary evidence, reported in Section 4.

Our contributions are threefold:

- **Representation-space next-scale generation.** We establish a controlled RAE-spatial-VAR baseline and improve it with RAE-VAR, reaching gFID 2.98 at 43 epochs and 2.58 at 200 epochs with the unchanged 310M-parameter VAR-d16 backbone.
- **Coarse-to-fine as constrained allocation.** We identify the tension between decodable content and autoregressive predictability, and instantiate the principle with a drop-in staircase tokenizer that jointly allocates spatial scale and channel rank using a nested PCA basis and Fourier resampling.
- **Evidence beyond reconstruction quality.** Ablations at epochs 8 and 16 consistently favor the structured PCA–Fourier design. Prefix probes show that early tokens already contain linearly accessible class information before their decoded images become recognizable, supporting token allocation as a design choice beyond final reconstruction quality alone.

2 Related Work

2.1 Visual autoregressive modeling

Next-token versus next-scale prediction. VQ-VAE (van den Oord et al., 2017) and VQ-GAN (Esser et al., 2021) enable two-stage image generation by compressing pixels into discrete codes that an autoregressive transformer models. Raster-order models such as LlamaGen (Sun et al., 2024) scale this next-token formulation, but require a sequential prediction for each token. VAR (Tian et al., 2024) instead uses a multi-scale residual tokenizer and predicts each scale in parallel under block-causal attention, reducing generation to roughly ten next-scale steps. Within this next-scale formulation, token prediction can be refined through bitwise codes (Han et al., 2025) or hybrid residual diffusion heads (Tang et al., 2025). Other work changes scale schedules (Jiao et al., 2025), conditioning structure (Zhang et al., 2026), or user control (Zhang et al., 2024). These directions modify how a hierarchy is modeled or used; our focus is which information the hierarchy assigns to each scale. Our work is complementary: we retain the VAR and next-scale factorization, but change the representation and the allocation of information within its token hierarchy.

A parallel line avoids vector quantization through continuous autoregression or masked prediction (Tschannen et al., 2024; Li et al., 2024; Ren et al., 2025; Shao et al., 2025). MAR (Li et al., 2024), for example, combines masked autoregressive conditioning with a diffusion loss for continuous tokens. These results establish that discretization is not required for strong image generation, but do not determine which interface is effective for next-scale prediction on high-dimensional representation features. Our continuous next-scale trials underperform at the tested budget (Appendix C.5), motivating the discrete interface studied here rather than a claim against continuous generation.

Alignment versus tokenizer restructuring. MEPA (Zhou et al., 2026) is closest in its use of pretrained semantics within VAR: it aligns aggregated intermediate-scale features to a visual encoder while retaining VQGAN tokens, and combines this with a mixture-of-experts architecture. We instead change the latent space itself and restructure its residual code. This distinction separates semantic supervision of the generator from semantic structure in its inputs; RAE-VAR requires neither an auxiliary generator-alignment loss nor a change to the transformer backbone.

2.2 Representation autoencoders and quantized representations

Pixel and reconstruction latents versus representation latents. REPA (Yu et al., 2025) shows that aligning diffusion-transformer features with a pretrained encoder improves generative learning, while REPA-E (Leng et al., 2025) extends representation-guided training to the autoencoder. RAE (Zheng et al., 2026) takes a structural approach: a pretrained representation encoder and a learned decoder define the latent space in which diffusion operates. In contrast, JiT (Li & He, 2026) directly denoises clean-image predictions in pixel space, without a tokenizer or pretrained representation. We study a different operating regime: few-step, next-scale autoregression in representation space, where the design of a discrete multi-scale interface becomes central.

RAE-AR (Yu et al., 2026) brings representation encoders to continuous raster-order autoregression and addresses token statistics and rollout instability through normalization and noise injection. Our approach changes the organization of the code rather than only its statistics: nested rank prefixes are distributed across spatial scales and modeled with next-scale prediction. The comparison therefore concerns distinct factorizations and token interfaces, not merely alternative encoders.

Discrete representations and their hierarchy. Quantized representation models address related challenges through dominant-subspace quantization (Jia et al., 2025), joint understanding and generation (Du et al., 2026), discrete diffusion (Wang et al., 2026), semantic regularization (Xiong et al., 2025), and region-adaptive codes (Zheng et al., 2025). SimVQ (Zhu et al., 2025), FSQ (Mentzer et al., 2024), and BSQ (Zhao et al., 2025) provide alternative quantization mechanisms. These methods motivate discrete semantic representations. Our contribution is the joint allocation of spatial scale and channel rank within a residual code. Matryoshka-style nesting (Kusupati et al., 2022) and band-limited scale decomposition (Burt & Adelson, 1983) supply its building blocks, while our ablations test whether the resulting prefixes support autoregressive generation.

3 Method

3.1 Next-scale prediction in representation space

VAR (Tian et al., 2024) encodes a latent map f into residual token maps at increasing spatial sides s_k. Starting with $e_1 = f$, each step quantizes a downsampled residual and returns its contribution to the full-resolution grid:

$$q_k = \Phi_k(\mathrm{up}_H(\mathcal{Q}(\mathrm{down}_{s_k}(e_k)))), \qquad e_{k+1} = e_k - q_k. \tag{1}$$

Here $\mathcal{Q}$ is codebook lookup and Φ_k refines the quantized contribution. A block-causal transformer learns $P(\mathbf{s}_k \mid \mathbf{s}_{<k}, y)$; decoding uses the cumulative latent $\hat{f}_k = \sum_{j \le k} q_j$.

We replace the VQGAN tokenizer with a DINOv2-S encoder (Oquab et al., 2024) and ViT-B decoder following RAE (Zheng et al., 2026). The encoder is lightly adapted under an L_2 anchor to its frozen pretrained copy, and a projection maps its 384 channels to a 256-channel quantization space. The tokenizer is then frozen before generator training. Using the original spatial schedule and pooled resampling gives **RAE-spatial-VAR**, our representation-space baseline. Adaptation losses and unsuccessful frozen-encoder and continuous-prediction alternatives are detailed in Appendices A.3 and C.5.

3.2 Coarse-to-fine as constrained allocation

Channel rank preserves content efficiently. We first ask how much information can be removed from a representation while still decoding a useful image. On a frozen DINOv2-B RAE, we truncate channel rank, Fourier frequency, or spatial resolution at matched retained-scalar budgets.

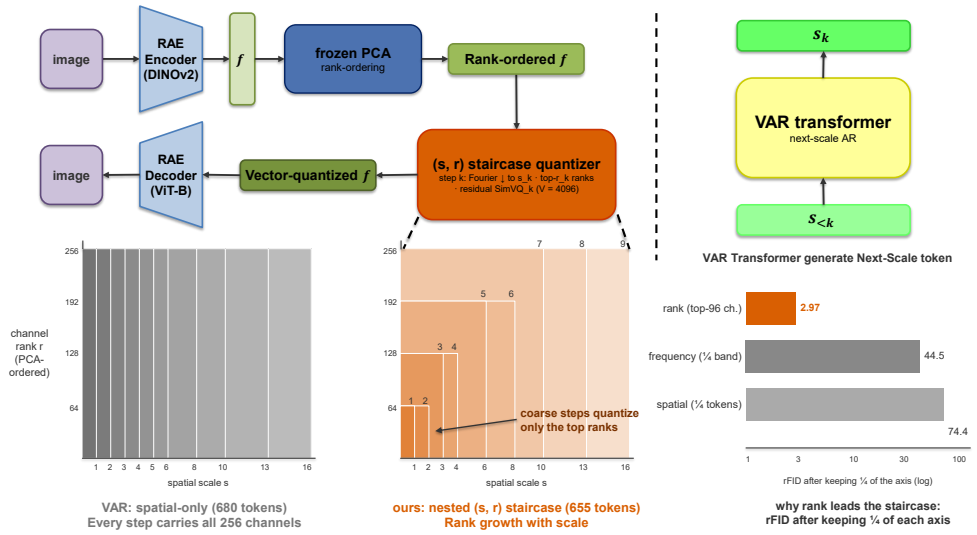

Figure 2: **RAE-VAR.** An anchored representation encoder feeds a frozen PCA projection and a residual staircase quantizer. Spatial scale and channel rank grow jointly; Fourier resampling supplies band-limited scale views. The VAR transformer is unchanged.

The same ordering observed in the frozen RAE (Table 1) also holds in our trained tokenizer (Fig. 2, right), where channel-rank truncation again preserves decoded content better than frequency or spatial truncation. These measurements motivate a channel-rank axis, although they also depend on the decoder's tolerance to truncated features (Appendices A.3 and A.4). These measurements motivate a channel-rank axis, although they also depend on the decoder's tolerance to truncated features (Appendices A.3 and A.4).

Easy to decode does not mean easy to predict. Reconstruction supplies the correct tokens from the encoder. Generation must instead predict them from the class label and previously sampled tokens. A low-rank code can therefore preserve a great deal of content when supplied correctly, yet be difficult to generate at the first step, when little context is available. Our rank-only experiments illustrate this distinction (Appendix C.4). In short, early scales should contain useful information, but not more information than what the generator can reliably predict from a short context.

Table 1: Decoded FID (5k) after truncating frozen RAE features at matched scalar budgets. Full-budget FID: 3.83.

Budget	Rank	Freq.	Spatial
6.25%	**79.6**	250.5	281.6
12.5%	**23.0**	175.0	235.2
25%	**7.1**	54.0	96.0
50%	**4.4**	8.0	82.8

Grow spatial scale and channel rank together. This is the intuition behind constrained allocation and the staircase in Fig. 2. We begin with a small spatial grid and a restricted set of channels. As the token prefix grows, later steps increase spatial resolution and introduce more channel directions. Information excluded from an early step remains in the residual for later steps, which have a richer conditioning context. Thus channel rank determines which feature directions a scale can encode, while spatial scale determines how finely it can describe them. Constrained allocation is a design principle for balancing these choices, not a separate optimization objective or a guarantee of predictability.

3.3 THE STAIRCASE QUANTIZER

We generalize the schedule to pairs (s_k, r_k). At step k, the residual is resampled to $s_k \times s_k$ and only its first r_k channels are quantized in a frozen PCA basis. The projection uses the teacher encoder's top 256 directions; the adaptation anchor keeps features tied to this coordinate system. Each rank group has a SimVQ codebook (Zhu et al., 2025) with 4096 entries. Codes are zero-padded to width

256 before reaching the transformer, so rank changes do not alter the backbone or its block-causal factorization.

Prefix interface. Let P_r select the first r channels in the PCA coordinate system and E_r embed an r-channel map back into 256 channels by zero-padding. They satisfy

$$P_r E_r = I_r, \qquad E_r P_r = \mathrm{diag}(\underbrace{1, \ldots, 1}_{r}, 0, \ldots, 0). \tag{2}$$

At scale k, codebook lookup operates on the r_k-channel residual view, while the accumulated contribution and transformer input retain the common width. Rank thus controls what each scale can encode, not the size of the generator's hidden state. Every spatial position still emits one categorical index, giving

$$N = \sum_{k=1}^{K} s_k^2, \qquad p(\mathbf{s}_1, \ldots, \mathbf{s}_K \mid y) = \prod_{k=1}^{K} p(\mathbf{s}_k \mid \mathbf{s}_{<k}, y). \tag{3}$$

The schedule changes the information assigned to these conditionals without changing the next-scale modeling objective.

The prefixes are nested (Kusupati et al., 2022): later steps revisit all earlier channels, so restricted rank defers rather than discards content. Nine steps increase rank from 64 to 256 and spatial side from 1 to 16, totaling 655 tokens. The exact schedule and basis provenance are in Appendix A.4 and Table 5.

Finally, we replace pooled resampling with a Fourier pair: retain low-frequency coefficients when downsampling and zero-fill when upsampling, with constant-preserving normalization. This supplies band-limited residual views instead of pooled approximations. The remaining residual is quantized at subsequent scales. This combination of nested rank prefixes, spatial progression, and Fourier resampling defines **RAE-VAR**; the experiments separate the representation-space baseline from the staircase package and test its basis and resampling choices.

The two structural choices play different roles. Rank restriction limits the channel subspace available at a step, whereas Fourier resampling determines its spatial frequency content. Neither is a substitute for the other: pooling can alter the residual passed to later scales even with the same rank schedule, and a random basis changes which directions a low-rank prefix can access even with the same resampling. This motivates separate basis and resampling ablations rather than attributing the entire spatial-to-staircase gain to rank alone.

4 Experiments

4.1 Experimental setup

We evaluate class-conditional ImageNet 256×256 generation (Deng et al., 2009). gFID (Heusel et al., 2017) uses 50k class-balanced samples and ADM reference statistics (Dhariwal & Nichol, 2021); rFID uses 50k validation reconstructions. Each family uses one fixed sampling configuration across epochs: $(\mathrm{cfg}, top{-}\mathrm{p}) = (1.2, 0.92)$ for RAE-VAR, $(3.0, 0.92)$ for RAE-spatial-VAR, and $(1.5, 0.96)$ for VQGAN-VAR, with top-k 900. These settings are chosen on final checkpoints, not retuned per milestone (Appendix B.3).

The RAE tokenizer has 111M trainable parameters versus 109M for VQGAN. The d16 generator has 310M parameters and follows VAR's augmented training recipe at global batch 768 on a 200-epoch cosine horizon. Our short-budget endpoint, 40+3, branches from epoch 40 into a three-epoch linear cooldown; all three matched systems use this protocol. We additionally train RAE-VAR-d16 for the full 200 epochs. Tokenizer training, auxiliary pretraining costs, sampling, and augmentation diagnostics are separated from generator epochs in AppendicesA.1, A.2, and A.3.

We select comparisons to answer three questions. First, does changing the tokenizer improve an otherwise unchanged next-scale generator? The matched VQGAN-VAR Tian et al. (2024), RAE-spatial-VAR, and RAE-VAR runs provide this controlled comparison, while released VAR checkpoints provide longer-training reference points. Second, how does our approach compare with other next-scale methods? MVAR Zhang et al. (2026) and MEPA Zhou et al. (2026) provide context

Table 2: **ImageNet** 256×256 **generation in context.** Parameters refer to the generator, excluding tokenizers and auxiliary guidance models; *MEPA reports activated MoE parameters. Steps count sampling iterations, not equivalent compute: MAR additionally uses a diffusion head. Detailed accounting and precision/recall are in Appendices B.1 and B.2.

Model	Gen. params	Train ep	Steps	gFID ↓	IS ↑
Diffusion / flow (reported)					
ADM-G (Dhariwal & Nichol, 2021)	554M	—	250	4.59	186.7
LDM-4-G (Rombach et al., 2022)	400M	—	250	3.60	247.7
DiT-XL/2 (Peebles & Xie, 2023)	675M	1400	250	2.27	278.2
JiT-L/16 (Li & He, 2026)	459M	600	50	2.36	298.5
JiT-G/16 (Li & He, 2026)	2.0B	600	50	1.82	292.6
SiT-XL/2 + REPA (Yu et al., 2025)	675M	800	250	1.42	305.7
RAE DiT-DH-XL (Zheng et al., 2026)	839M	800	50	1.13	262.6
Autoregressive / masked autoregressive (reported)					
LlamaGen-L (Sun et al., 2024)	343M	300	576	3.07	256.1
LlamaGen-3B (Sun et al., 2024)	3.1B	300	576	2.18	263.3
MAR-B (Li et al., 2024)	208M	800	256	2.31	281.7
MAR-H (Li et al., 2024)	943M	800	256	1.55	303.7
RAE-AR (SigLIP2) (Yu et al., 2026)	—	300	—	6.09	241.8
Next-scale autoregressive (reported)					
MVAR-d16 (Zhang et al., 2026)	310M	300	10	3.09	285.5
MEPA-d12 (Zhou et al., 2026)	252M*	200	10	2.86	288.4
MEPA-d16 (Zhou et al., 2026)	585M*	200	10	2.32	311.3
VAR-d16 comparisons under our evaluation (50k, ADM reference)					
VQGAN-VAR (repro.)	310M	40+3	10	4.88	215.8
RAE-spatial-VAR	310M	40+3	10	3.88	308.3
RAE-VAR	310M	**40+3**	9	**2.98**	270.0
Official VAR-d16 (Tian et al., 2024)	311M	200	10	3.15	265.0
RAE-VAR	310M	200	9	**2.58**	295.6

within this modeling family. Third, where does RAE-VAR sit among broader image-generation approaches? Diffusion/flow models, raster-order AR models, and masked AR models contextualize generation quality alongside generator size, training epochs, and sampling iterations. These cross-family results describe different operating regimes; they do not isolate the effect of our tokenizer.

What is controlled. The VQGAN baseline uses the released, fully trained tokenizer; only its generator is restricted to our short training budget. RAE-spatial-VAR changes the representation tokenizer while retaining the 680-token spatial schedule. RAE-VAR then reorganizes the tokens into the 655-token staircase. All three share generator depth, global batch, augmentation, and the cooldown protocol. Their guidance settings differ because the families have different sampling optima, but each setting is held fixed over training. This design separates a tokenizer-level improvement from changes in generator capacity or checkpoint-specific sampling searches.

Comparison protocol and interpretation. The re-evaluated VAR checkpoints provide head-to-head comparisons under one metric pipeline; published results from other model families provide context, not a compute-matched ranking. Generator epochs exclude encoder pretraining and tokenizer training, and epoch ratios are not wall-clock speedups. Parameter counts and sampling steps describe different aspects of cost and should not be treated as interchangeable. The short-budget d20/d24 runs test proximity to released endpoints, not equal asymptotic performance or statistical equivalence. Ablation results are single-run measurements at the stated budgets. Detailed accounting is collected in Appendices A.2 and B.1.

4.2 Main result: quality at a small VAR budget

At the same 43-epoch d16 budget, moving from VQGAN to representation space improves gFID from 4.88 to 3.88; the staircase tokenizer further improves it to $\mathbf{2.98}$ using nine rather than ten sampling steps (Table 2). This already surpasses the released 200-epoch VAR-d16 checkpoint (3.15

under our pipeline), with $4.65\times$ fewer generator epochs. At 200 epochs, RAE-VAR reaches $\mathbf{2.58}$, improving by 0.57 at the same backbone scale and schedule length. Recall also improves in the matched comparison; complete precision/recall and sampling hyperparameters are retained in Appendix Table 6.

Learning trajectory. The advantage is visible before the cooldown: at epoch 8, VQGAN-VAR, RAE-spatial-VAR, and RAE-VAR read 14.40, 7.58, and 6.05 gFID, respectively. By epoch 16, RAE-VAR reaches 4.70, already below the reproduced baseline's 43-epoch endpoint of 4.88. Figure 1 distinguishes the cooldown branch from uninterrupted training: the latter reaches 3.14 at epoch 100 and 2.58 at epoch 200. Thus the short-budget result is not presented as an intermediate point on a different learning-rate schedule; both trajectories are reported explicitly.

The broader comparison highlights the operating regime rather than an overall FID record. Our 310M model obtains lower reported FID than the similarly sized MVAR-d16 and MEPA-d12 endpoints, while larger MEPA and diffusion systems achieve stronger absolute results. Our strongest evidence is the matched VAR-d16 experiment: better quality with the same generator architecture and training budget, fewer tokens, and fewer next-scale passes.

Quality and coverage. Improved FID does not coincide with reduced recall in the controlled short-budget comparison: recall increases from 0.452 for VQGAN-VAR to 0.488 for spatial and 0.542 for the staircase, with corresponding precision 0.765, 0.735, and 0.753. We retain these diagnostics in the Appendix B.2 rather than ranking methods by Inception Score alone, since guidance changes both class confidence and coverage. The reported benefit is a better measured quality–coverage balance at the selected family sampling settings, not immunity to sampler-dependent trade-offs.

4.3 Scaling with depth

Table 3 pairs each short-budget RAE-VAR endpoint with the released VAR model at the same depth. Using the same tokenizer, RAE-VAR-d20 (601M) reaches 2.46 after 50+4 epochs, close to official VAR-d20's 2.43 after 250. RAE-VAR-d24 (1.03B) reaches 2.16 after 70+5 epochs, versus 2.00 after 350. Thus the deeper models approach the released checkpoints with approximately $4.6\times$ fewer generator epochs. Full metrics and intermediate milestones are in Appendices B.2 and B.4.

Table 3: **Depth-matched comparison.** 50k gFID under our pipeline; VAR denotes released checkpoints. Epochs count generator training only.

Depth	Model	Epochs	gFID ↓
d16	Ours	40+3	**2.98**
	VAR	200	3.15
d20	Ours	50+4	2.46
	VAR	250	**2.43**
d24	Ours	70+5	2.16
	VAR	350	**2.00**

The tokenizer is held fixed across depths, so this comparison tests whether its interface remains useful as generator capacity increases, rather than combining a larger generator with a separately improved autoencoder. The near-parity endpoints are accompanied by similar recall: 0.565 versus 0.575 at d20 and 0.582 versus 0.603 at d24. These results complement the d16 experiment: d16 establishes improved quality under both short and equal-length schedules, while the deeper runs establish proximity to strong released checkpoints at substantially shorter training lengths.

Across the three depth-matched pairs, the generator-epoch ratios are $4.65\times$, $4.63\times$, and $4.67\times$. The corresponding FID differences are -0.17, $+0.03$, and $+0.16$ for ours minus VAR. The same tokenizer therefore supports both the d16 improvement and competitive short-budget endpoints at larger depths. Within the available compute budget, we prioritized full-length training at d16 and short-budget scaling tests at d20 and d24; full-length training at the larger depths remains for future evaluation.

4.4 Ablations

Two conclusions emerge from Table 4. First, both the PCA basis and Fourier resampling matter for early generation quality. Second, the gains are not explained by final tokenizer reconstruction quality or training length alone. We distinguish the system-level spatial comparison from the controlled variants that test individual design choices.

Table 4: **Tokenizer ablations at epochs 8 and 16.** All generators are VAR-d16; metrics use 50k samples. Controlled variants use 105k-step tokenizers and fixed $(\mathrm{cfg}, top{-}\mathrm{p}) = (1.2, 0.92)$. The final row controls tokenizer training length. Full settings are in Appendix C.1.

system	basis	resampling	tokens	tok. steps	rFID	gFID ep8	gFID ep16
VQGAN-VAR (baseline)	- (VQGAN)	pool	680	released	3.39	14.40	8.66
RAE-spatial-VAR	learned proj.	pool	680	200k	1.094	7.58	5.37
RAE-VAR	400-image PCA	Fourier	655	200k	**0.955**	**6.05**	**4.70**
random basis	random orthonormal	Fourier	655	105k	1.39	10.37	6.18
pooled resampling	train PCA	pool	655	105k	1.39	7.99	6.12
PCA basis (ref.)	train PCA	Fourier	655	105k	1.12	6.36	5.29
PCA basis (longer training)	train PCA	Fourier	655	200k	0.963	6.28	5.36

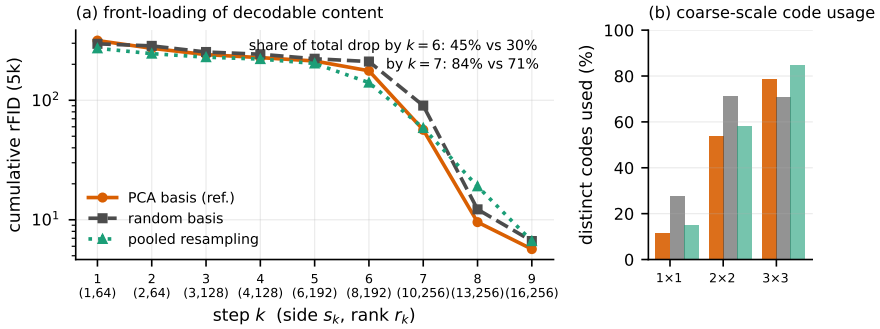

Figure 3: **Structured prefixes expose decodable content earlier.** PCA and random-basis reconstructions separate at intermediate scales (left); their coarse-scale code usage also differs (right). Measurements use 5k images; details are in Appendix C.1.

Both the structured basis and Fourier resampling matter. Three controlled staircase variants share the same 105k-step tokenizer training budget. Starting from the train-PCA/Fourier reference, one replaces the PCA basis with a random orthonormal basis; another replaces Fourier resampling with area pooling and bicubic upsampling. At epoch 8, the reference achieves gFID 6.36, compared with 10.37 for the random basis and 7.99 for pooled resampling. **Although the gaps narrow with longer training, the reference's advantage is unchanged: it outperforms both variants at epochs 8 and 16.** At epoch 16, the reference achieves 5.29, compared with 6.18 and 6.12, respectively. The random-basis comparison tests subspace selection and ordering together, rather than isolating ordering alone.

Reconstruction quality and training length are not sufficient explanations. Both ablated variants have the same rFID of 1.39, yet their epoch-8 generation quality differs substantially. Final reconstruction quality therefore does not predict their generation performance. We also extend the train-PCA tokenizer to 200k steps, matching the main tokenizer's training length. Its rFID improves to 0.963, close to the main tokenizer's 0.955, but its epoch-16 generator still achieves 5.36 rather than 4.70. Matching training length and nearly matching reconstruction quality do not reproduce the main result. The remaining difference involves basis provenance; this control does not identify which property of that basis causes the gap.

Spatial tokenization remains an important system-level ablation. RAE-spatial-VAR tests whether replacing the representation alone is enough. It improves over VQGAN-VAR, but RAE-VAR improves further: epoch-16 gFID decreases from 5.37 to 4.70, and the 40+3 endpoint from 3.88 to 2.98. This comparison supports the full staircase design, since projection, allocation, and resampling change together. The pooled staircase also remains behind the spatial baseline at epochs 8 and 16, showing that fewer tokens and a rank schedule alone do not recover the full gain. The controlled variants above provide the more targeted evidence for the basis and resampling choices. Exact checkpoint-step accounting, including the pooled variant's 1.2% extra generator steps, is reported in Appendix C.1.

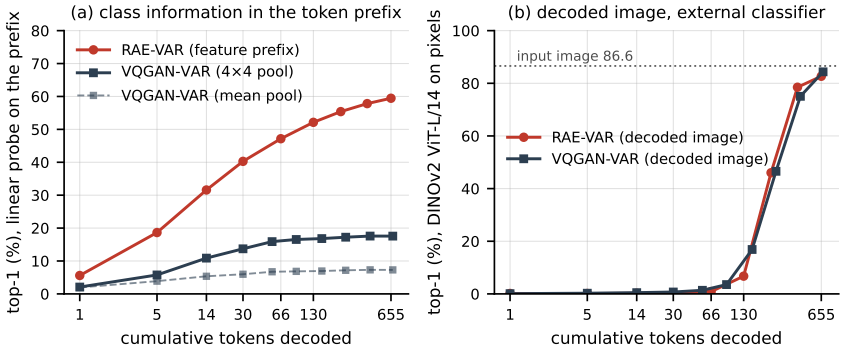

Figure 4: **Class information precedes pixel recognizability.** Absolute ImageNet Top-1 accuracy from per-prefix linear probes (left, 50k validation images) and a frozen DINOv2-L classifier on decoded reconstructions (right, 10k). Protocols and numerical results are in Appendix E.1.

4.5 Analysis: what the structured code preserves

Figure 3 measures reconstruction from cumulative ground-truth prefixes. By steps 6 and 7, the PCA reference realizes 45% and 84% of its total reconstruction gain, versus 30% and 71% for the random basis. Their coarse-scale code usage also differs: at the first two scales, PCA uses 11.5% and 53.8% of the vocabulary, compared with 27.5% and 71.2% for the random basis. This is consistent with the structured basis concentrating decodable content into early, weakly conditioned steps.

Neither reconstruction nor teacher-forced accuracy alone explains generation quality. The PCA reference reconstructs slightly worse than the spatial tokenizer but generates better, while the random-basis model has lower teacher-forced cross-entropy yet generates worse. Both tokenizers eventually use the full vocabulary, so final utilization is likewise insufficient. Per-step results and failure analyses are in Appendices C.1, C.4, and F.1.

Figure 4 separates class information in the latent prefix from recognizability after decoding to pixels. At matched early token budgets, RAE-VAR prefixes achieve higher linear-probe accuracy than either VQGAN readout even while their decoded images remain near chance; pixel recognizability emerges later. Thus, coarse tokens can carry class-related information without forming recognizable thumbnails.

These readouts are complementary rather than capacity-matched: the probes differ in pooling and dimensionality, and the pixel classifier shares the tokenizer's DINOv2 family. We therefore interpret the result only as evidence of linearly accessible class information in coarse prefixes, not that all later tokens encode only texture or that the probe exhausts representation semantics. Full protocols and numerical values are in Appendix E.1.

5 Conclusion

RAE-VAR organizes representation features into a staircase of spatial scales and channel ranks while preserving the VAR backbone. With the unchanged 310M-parameter generator, it reaches gFID 2.98 in 43 epochs, surpassing released VAR-d16 after 200 epochs, and improves to 2.58 with full-length training. Deeper models approach released VAR checkpoints with roughly one fifth of their generator epochs. These gains require neither a larger transformer nor an auxiliary generator-alignment loss, and use nine next-scale passes rather than ten. Ablations show that final reconstruction quality alone does not explain generation performance, while probes reveal class information before decoded prefixes become recognizable. The central lesson is that representation-space coarse-to-fine generation should allocate both spatial and channel capacity: early tokens should preserve useful content while leaving difficult predictions to steps with richer context.

Limitations. We focus on class-conditional ImageNet-256 to compare tokenization designs under a common dataset, resolution, and evaluation protocol, using released VAR checkpoints as reference points. This setting supports controlled comparisons, but transfer to other datasets, resolutions, and conditioning tasks remains untested. Within the available compute budget, we completed full-length training at d16 and shorter-horizon evaluations of deeper models and controlled ablations, leaving their asymptotic performance unresolved. Future work can isolate these effects and extend the allocation principle to other representations and tasks.

AI USE STATEMENT

Generative AI tools were used as a coding and analysis assistant throughout this work: writing and refactoring training and evaluation code, orchestrating experiment launches and monitoring, and drafting and editing portions of this manuscript. All AI-assisted code was executed and its outputs checked against independent measurements where possible; in particular, evaluation-pipeline changes were validated by reproducing previously measured numbers, and one measurement bug found this way is described in Appendix. C.1. Research questions, experimental design decisions, and all reported conclusions were determined by the authors. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

All quantitative claims are measured under a single evaluation pipeline described in §4.1, with the tokenizer construction in §3.3. Appendix. A.3 gives the tokenizer implementation (rank schedule, per-rank quantizers, the provenance and measured quality of the frozen projection basis, and the construction of the ablation arms) and Appendix. A.1 the training and evaluation protocol, including the fixed-coordinate sampling rule applied to every curve point and the reference statistics used for FID. Because FID pipelines disagree across papers, we re-evaluate the officially released baseline checkpoints under our own pipeline (§4.2) so that every number in the head-to-head comparison is produced by the same code path. The complete per-milestone curves behind Fig. 1, the final-checkpoint sampling grids, and the per-step, per-scale readouts behind the ablation are reported in Appendix. B.3 and C.1 rather than summarized, so that any single number can be located in its curve.

REFERENCES

Peter J. Burt and Edward H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983.

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

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

Sinan Du, Jiahao Guo, Bo Li, Shuhao Cui, Zhengzhuo Xu, Yifu Luo, Yongxian Wei, Kun Gai, Xinggang Wang, Kai Wu, et al. Vqrae: Representation quantization autoencoders for multimodal understanding, generation and reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30322–30334, 2026.

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

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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

Mingkai Jia, Mingxiao Li, Zhijian Shu, Anlin Zheng, Liaoyuan Fan, Jiaxin Guo, Tianxing Shi, Dongyue Lu, Zeming Li, Xiaoyang Guo, et al. Dino-tok: Adapting dino for visual tokenizers. *arXiv preprint arXiv:2511.20565*, 2025.

Siyu Jiao, Gengwei Zhang, Yinlong Qian, Jiancheng Huang, Yao Zhao, Humphrey Shi, Lin Ma, Yunchao Wei, and Zequn Jie. Flexvar: Flexible visual autoregressive modeling without residual prediction. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 8955–8978.

Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/0d0d75da30e2d81b928d793afe389ad4-Paper-Conference.pdf.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning of latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18262–18272, October 2025.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *International Conference on Learning Representations (ICLR)*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. FlowAR: Scale-wise autoregressive image generation meets flow matching. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 51489–51502. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/ren25e.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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

Chenze Shao, Fandong Meng, and Jie Zhou. Continuous visual autoregressive generation via score maximization. In *International Conference on Machine Learning (ICML)*, 2025.

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

Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. Hart: Efficient visual generation with hybrid autoregressive transformer. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 68413–68432, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ab4e1e704c68b0f476c265996f08d283-Paper-Conference.pdf.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

Michael Tschannen, Cian Eastwood, and Fabian Mentzer. GIVT: Generative infinite-vocabulary transformers. In *European Conference on Computer Vision (ECCV)*, 2024.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

Yuqing Wang, Chuofan Ma, Zhijie Lin, Yao Teng, Lijun Yu, Shuai Wang, Jiaming Han, Jiashi Feng, Yi Jiang, and Xihui Liu. Cubic discrete diffusion: Discrete visual generation on high-dimensional representation tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36072–36081, June 2026.

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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Jiawei Yang, Zhengyang Geng, Xuan Ju, Yonglong Tian, and Yue Wang. Representation fréchet loss for visual generation, 2026.

Hu Yu, Hang Xu, Jie Huang, Zeyue Xue, Haoyang Huang, Nan Duan, and Feng Zhao. Rae-ar: Taming autoregressive models with representation autoencoders. *arXiv preprint arXiv:2604.01545*, 2026.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations (ICLR)*, 2025.

Jinhua Zhang, Wei Long, Minghao Han, Weiyi You, and Shuhang Gu. Mvar: Visual autoregressive modeling with scale and spatial markovian conditioning. In *International Conference on Learning Representations*, volume 2026, pp. 32889–32917, 2026.

Qian Zhang, Xiangzi Dai, Ninghua Yang, Xiang An, Ziyong Feng, and Xingyu Ren. Var-clip: Text-to-image generator with visual auto-regressive modeling. *arXiv preprint arXiv:2408.01181*, 2024.

Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 90844–90868, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/e25198b6a75f74277ee3a2bd4165d9ef-Paper-Conference.pdf.

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

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0u1LigJaab.

Nuoyan Zhou, Zhijun Tu, Lei Yu, Kun Cheng, Jie Hu, Nannan Wang, and Xinghao Chen. Multi-scale representation alignment for visual autoregressive modeling with mixture of experts. In *European Conference on Computer Vision*, pp. 77–96. Springer, 2026.

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

GUIDE TO THE APPENDIX

The appendix is organized into six thematic sections. Experimental protocols and implementation details come first; the remaining sections collect supporting results, analyses, and visual examples.

Appendix A: Experimental and implementation details. Training recipes, evaluation protocols, compute accounting, tokenizer implementation, and the exact staircase schedule.

Appendix B: Extended results and sampling analysis. Cross-paper comparison conventions, full system metrics, sampling grids, epochs to target quality, stronger-guidance curves, and the FDr^6 metric.

Appendix C: Ablations and failure analyses. Controlled tokenizer ablations, augmentation and memorization diagnostics, rank-only schedules, and continuous-prediction trials.

Appendix D: Representation structure and compaction. Channel and spatial statistics across encoders, and decoded-image comparisons of rank and spatial truncation.

Appendix E: Semantic information in token prefixes. Per-prefix probe protocols and accuracies, plus decoded-image classification curves for reconstruction and generation.

Appendix F: Qualitative reconstructions and samples. Per-scale reconstruction and generation decodes, followed by uncurated sample grids.

A EXPERIMENTAL AND IMPLEMENTATION DETAILS

A.1 RECIPE AND EVALUATION DETAILS

Precision. All training runs use full fp32 weights with TF32 matmul (a $4.5\times$ throughput gain over strict fp32 on our hardware with no measurable quality change); evaluation precision is recorded per run and never mixed within a comparison.

Step matching. All cross-system claims are matched in optimizer steps at fixed global batch 768 (generators) / 256 (tokenizers). We document one trap: an epoch-matched comparison across different batch sizes briefly produced a false $2\times$ speedup for one variant, which vanished under step matching.

Generator hyperparameters. AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$), peak learning rate $3\text{e}{-4}$, weight decay 0.05, gradient clipping 1.0, label smoothing 0.1, conditional dropout 0.1, drop-path $0.1 \cdot \text{depth}/24$, attention ℓ_2-normalization, 2k warmup steps, cosine decay over the stated horizon; global batch 768 at every depth (local batch $\times$ gradient accumulation).

Augmentation. Generators use VAR's training transform: resize the shorter side to $1.125\times$ the target resolution (Lanczos), random 256×256 crop, horizontal flip off as in VAR's default; RAE-side encoders see the crop resized to 224. Encoding is done on the fly ($\approx 4\%$ throughput cost) since the crop changes every epoch.

Decay tails. All d16 systems execute the same 200-epoch cosine schedule; the matched-budget point branches from the epoch-40 checkpoint and decays the learning rate linearly to zero over 3 epochs (no EMA, no schedule recompression), recovering 7–20% gFID ($5.85 \rightarrow 4.88$ baseline, $4.18 \rightarrow 3.88$ spatial, $3.71 \rightarrow 2.98$ ours). d20 branches at epoch 50 with a 4-epoch tail on the same 200-epoch horizon; d24 branches at epoch 70 with a 5-epoch tail on the official 350-epoch horizon. The d16 200-epoch number is the natural end of its cosine schedule and has no tail.

Sampling. One (cfg, top-p) coordinate per family, chosen once on the family's final checkpoint by a grid over cfg $\in \{1.0, 1.2, 1.5, 2.0, 2.5, 3.0, \ldots\}$ and top-$p \in \{0.88, \ldots, 0.98, \text{off}\}$ at top-k 900, and then used unchanged for every epoch and every depth of that family (Appendix B.3). No per-checkpoint re-tuning is performed anywhere; re-tuning the sampler at every epoch would turn a convergence curve into an envelope of different models. The released VAR checkpoints are read at their own calibrated optimum, $(1.5, \text{off})$ for d16–d24 and $(1.5, 0.96)$ for d30.

Metrics. IS uses 10 shuffled splits on unbiased logits; precision/recall use $k = 3$ nearest neighbors in Inception-2048 features against 50k ImageNet-val images; both are computed from each model's 50k samples at its family sampling setting on one pipeline. Our harness reproduces the ADM evaluator to within 0.002 FID on the same samples. rFID reconstructs the full 50k validation set and compares against the originals (paired; VQ tokenizers go through the full quantized round trip).

Compute. Measured wall-clock on $4\times$ RTX PRO 6000 (96 GB), inclusive of teacher-encoder forwards and inline milestone evaluations: staircase tokenizer $\approx$300 GPU-hours to 200k steps; VAR-d16 $\approx$240 GPU-hours for 40 epochs plus tail and $\approx 5\times$ that for the 200-epoch run; d20 $\approx 1.6\times$ the d16 cost per epoch, d24 scaling further with its parameter count. Each ablation arm (105k-step tokenizer, 8-epoch generator, evaluation) costs about two days on four GPUs.

A.2 Additional experimental protocol and metric accounting

Training budgets and adaptation. The trainable RAE tokenizer stack comprises the adapted encoder, projection, quantizer, and ViT-B decoder (111M parameters), versus 109M for VQGAN. Tokenizer training uses global batch 256 and a 100-epoch cosine horizon; main checkpoints are taken at 200k steps. The spatial and staircase tokenizers reach rFID 1.094 and 0.955. The generator uses the VAR optimizer, regularization, warmup, and crop augmentation at global batch 768 (Appendix A.1). At d16 all matched systems share a 200-epoch cosine horizon; the 40+3 branch applies a linear three-epoch decay from epoch 40. This improves gFID from 5.85 to 4.88 for VQGAN-VAR, 4.18 to 3.88 for spatial, and 3.71 to 2.98 for RAE-VAR. The d20 and d24 endpoints use 50+4 and 70+5 tails; d24 follows the official 350-epoch horizon. Epoch ratios count generator data passes, excluding tokenizer training and pretrained-encoder construction; they are not FLOP or wall-clock speedup estimates.

RAE-side systems follow the RAE resizing protocol (224 encoder input, 256 decoder output), whereas VQGAN is 256-native. All reported image metrics use 256×256 outputs. Augmentation is retained for every headline generator; the unaugmented runs and their train–validation gaps are reported separately in Appendices C.2 and C.3.

Diversity metrics. Inception Score (Salimans et al., 2016) uses ten shuffled splits. Improved precision and recall (Kynkäänniemi et al., 2019) use $k = 3$ in Inception-2048 space against 50k ImageNet validation images. At the matched budget, recall rises from 0.452 (VQGAN-VAR) to 0.488 (spatial) to 0.542 (RAE-VAR), alongside improved gFID. At 200 epochs, RAE-VAR has recall 0.563 versus 0.524 for the official d16 checkpoint at its separately calibrated metric coordinate. These measurements check diversity but do not eliminate sampler-dependent tradeoffs. In particular, the official checkpoint's curve-coordinate FID is 3.15, while its IS/precision/recall are recorded at the coordinate giving FID 3.18; the main table leaves its IS blank to avoid pairing metrics from different coordinates.

Figure 1 accounting. Figure 1 shows milestones from epoch 4 onward. The 40+3 cooldown branch and uninterrupted 200-epoch trajectory are distinct runs from the epoch-40 state. The cooldown endpoint reaches 2.98 at epoch 43, while uninterrupted training reaches 3.14 at epoch 100. MEPA-d12 and RAE-AR are reported final-only context, not runs under our evaluation.

A.3 Tokenizer Implementation Details

PCA basis. The rank-ordering basis is the eigenbasis of $\sim 10^5$ patch tokens (1.024×10^5) of the *frozen teacher* encoder. The $384 \to 256$ input projection is initialized to the top-256 eigendirections and kept frozen; validity under encoder fine-tuning is maintained by the distillation anchor (§3.1). Measured in the main space, removing the tail *improves* decoding (rFID 4.09 full $\to$ 3.56 top-256 $\to$ 2.97 top-96), so the frozen cut discards noise-like directions.

Provenance and quality of the main basis. The 400 images are the first 400 of the validation split in sorted order, which span only 8 of the 1000 classes. The basis is therefore a coarse estimate of the ImageNet variance ordering, and we report its quality against the *true* distribution rather than against its own fitting sample. Writing Σ for the covariance of the full training distribution, the retained variance of the top-k prefix of a basis B is $\|\Sigma^{1/2} B_{:k}\|_F^2 / \operatorname{tr} \Sigma$:

k	class-uniform train PCA	main (400-image)	random orthonormal basis
32	28.4%	18.5%	8.3%
64	43.3%	31.5%	16.7%
128	63.6%	50.4%	33.3%
256	**88.0%**	**77.0%**	66.7%

The main basis recovers about half of the achievable gain over a random basis (48.5% at k=256, 50.4% at k=32): it is a degraded, not a random, ordering. Subspace diagnostics agree — prefix overlap with the class-uniform train basis is $0.42/0.51/0.61/0.77$ at k=$32/64/128/256$ (random baseline $0.08/0.17/0.33/0.67$), while the mean $|\cos|$ between corresponding top-32 axes is only 0.08: the prefixes span similar subspaces, the individual axes do not. The basis is a property of the representation rather than of the split — a class-uniform *validation* basis agrees with the class-uniform *training* basis at prefix overlap 0.98 and mean axis $|\cos|$ 0.77 — so the issue is class coverage, not train/test leakage. Its flattering 96.0% retained variance when measured on its own 8-class sample is a coverage artifact.

All headline results use the main basis. The ablation arms of §4.4 use a controlled tokenizer family built on the class-uniform training-split basis: identical recipe and schedule, stopped at 105k steps of the 100-epoch horizon — a point inside the reference tokenizer's rFID plateau band (0.96–1.13 from 80k steps on), chosen once from the reference arm's curve and applied to every arm so that training length is not a variable — with the random-basis arm replacing the PCA basis by a random orthonormal basis (seed 0; both the retained subspace and the order are random) and the pooled arm replacing the Fourier pair by area pooling with bicubic upsampling. We did not retrain the headline models on the class-uniform basis for want of compute. The headline numbers were therefore obtained with a degraded estimate of the intended variance structure; whether an exact class-uniform basis improves generation remains to be tested.

Encoder adaptation. Quantizing the frozen DINOv2 features fails in our runs with SimVQ, FSQ and plain VQ: rFID exceeds 100 within 5k steps. We attribute this behavior to the near-full-rank features (Table above) and to the pretrained decoder expecting continuous inputs. The encoder is therefore fine-tuned at a $10\times$ lower learning rate under an L_2 self-distillation anchor to a frozen copy of itself, which keeps linear-probe accuracy within 0.05 points of the frozen encoder while making the space quantizable.

Per-rank quantizers. Each distinct rank $r \in \{64, 128, 192, 256\}$ has its own SimVQ codebook ($V = 4096$, reparametrized through one linear layer) and its own per-scale refinement convolution Φ. Codes are zero-padded to $C = 256$ in the rotated (PCA) space before entering the transformer, so token count, attention mask, and embedding width are rank-independent and the VAR model is byte-identical to the released implementation.

Fourier resampling. Downsampling to $s \times s$ crops the centered $s \times s$ window of the orthonormal 2D FFT, scaled by s/H so constant maps are preserved; upsampling zero-pads the same window. At even sizes the centered window is Nyquist-asymmetric; we take the real part after the inverse transform (Hermitian symmetrization). Because the up-path adds no information, down$\circ$up is the identity on band-limited maps (verified to 5×10^{-7}), so each coarse step sees an exact low-pass of the running residual and cannot alias fine content. Ringing from the sharp cutoff remains in the residual and is re-quantized at finer steps.

Schedule. The staircase is $[(1, 64),\ (2, 64),\ (3, 128),\ (4, 128),\ (6, 192),\ (8, 192),\ (10, 256),\ (13, 256),\ (16, 256)]$, 655 tokens; the 5×5 scale of the baseline ladder is omitted (the entirety of the $680 \to 655$ reduction). The structure is nested: later steps revisit all channels.

Autoencoder protocol. DINOv2-S with registers (patch-14, 384-d); 256×256 inputs bicubic-resized to 224; final layernorm without affine; CLS and register tokens dropped; ImageNet-statistic channel standardization; ViT-B decoder to 256×256. The encoder is fine-tuned at lr 10^{-5} under an L2 self-distillation anchor (weight 1.0) to a frozen teacher copy; the decoder trains through the quantized path; everything is frozen before generator training.

Table 5: The staircase schedule, which grows channel rank jointly with spatial scale and totals 655 tokens against the 680 of the spatial baseline.

step k	1	2	3	4	5	6	7	8	9
side s_k	1	2	3	4	6	8	10	13	16
rank r_k	64	64	128	128	192	192	256	256	256

A.4 Allocation and tokenizer details

Truncation protocol. The frozen-RAE measurement in Table 1 uses the pretrained DINOv2-B RAE and its official decoder. We retain either the top-r PCA channels, a radial low-pass region of the 2D Fourier transform, or an area-pooled spatial grid. Each truncated representation is returned to the full 16×16 grid before decoding, and FID is evaluated on 5k images. The full-budget reference is 3.83. Because the decoder was trained on full features, the measurement combines feature compaction with decoder robustness to truncation. Repeating the comparison within the adapted main autoencoder reproduces the ordering at 50k rFID: at a quarter scalar budget, rank, frequency, and spatial truncation yield 2.97, 44.5, and 74.4 respectively. Dropping the PCA tail also improves this decoder readout from 4.09 at 384 channels to 3.56 at the top 256 (Appendix A.3).

Exact staircase schedule. Each distinct rank owns a low-dimensional SimVQ codebook with vocabulary 4096 and a refinement module. Quantized contributions are zero-padded to width 256 in the PCA coordinate system before being passed to the generator. Later ranks contain all earlier channels; a coarse restriction therefore delays residual information instead of permanently removing it. The main PCA is fitted on 400 images from the frozen teacher encoder; its provenance and the class-uniform train-PCA controls are documented in Appendix A.3. The teacher anchor constrains encoder drift but does not make the adapted features' covariance identical to the teacher's.

Fourier resampling. Downsampling retains the low-frequency block of an orthonormal 2D FFT, and upsampling zero-fills omitted frequencies with constant-preserving scaling and real-valued output. Unlike area pooling and bicubic interpolation, this constructs a band-limited scale view. A sharp cutoff can introduce spatial ringing; residual quantization carries remaining discrepancies to later steps, and the refinement modules can model local corrections. Implementation conventions at even-sized frequency boundaries matter, so the empirical Fourier-versus-pooling ablation, rather than an idealized identity argument alone, supports this choice.

B Extended Results and Sampling Analysis

B.1 Cross-paper comparison conventions

Table 2 separates published context from our shared evaluation. Training epochs exclude encoder pretraining, tokenizer training, and any auxiliary guidance-model training. Reported parameter counts refer to generators, not total system storage or training cost. MEPA's MoE counts are activated parameters rather than total stored parameters (Zhou et al., 2026); these are not directly equivalent to dense-model counts.

The steps column records algorithmic sampling iterations, not a uniform NFE measure. VAR uses one transformer prediction per scale, before classifier-free guidance. LlamaGen's reported high-quality configuration uses 576 raster tokens and resizes its 384×384 outputs for ImageNet-256 evaluation (Sun et al., 2024). MAR's best reported results use 256 autoregressive iterations, rather than the 64-iteration default used in its ablations, with additional diffusion-head evaluations per iteration (100 denoising steps by default) (Li et al., 2024). JiT (Li & He, 2026) uses 50 Heun steps, which require more model evaluations than 50 Euler steps. REPA uses 250 Euler–Maruyama steps with guidance intervals for the reported 1.42 result (Yu et al., 2025). RAE uses a 50-step Euler sampler and AutoGuidance; the auxiliary guiding model is excluded from the 839M generator count (Zheng et al., 2026). Neither this column nor a generator-parameter count establishes equal sampling cost.

For ADM we report the single-stage guided model (gFID 4.59), not the two-stage guided-plus-upsampling result (3.94) (Dhariwal & Nichol, 2021). The 554M generator count excludes the clas-

sifier. LDM-4-G reports 178k optimizer steps at batch 1200 rather than a directly reported epoch count; we leave its epoch cell blank (Rombach et al., 2022). Missing accounting fields are left blank rather than inferred.

Our main claim is the measured improvement within the matched VAR-d16 setup. Published diffusion, masked AR, and larger next-scale systems can have better absolute FID, so the table does not establish a universal state-of-the-art claim or a compute-matched ranking.

B.2 Extended system comparison

Table 6 supplements the main comparison with tokenizer sizes, precision/recall, and depth-dependent endpoints. Published context is separated from our re-evaluated checkpoints; the sampling-coordinate distinction for official VAR-d16 is stated explicitly.

Table 6: **Extended ImageNet** 256×256 **comparison and diversity metrics.** "steps" counts sampling iterations before guidance overhead; "#params" lists tokenizer / generator parameters. The REPA row uses constant guidance (the main table instead reports its guidance-interval result); RAE uses AutoGuidance. See Appendix B.1 for accounting. Only the lower two panels support quantitative claims. The middle panel samples every family at one fixed (cfg, top-p) chosen once on its final checkpoint (RAE-VAR $(1.2, 0.92)$, RAE-spatial-VAR $(3.0, 0.92)$, VQGAN-VAR $(1.5, 0.96)$; top-k 900), the same coordinate used for every point of Fig. 1; "40+3" is the epoch-40 checkpoint plus a 3-epoch decay tail (§4.1). †The released d16 checkpoint achieves 3.15 at the fixed coordinate of Fig. 1 and 3.18 at its own calibrated optimum $(1.5,\ \text{off})$, where its IS / precision / recall are measured. IS / precision / recall are reported for the 40+3 and final checkpoints; the 100-epoch row is gFID only.

system	tokenizer	#params	train ep	steps	gFID	IS	Prec.	Rec.
As reported by the respective papers, on their own evaluation pipelines (context only).								
VAR-d16 (Tian et al., 2024)	VQGAN	109M / 310M	200	10	3.30	274.4	0.84	0.51
MVAR-d16 (Zhang et al., 2026)	VQGAN	109M / 310M	300	10	3.09	285.5	0.85	0.51
MEPA-d16 (Zhou et al., 2026)	VQGAN	109M / 585M (MoE)	200	10	2.32	311.3	0.82	0.57
SiT-XL/2 + REPA (Yu et al., 2025)	SD-VAE	∼84M / 675M	800	250	1.80	284.0	0.81	0.61
RAE DiT-DH-XL (Zheng et al., 2026)	RAE	∼500M / 839M	800	50	1.13	262.6	0.78	0.67
Matched budget at d16, measured under our pipeline (50k, ADM ref.), one fixed sampling configuration per family.								
VQGAN-VAR (repro.)	VQGAN	109M / 310M	40+3	10	4.88	215.8	0.765	0.452
RAE-spatial-VAR	RAE, spatial	111M / 310M	40+3	10	3.88	308.3	0.735	0.488
RAE-VAR	RAE, staircase	111M / 310M	40+3	9	**2.98**	270.0	0.753	0.542
RAE-VAR	RAE, staircase	111M / 310M	100	9	3.14	—	—	—
RAE-VAR	RAE, staircase	111M / 310M	200	9	**2.58**	295.6	0.748	**0.563**
official VAR-d16 (Tian et al., 2024)	VQGAN	109M / 311M	200	10	3.15†	265.0	0.753	0.524
Depth, measured under our pipeline; released checkpoints at their own calibrated optimum.								
official VAR-d20 (Tian et al., 2024)	VQGAN	109M / 601M	250	10	2.43	289.4	0.733	0.575
RAE-VAR-d20	RAE, staircase	111M / 601M	**50+4**	9	2.46	**300.7**	0.745	0.565
official VAR-d24 (Tian et al., 2024)	VQGAN	109M / 1.03B	350	10	2.00	299.2	0.718	0.603
RAE-VAR-d24	RAE, staircase	111M / 1.03B	**70+5**	9	2.16	**312.3**	0.740	0.582
official VAR-d30 (Tian et al., 2024)	VQGAN	109M / 2.01B	350	10	1.96	305.2	0.734	0.589

B.3 Sampling Grids of the Final Checkpoints

The fixed family sampling settings of §4.1 were chosen on the final checkpoints and sit within 0.02 gFID of each final grid minimum: RAE-VAR-d16 (200 ep) achieves 2.575 at $(1.2, 0.92)$ against a grid minimum of 2.564 at $(1.2, 0.90)$; RAE-VAR-d20 (50+4) achieves 2.459 at $(1.2, 0.92)$, its grid minimum; RAE-VAR-d24 (70+5) achieves 2.161 at $(1.2, 0.92)$ against 2.157 at $(1.5, 0.94)$. RAE-spatial-VAR's surface is interior on both axes at its coordinate (epoch 40: cfg $1.2/1.5/2.5/3.0/3.5/4.0$ at top-p $0.92 = 6.75/5.83/4.31/4.18/4.35/4.62$; top-$p$ $0.88/0.92/0.96$ at cfg $3.0 = 4.45/4.18/4.62$), as is the VQGAN baseline's after its tail (cfg $1.0/1.5/2.0/2.5$ at $0.96 = 6.16/4.81/4.88/5.68$; top-$p$ $0.94/0.96/0.98/\text{off}$ at cfg $1.5 = 4.90/4.81/4.84/4.85$; the fixed-sampler re-measurement on a fresh 50k sample set achieves 4.88, the 0.06 difference being sample-set variance). The guidance optimum differs by family — $\approx$1.2 for the staircase, $\approx$3 for the

spatial pyramid, 1.5 for VQGAN tokens — which is why one inherited constant would mis-rank the families and why each is read at its own coordinate.

B.4 Epochs to Quality

Table 7 reads Fig. 1 along iso-quality lines: for each reference level, the epoch at which every family first reaches it. The epoch-1 tie between RAE-VAR-d16 and the VQGAN baseline (54.4 against 55.2) shows that the advantage is not an initialization effect; it opens from epoch 2 onward and compounds with depth.

Table 7: Epochs at which each family first reaches a reference gFID (fixed family sampling settings, 50k). The matched-epoch gap between RAE-VAR-d16 and the VQGAN baseline is 2.4× at epoch 8, 1.8× at epoch 16 and 1.6× at epoch 40.

reference level (gFID)	VQGAN-VAR d16	RAE-VAR d16	d20	d24	reading
4.9 (baseline, annealed 40+3)	43	16	8	<8	2.7× / 5.4× fewer epochs
3.3	—	55	25	16	fixed ratio along the family
3.15 (released VAR-d16, 200 ep)	200	100	50	40	2× / 4× / 5×
2.58 (RAE-VAR-d16, end of schedule)	—	200	—	70	depth buys 2.9× epochs

B.5 Training-Time Curves at a Stronger Guidance

The inline milestone evaluations logged during training use cfg 3.0 and top-p 0.96, a stronger guidance than the fixed main coordinate. Table 8 lists them for completeness; they are not comparable to Fig. 1. The two settings cross at about epoch 100 for d16: the stronger guidance is better early (epoch 40: 3.50 against 3.71) and worse late (epoch 200: 3.08 against 2.575), because the optimal guidance strength falls as the model converges and the learning rate anneals — the same effect that moves the VQGAN baseline's guidance optimum from 2.0 to 1.5 across its decay tail (Appendix B.3). This is why every curve in the paper is read at one coordinate chosen on the final checkpoint rather than at the training-time setting. Mid-training checkpoints also fluctuate genuinely under either setting (d16 at the main coordinate: epoch 55 3.34 → 70 3.70 → 85 3.41, far above the ±0.03 50k-FID noise floor); monotone improvement resumes after epoch 100 as the learning rate decays.

Table 8: Training-time inline evaluations (50k, cfg 3.0, top-p 0.96). Compare with the fixed-sampler curve of Fig. 1: d16 achieves 3.71 at epoch 40 and 2.575 at epoch 200 there.

inline gFID at cfg 3.0, top-p 0.96	ep1	ep2	ep4	ep8	ep16	ep40	ep100	ep200
RAE-VAR-d16	47.61	19.07	7.92	4.66	3.98	3.50	3.34	3.08
RAE-VAR-d20	40.0	14.3	6.45	3.99	3.77	3.21	—	—
RAE-VAR-d24	34.08	10.71	5.28	3.78	3.37	2.73	—	—

B.6 A Second Distributional Metric: FDr^6

Because FID with an Inception backbone is one view of sample quality, we also report the representation Fréchet ratio of Yang et al. (2026): for each of six frozen encoders ϕ (Inception, ConvNeXt-V2, MAE, DINOv2, SigLIP2, CLIP), $\mathrm{FDr}_\phi(G) = \mathrm{FD}_\phi(G, \mathrm{train})/\mathrm{FD}_\phi(\mathrm{val}, \mathrm{train})$, the arithmetic mean over the six being FDr^6; real held-out data scores exactly 1.0. Reference moments come from the full 1.28M training split and 50k samples per model. Table 9 orders every final checkpoint by it.

Three readings. First, the ordering is consistent with gFID within each family and separates the matched-budget systems more strongly (1.9× between RAE-VAR-d16 and the VQGAN baseline at 40+3 where gFID separates them 1.6×). Second, FDr is *not* a memorization-proof metric: because its reference is the training set, the unaugmented models score *better* than their augmented counterparts at every depth (6.46 against 6.69 at d16, 5.32 against 6.43 at d20, 5.37 against 5.74 at d24) even though Appendix C.3 shows them to be worse on held-out data, so it cannot replace the train–validation cross-entropy gap as the memorization control. Third, the DINOv2 column is

Table 9: Per-encoder Fréchet ratios and their mean on every final checkpoint (50k samples; training-split reference; 1.0 = held-out real data). Published base models span $\mathrm{FDr}^6 \approx 6.9$–$8.4$ on this metric.

model	ep	gFID	Incep.	ConvNeXt	MAE	DINOv2	SigLIP2	CLIP	FDr^6
official VAR-d24	350	2.00	1.23	2.69	7.14	6.01	6.90	6.93	**5.15**
RAE-VAR-d20, unaugmented	40+3	2.14	1.31	4.20	7.29	6.58	6.94	5.59	5.32
RAE-VAR-d24, unaugmented	40+3	2.06	1.41	4.23	7.10	6.78	7.06	5.66	5.37
RAE-VAR-d24	70+5	2.16	1.37	4.77	9.60	5.87	6.51	6.32	5.74
official VAR-d20	250	2.43	1.52	3.17	7.71	7.89	7.91	7.67	5.98
RAE-VAR-d20	50+4	2.46	1.55	5.20	10.46	7.10	7.34	6.89	6.43
RAE-VAR-d16, unaugmented	40+3	2.28	1.50	4.77	9.45	8.28	8.15	6.58	6.46
RAE-VAR-d16	200	2.58	1.62	5.44	10.70	7.51	7.74	7.10	6.69
RAE-VAR-d16	40+3	2.98	1.89	6.02	11.77	9.01	8.59	7.63	7.49
official VAR-d16	200	3.18	2.02	4.20	9.27	11.40	9.67	8.66	7.54
VQGAN-VAR d16 (repro.)	40+3	4.88	3.23	6.22	10.73	16.18	11.86	9.97	9.70

not a neutral judge for models that generate in DINOv2 space and should be read together with the others (our models are not favored on the MAE column, which dominates the mean through its small normalizer).

C Ablations and Failure Analyses

C.1 Ablation Details

Arms and cut. The three arms of Table 4 share the recipe of Appendix A.3 and differ in one setting each (basis file; resampling operator). Their tokenizers are stopped at exactly 105,000 optimizer steps of the 100-epoch cosine horizon; the cut was fixed from the reference arm's inline rFID curve ($3.30 \to 1.73 \to 1.21 \to 1.07$ at 25k/50k/80k/105k steps, entering the reference plateau band) before any other arm was evaluated, and the same cut is applied to every arm. The random-basis tokenizer follows a roughly parallel curve ($3.67 \to 1.98 \to 1.51 \to 1.42$ at 25k/50k/75k/100k), indicating a persistent structured-basis advantage over this interval. The generators are VAR-d16 with the recipe of §4.1. The completed epoch-8 comparison uses 13,344 optimizer steps; the completed epoch-16 extension uses 26,689 recorded steps for the random and PCA-reference arms with the same frozen 105k-step tokenizers, optimizer state and original 200-epoch learning-rate horizon. No cooldown is added at either readout. All three controlled ablation variants are evaluated at the same fixed coordinate $(1.2, 0.92)$ on 50k samples, with top-k 900. RAE-spatial-VAR is included as a system-level tokenizer ablation at its fixed $(3.0, 0.92)$ coordinate; its existing longer run supplies the matching checkpoints. The main RAE-VAR and VQGAN baseline provide additional references, with their family sampling settings unchanged across epochs. Epoch-16 readouts at the same coordinates: random basis 6.18, pooled 6.12, PCA reference 5.29, RAE-spatial-VAR 5.37, RAE-VAR 4.70, VQGAN-VAR 8.66 (inline cfg-3.0 values: random 4.60, pooled 4.40, RAE-VAR 3.98; the reference arm's inline epoch-16 point was lost to a cut). The pooled arm was moved from two to four GPUs at step 22,000 and resumed from a mid-epoch state; the trainer counts epochs by loop index, so its epoch-16 checkpoint is at step 27,004 rather than 26,689, i.e. 315 additional steps (1.2% more training). We retain the measured result without a post-hoc correction; its exact effect on gFID has not been isolated. The 200k-step train-PCA tokenizer is a separate, length-matched experiment: continued from the reference arm's 130k checkpoint to 200k steps (stand-alone rFID 0.963 against the headline's 0.955; front-loading 40.9% / 77.2% by $k = 6$ / 7), its d16 generator achieves 6.28 and 5.36 at epochs 8 and 16 against 6.05 and 4.70 for the main 400-image basis. Matching tokenizer training length does not eliminate the gap to the main system. The basis that retains more of the encoder's variance (88% against 77% at rank 256, Appendix A.3) generates slightly worse; retained variance alone does not predict the generation ordering in this comparison, and we do not have an explanation for the direction of the difference.

Per-step readouts. Decoding the first k steps with ground-truth tokens and scoring the reconstruction (5k validation images) gives the front-loading curve of Fig. 3:

k	1	2	3	4	5	6	7	8	9
(s_k, r_k)	(1,64)	(2,64)	(3,128)	(4,128)	(6,192)	(8,192)	(10,256)	(13,256)	(16,256)
cumulative tokens	1	5	14	30	66	130	230	399	655
PCA basis (ref.)	316.0	272.3	241.0	227.5	213.1	**176.5**	**56.6**	9.55	5.68
random basis	296.8	285.2	253.5	242.8	222.6	210.9	90.3	12.2	6.64
pooled resampling	272.1	245.6	229.4	221.2	203.3	140.5	58.6	19.1	6.63
RAE-VAR (headline, 200k steps)	309.5	275.6	241.7	225.2	208.4	183.3	89.9	9.57	5.23

The pooled arm shows the opposite failure from the random basis: because area pooling yields pixel-plausible thumbnails, its curve sits *below* the Fourier reference through $k = 6$ (49.6% of its total drop against 45%), but the aliased fine residual is charged at the fine steps, where it is twice as bad (19.1 against 9.6 at $k = 8$) and its endpoint matches the random basis (6.6). Under generation it achieves 7.99 gFID at epoch 8 (Table 4): the aliased fine residual costs 1.6 gFID for 0.27 rFID, a proportional loss, whereas the random basis costs 4.0 gFID for the same rFID. The headline tokenizer, whose 400-image basis retains about half of the achievable ordering gain (Appendix A.3), front-loads between the two basis arms (41.5% / 72.2% of its total drop by $k = 6$ / 7, against 45% / 84% and 30% / 71%); its lower endpoint reflects its longer training (200k against 105k steps), and no generation ordering is read off this row.

Per-scale token entropy (bits, $\log_2 V = 12$) for the PCA / random arms is $5.90/8.55$ at 1×1, $9.94/10.82$ at 2×2, $8.39/8.74$ at 3×3, $10.66/11.15$, $9.69/9.47$, $11.34/11.54$, $10.80/9.31$, $11.59/11.73$ and $11.51/11.71$ at the remaining scales (token-weighted means 11.22 and 11.14; the pooled arm achieves 11.39, the least concentrated of the three); distinct-code usage over 5k images is $11.5/53.8/78.9/98.4/99.9\%$ (PCA) against $27.5/71.2/70.8/96.2/95.3\%$ (random) and $15.1/58.2/85.0\%$ at the first three scales for the pooled arm, $\geq 99.9\%$ thereafter, with all 4096 codes used and none dead in any of them. The generators' inline gFID at cfg 3.0 achieves $44.9/19.9/9.5/4.8$ (PCA), $61.4/24.9/11.2$ (random; its epoch-8 point was lost to the cut) and $73.6/23.2/8.8/5.8$ (pooled) at epochs 1/2/4/8, so the pooled arm briefly overtakes the reference at epoch 4 under the training-time setting and falls back by epoch 8; their teacher-forced cross-entropy at 13.3k steps is 5.20 / 5.12 / 5.15 nats (token accuracy 18.3% / 17.6% / 16.7%), which orders none of them by generation.

A measurement note. The per-step and 50k rFID readouts above are computed by a stand-alone evaluator that rebuilds the tokenizer from its checkpoint. An earlier version of that evaluator did not forward the resampling operator and silently rebuilt every Fourier tokenizer with pooling; because resampling has no learnable parameters, no state-dictionary key was missing and no warning was raised. It inflated rFID by 0.2–0.3 (the headline tokenizer read 1.17 instead of 0.955) and distorted every per-step curve, while leaving generator training and every gFID untouched, since those paths construct the tokenizer correctly. All tokenizer-level numbers in this paper were re-measured after the fix, and inline (training-time) and stand-alone rFID values are never mixed within a comparison (they differ by ≈ 0.05 from sharding).

C.2 Results Without Training Augmentation

Every generator in the main text is trained with VAR's crop augmentation. The earlier runs of this project were not, and we report them here because the design ladder was first established on them. They are not comparable to the main-text numbers for two reasons. First, without augmentation the generator memorizes the training set (Appendix C.3), and FID against a training-set reference rewards memorization, so every number below is optimistic by an amount that grows with depth. Second, these runs were sampled at a per-model calibrated (cfg, top-p) optimum rather than at one fixed coordinate per family, which favors the systems whose optimum sits in the corner of the grid. The ordering they established, spatial $<$ Fourier staircase, both ahead of the VQGAN baseline, is reproduced under augmentation in §4.2 and §4.4, and the augmented gaps are the ones we claim. The pooled-staircase row below is a 200k-step tokenizer on the main basis and is not the 105k-step pooled arm of Table 4; under augmentation and at a fixed coordinate that arm lands slightly above RAE-spatial-VAR (7.99 against 7.58), so the pooled-staircase-beats-spatial reading of this table is not reproduced.

Table 10: Unaugmented milestone curves on the identical VAR-d16 backbone (50k gFID, per-milestone best cfg; †fixed cfg 5, conservative). Superseded by Table 2 and Fig. 1; retained for provenance.

system (unaugmented, VAR-d16)	rFID	ep1	ep2	ep4	ep8	ep16	ep40	40+3
VQGAN-VAR	3.39	—	—	—	9.02	6.87	5.43	4.81
RAE-spatial-VAR	1.094	88.8	37.2	13.2	6.40	4.72	4.07	3.33
rank-only + spatial warm-up (Appendix C.4)	0.977	63.9	27.7	10.4	7.89	6.71	5.19	4.59
staircase, pooled resampling	$\sim$1.2	48.5†	18.9†	7.60†	5.48†	4.45†	2.96	2.44
staircase, Fourier resampling	0.954	—	—	—	4.61	3.75	2.76	2.28

The unaugmented deeper runs read 1.93 at d20 (40+3, optimum $(2.5, 0.90)$) and 2.08 at d24 (40+3, optimum $(2.5, 0.88)$), respectively below and above the released d20 and d24 checkpoints (2.43, 2.00); their train–val cross-entropy gaps of 1.03 and 2.06 nats (Appendix C.3) show the margin to be memorization, and under augmentation the same depths read 2.46 (50+4) and 2.16 (70+5).

C.3 Memorization Without Augmentation

We measure the teacher-forced cross-entropy of each generator on 25k training and 25k validation images (nats per token, averaged over the 655 or 680 tokens; no guidance).

Table 11: Train–validation cross-entropy gap of RAE-VAR generators with and without VAR's crop augmentation (nats). Unaugmented, the gap grows with depth and the validation loss *rises* while the training loss collapses; augmented, the gap is flat at 0.07–0.10.

generator	val CE	train CE	gap	val acc.	per-scale gap $s_1 \ldots s_9$
d16, unaugmented, 40+3	4.370	3.865	**0.505**	20.9%	.05 .29 .61 .77 .63 .74 .44 .48 .45
d16, augmented, 200 ep	4.018	3.951	**0.067**	23.7%	.05 .18 .17 .14 .12 .11 .08 .06 .04
d20, unaugmented, 40+3	4.522	3.488	**1.034**	20.0%	.04 .42 1.29 1.72 1.20 1.52 .87 1.00 .94
d20, augmented, 50+4	4.035	3.966	**0.069**	23.6%	.05 .18 .17 .14 .12 .11 .08 .07 .04
d24, unaugmented, 70+5	4.915	2.856	**2.059**	18.2%	.05 .52 2.21 3.18 2.00 2.80 1.62 2.06 2.01
d24, augmented, 70+5	3.941	3.843	**0.098**	24.3%	.04 .25 .28 .23 .18 .17 .11 .09 .05

Without augmentation the gap grows with depth ($0.51 \to 1.03 \to 2.06$), the training loss falls below 3 nats at d24 while the validation loss rises ($4.37 \to 4.52 \to 4.92$), and the per-scale profile shows the memorization concentrated in the middle scales (s_3–s_6), exactly where the unaugmented gFID gains appeared. Deeper unaugmented models are therefore worse on held-out images while their FID against the training-set reference improves. With augmentation the gap is flat, validation loss improves or holds with depth, and the same-schedule pairs read $0.505 \to 0.045$ (d16, 43 epochs), $1.034 \to 0.061$ (d20, 40+3) and $2.059 \to 0.098$ (d24, 70+5). Over the 200-epoch d16 run the augmented gap grows only from 0.045 (epoch 40) to 0.067 (epoch 200).

C.4 Failure Taxonomy: Rank-Only Schedules

The experiments in this section predate the augmentation of Appendix C.2 and were sampled at per-model calibrated optima; the ordering they show is structural and does not depend on either.

The collapse and its diagnosis. Pure-compaction schedules (rank-only coarse-to-fine, every step at 16×16, 1024 tokens) reach near-best reconstruction (rFID 0.93–1.05) yet collapse at generation (gFID 57.7–72). Progressive teacher-forcing isolates the mechanism: giving the generator the *oracle* first rank band produces photorealistic rollouts, while the freely-sampled first band produces texture soup — the first band is high-entropy, weakly class-determined, and must be predicted nearly unconditionally. The failure is structural (realizability), not capacity.

Band-truncation analysis. Prepending a spatial warm-up pyramid (rank-only + spatial warm-up, 1109 tokens) repairs the collapse but still trails the plain spatial pyramid (Table 10). Decoding the same rollouts at cumulative band k (GT side: per-step rFID with real tokens) shows why:

Table 12: Band truncation of the rank-only + warm-up schedule: the final full-rank residual band improves reconstruction with ground-truth tokens but contributes *nothing* under generation — 23% of the tokens are dead at sampling time. Rank bands transfer under AR up to $r \approx 128$; only fine-rank residual at full spatial resolution is unrealizable.

cumulative	GT rFID(5k)	Δ	GEN gFID(10k)	Δ
warm-up (32)	63.0	—	79.4	—
+band 64	10.0	−53	13.9	−66
+band 128	6.3	−3.7	9.07	−4.8
+band 256	5.4	−0.9	9.02	−0.04

This is the token-level version of the dissociations in §4.5: reconstruction lives partly in tokens the AR cannot materialize. It directly motivates the staircase: allocate rank jointly with scale so that late steps carry fine-*spatial* residual (the realizable kind), not fine-rank residual at full resolution.

The teacher-forced accuracy trap. The rank-only + warm-up generator reaches roughly $3\times$ the token accuracy of the spatial baseline yet trails it in gFID at every epoch past 8 — teacher-forced predictability and rollout quality diverge, the same dissociation that the random-basis arm of §4.5 shows at the level of cross-entropy.

C.5 Continuous RAE-VAR Fails at Matched Budget

Direct continuous next-scale regression on RAE features diverges at rollout: the statistics of generated latents drift out of the decoder's training domain, and per-scale errors compound over the ten-scale trajectory—the classic exposure-bias failure, here fatal rather than merely degrading. A FlowAR-style flow-matching scale-wise head fares no better at matched budget, reaching gFID $\approx$127 at epoch 16 on RAE features, versus $\approx$5 for our discrete path at matched epochs. We are explicit about what this evidence does and does not show. The same head also drastically underperforms on *VAE* latents (65.6 at epoch 30 with a 172M head)—far from what FlowAR reports at its much longer published schedules—so our negative results cannot separate representation-space difficulty from implementation and budget: the supportable claim is that *our continuous implementations fail in both spaces at this budget*, not that continuous prediction on representations is impossible. What the experiments do establish is a practical asymmetry: at the matched budget the discrete route works immediately and well, while both continuous variants are far from usable. We therefore pursue the discrete route here, and regard well-tuned continuous heads on representations (Shao et al., 2025) as future work rather than refuted.

D Representation Structure and Compaction

D.1 Why the Rank Axis: Scale Structure of the Representation

The compaction ordering of §3.2 is a property of the representation, and Table 13 shows where it comes from. Measured on 5k validation images at the encoder output (16×16 patch tokens), DINOv2 features carry $28.6\times$ more power in the low spatial-frequency half than in the high half, against $1.95\times$ for VAR's VQGAN pre-quantization features, and they spread variance over far more channels: the top 8 principal components hold 17% of the variance (34% for VQGAN) and 90% of the variance needs 67% of the channels (88% for VQGAN, whose 32 channels are all in use). A semantic feature map is thus spatially smooth and channel-rich, which is exactly the regime in which a spatial thumbnail is a poor coarse code and a channel prefix is a good one. Across pretrained encoders the smoothness is specific to DINOv2: LeJEPA-S, SigLIP2-B and MAE-B have a similar or lower channel compaction but a 3–4$\times$ lower low/high spatial-frequency ratio, so the staircase's assumption that coarse spatial scales carry content is most strongly satisfied by the encoder we deploy; whether the layout transfers to smoother-or-not encoders is the natural next test. (The two-NN intrinsic dimension in the table is reported for completeness; it is an unreliable estimator at this feature dimensionality and both spaces have a maximum-likelihood intrinsic dimension of $\approx$20–26.)

Table 13: Scale structure of frozen encoder outputs on 5k ImageNet-val images: cumulative variance captured by the top-k principal channels, ratio of low- to high-spatial-frequency power, number of principal channels needed for 90% / 95% of the variance, and two-NN intrinsic dimension.

encoder	C	cum. variance @32 / 64 / 128 / 256	low/high spatial-freq. power	PCA rank @90% / 95%	two-NN id
DINOv2-S (main)	384	28.4 / 43.3 / 63.6 / **88.0**%	**27.4**	268 / 313	6.7
LeJEPA-S	384	25.8 / 40.3 / 60.6 / 86.1%	8.9	280 / 323	14.4
SigLIP2-B	768	23.9 / 35.0 / 50.1 / 69.4%	7.0	497 / 600	12.0
MAE-B	768	17.8 / 27.8 / 42.5 / 62.8%	6.8	542 / 632	18.4
VQGAN (VAR's tokenizer, pre-quant.)	32	—	1.95	28 / —	27.1

D.2 Decode-Grounded Axis Comparison

Table 1 of the main text truncates at matched kept-scalar budgets. Table 14 reports the underlying measurement on the frozen DINOv2-B RAE with its official ViT-XL decoder (5k validation images, so absolute rFID is inflated relative to 50k and only the relative shape matters): keeping all 256 positions and the top-k principal channels (channel axis) against keeping all channels and the cumulative spatial scales $1..s$ of VAR's schedule (spatial axis), each decoded and scored against the originals. At every matched reconstruction level the channel axis needs 3–6$\times$ fewer retained scalars (rFID $\approx$19: 33k scalars along channels against 196k along scales; $\approx$10: 53k against $\approx$250–320k; $\approx$6–7.6: 98k against 326k). Two further readings shaped the design: the decoder is sensitive to low-variance principal components, so the useful channel budget is $\approx$200–384 rather than the tens that an intrinsic-dimension estimate suggests (hence $C = 256$); and the fine spatial scales, although they hold little latent energy, are perceptually critical ($16.1 \to 7.6 \to 4.35$ from scale 10 to 13 to 16), so the schedule keeps the full spatial ladder and reallocates rank within it rather than dropping scales.

Table 14: Decoded rFID (5k) after truncating the frozen DINOv2-B RAE latent along the channel axis or the spatial axis, with the number of retained scalars. Full reconstruction achieves 4.35 on this 5k protocol.

channel axis: top-k principal channels				spatial axis: cumulative scales $1..s$			
k	cum. var.	rFID	scalars	s	tokens	rFID	scalars
16	0.115	185.5	4 096	1	1	301.2	768
32	0.182	121.7	8 192	2	5	267.0	3 840
64	0.281	57.6	16 384	3	14	249.3	10 752
96	0.359	29.0	24 576	4	30	225.8	23 040
128	0.426	19.1	32 768	5	55	171.3	42 240
206	0.557	10.2	52 736	6	91	121.6	69 888
384	0.766	6.0	98 304	8	155	47.4	119 040
768	1.000	**4.35**	196 608	10	255	16.1	195 840
				13	424	7.6	325 632
				16	680	**4.35**	522 240

E SEMANTIC INFORMATION IN TOKEN PREFIXES

E.1 CLASS INFORMATION AND RECOGNIZABILITY OF TOKEN PREFIXES

Latent-prefix probes (route B). We train a separate linear classifier for every reconstruction prefix of each tokenizer, with 100,000 ImageNet training images and evaluation on all 50,000 validation images. RAE-VAR prefixes are globally averaged into 384-dimensional features; VQGAN-VAR prefixes use global averaging (32 dimensions) or adaptive 4×4 average pooling followed by flattening (512 dimensions). Each feature coordinate is standardized with its training-set mean and standard deviation, also applied at validation time. Probes are trained for 12 epochs with AdamW, learning rate 2×10^{-3}, weight decay 10^{-3}, batch size 4096, cosine learning-rate decay, and seed 0. Unlike the earlier exploratory probe trained only on unquantized features, each probe here is trained and evaluated on the corresponding quantized prefix, avoiding that particular distribution mismatch. The readouts have different dimensions and pooling operators; their accuracy measures linearly accessible class information under these protocols, not all information present in the tokens. They evaluate reconstruction prefixes, not generated prefixes or changes across generator-training checkpoints.

Table 15: Absolute validation accuracy of the per-prefix probes in Fig. 4. Spatial scale s denotes an $s \times s$ step. Token budgets coincide through scale 4 and differ thereafter. Results are single-run measurements without confidence intervals.

	Cumulative tokens		Linear-probe Top-1 (%)		
Scale	VAR	Ours	VAR (mean)	VAR (4×4)	Ours (mean)
1	1	1	1.95	2.08	5.59
2	5	5	3.87	5.76	18.64
3	14	14	5.35	10.87	31.58
4	30	30	5.97	13.70	40.28
5	55	—	6.73	15.89	—
6	91	66	6.85	16.54	47.16
8	155	130	6.95	16.81	52.14
10	255	230	7.17	17.23	55.39
13	424	399	7.31	17.57	57.83
16	680	655	7.30	17.56	59.46

For reference, independently trained probes on unquantized RAE features and pre-quantization VQGAN latents achieve 71.22% (RAE mean pooling), 7.60% (VQGAN mean pooling), and 17.94% (VQGAN 4×4 pooling). These are separate representation readouts, not theoretical ceilings, and no curve is normalized by them. The early RAE-VAR probe accuracy is substantially above chance while pixel classification remains near chance, indicating that linearly accessible class information need not yet be expressed as a recognizable decoded image. The probe and pixel evaluations use 50k and 10k validation images respectively, so the comparison is between aggregate readouts rather than a paired per-image analysis.

Decoded-image classification (route A). Reconstruction uses 10 validation images per ImageNet class (10,000 total), selected in sorted file order, with the same images for both tokenizers. Generation uses 10 samples per class, the released VAR-d16 and our 200-epoch RAE-VAR-d16, at $(\mathrm{cfg}, \mathit{top}{-}\mathrm{p}) = (1.5, 0.96)$ and $(1.2, 0.92)$ respectively, top-k 900, and base seed 0. Each cumulative prediction is decoded to pixels, clamped to $[0, 1]$, resized to 224×224 with antialiased bilinear interpolation, and normalized with ImageNet channel statistics. The evaluator is the frozen `dinov2_vitl14_lc` classifier with its default pretrained linear head. Top-1 targets are the source class for reconstruction and conditioning class for generation. Original input images score 86.59%.

Both reconstruction and generation prefixes begin near chance under the pixel classifier, with VAR ahead at coarse spatial scales and RAE-VAR becoming recognizable at later scales. This does not contradict class information being accessible to a latent probe earlier: the decoder and external classifier impose a different readout. We do not interpret these curves as evidence of uniformly faster semantic acquisition or compare their slopes after normalization to separate final accuracies.

Table 16: Absolute pixel-classification Top-1 accuracies underlying Fig. 4 (reconstruction panel) and Fig. 5. A dash denotes our absent 5×5 step. These are single-run measurements without confidence intervals.

Scale	Cumulative tokens VAR	Cumulative tokens Ours	Recon. Top-1 (%) VAR	Recon. Top-1 (%) Ours	Gen. Top-1 (%) VAR	Gen. Top-1 (%) Ours
1	1	1	0.10	0.16	0.12	0.21
2	5	5	0.26	0.15	0.23	0.15
3	14	14	0.48	0.22	0.36	0.26
4	30	30	0.72	0.24	0.64	0.30
5	55	—	1.37	—	1.58	—
6	91	66	3.54	0.87	4.64	0.76
8	155	130	16.87	6.74	18.33	4.68
10	255	230	46.57	46.03	48.29	43.18
13	424	399	75.05	78.50	74.47	82.72
16	680	655	84.35	82.67	85.64	88.71

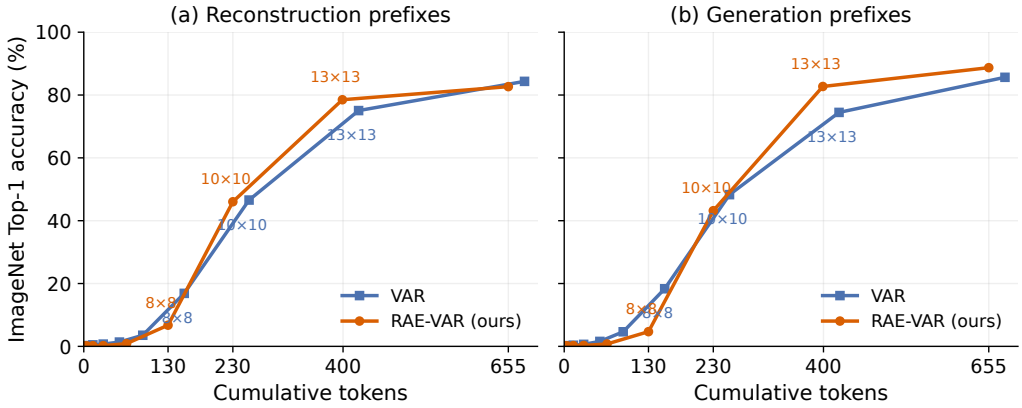

Figure 5: **Pixel recognizability during reconstruction and generation.** Absolute Top-1 accuracy of the frozen DINOv2-L classifier on decoded prefixes. Labels indicate spatial scales. These complementary readouts measure recognizability after decoding, rather than class information directly accessible in the latent prefix.

F Qualitative Reconstructions and Samples

F.1 Per-Scale Decodes on Images

The following figures show cumulative decoded prefixes rather than isolated residuals. Figure 6 uses generated tokens, Fig. 7 uses ground-truth reconstruction tokens, and Fig. 8 provides additional generated examples.

F.2 Qualitative Samples

Figure 9 shows uncurated 8×8 sample grids (classes 0–63, one sample each, identical seeds) at each model's family sampling setting, taken from the same 50k sample sets that produced Table 2; the released checkpoints are sampled at their own calibrated optimum.

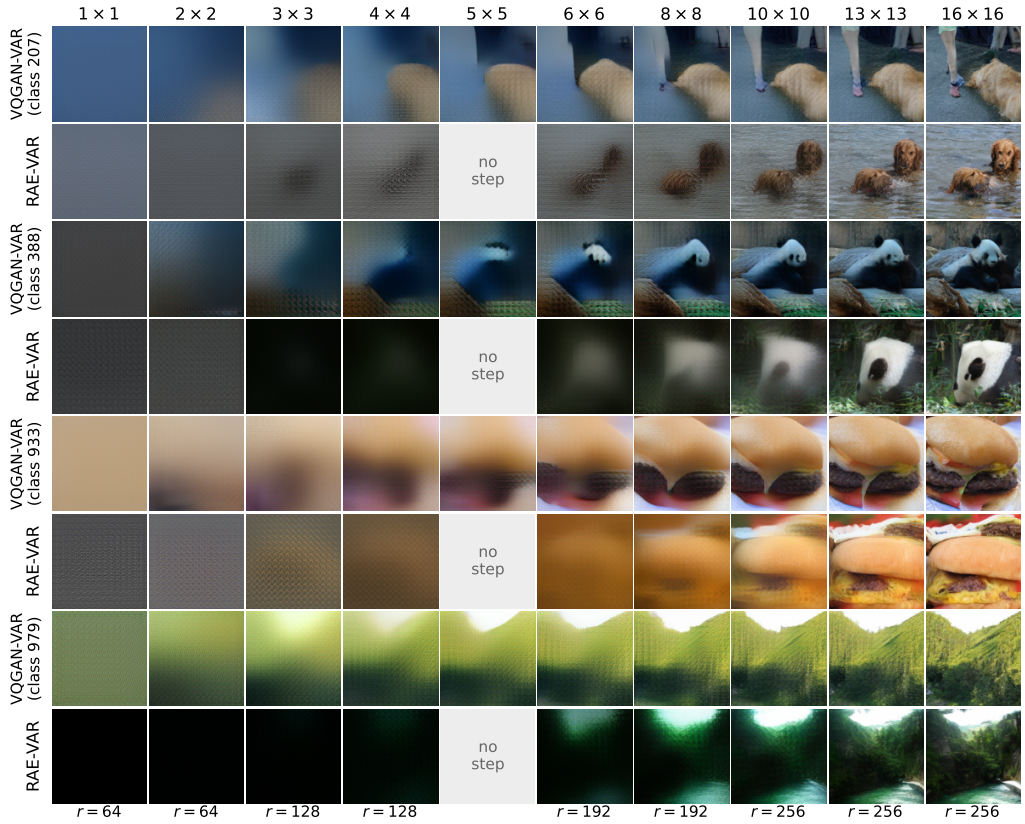

Figure 6: **What each scale generates.** The running prediction decoded after every scale, for the released VQGAN-VAR-d16 (200 epochs, $(1.5, 0.96)$) and RAE-VAR-d16 (200 epochs, $(1.2, 0.92)$) on the same classes and seed; columns are the spatial side of the step, the staircase has no 5×5 step, and its rank is given under each column. VQGAN-VAR commits to a low-pass pixel layout from the first step and refines it; RAE-VAR's coarse steps carry rank-limited semantic content that the decoder cannot yet render, and the image snaps in at the 8×8–13×13 steps, once rank reaches 192–256, which is where the curves of Fig. 3 separate.

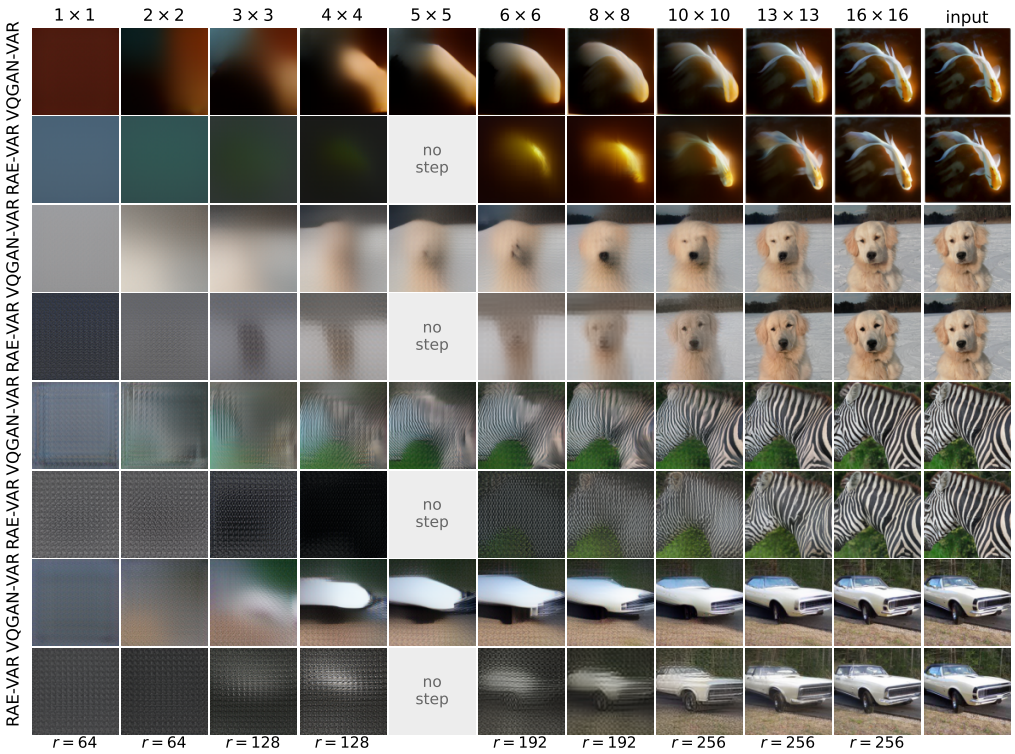

Figure 7: **What each scale contributes**, decoded with ground-truth tokens after the first k steps on the same validation images (columns: spatial side of step k; the staircase has no 5×5 step, and its rank r_k is given under each column). VAR's VQGAN pyramid writes a low-pass pixel thumbnail from the first step on, so its coarse scales are blurry but pixel-plausible. The staircase's coarse steps carry rank-limited *semantic* content ($r \leq 128$) that the pixel decoder cannot yet render, so they decode to little; the image snaps in at the 8×8 and 10×10 steps where rank reaches 192–256, which is exactly where the cumulative-rFID curves of Fig. 3 separate, and it is essentially complete by 13×13 (399 of 655 tokens). The coarse steps are therefore not thumbnails but the conditioning prefix for the scales that render.

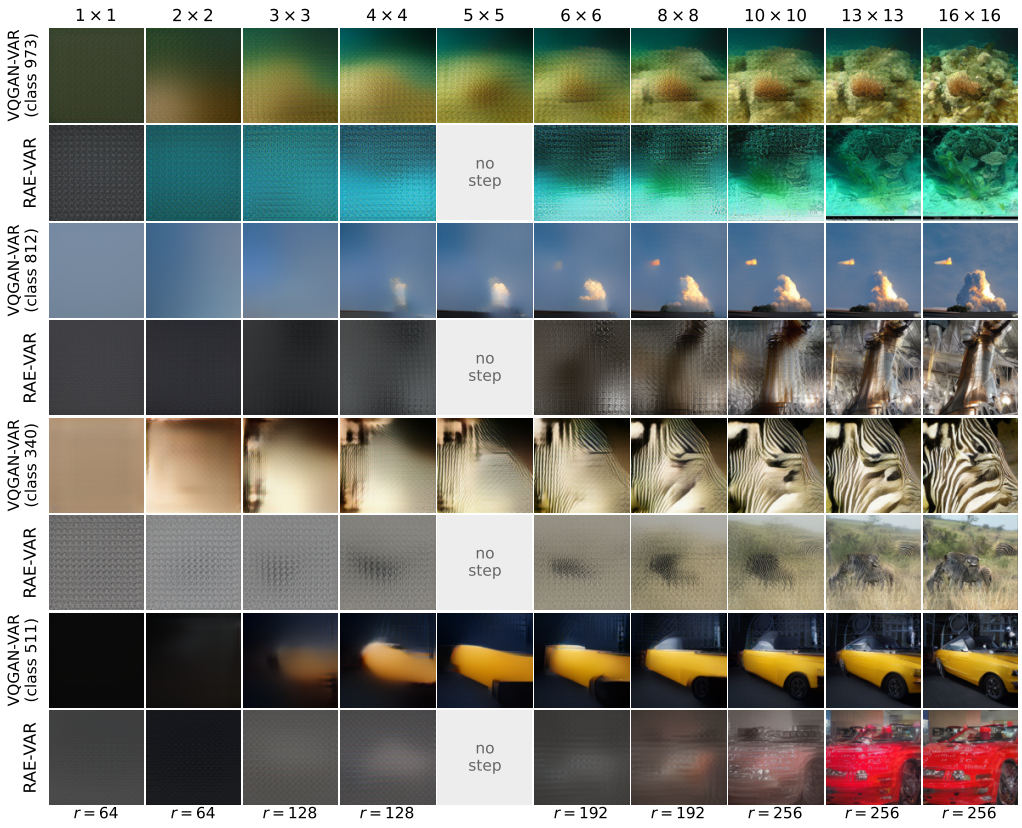

Figure 8: Further generation rollouts in the format of Fig. 6 (classes 973, 812, 340, 511; a different seed), decoded after every scale for the released VQGAN-VAR-d16 and RAE-VAR-d16 at their family sampling settings.



(a) RAE-VAR-d16, 200 ep, gFID 2.58 (b) released VAR-d16, 200 ep, 3.18 (c) VQGAN-VAR d16 (repro.), 40+3



(d) RAE-VAR-d20, 50+4, 2.46 (e) RAE-VAR-d24, 70+5, 2.16 (f) released VAR-d30 (2B), 350 ep, 1.96

Figure 9: Uncurated class-conditional samples (classes 0–63). Top row: the d16 comparison at equal schedule (a, b) and at matched budget (c). Bottom row: our deeper models next to the largest released VAR.