WITHINVAE: COMPACT VISUAL AUTOENCODING WITH SELF-DISTILLED SEMANTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual autoencoders must balance reconstruction fidelity, compactness, and semantic structure. Existing semantic tokenizers typically inherit semantics from pretrained recognition encoders, whose invariances can discard the appearance detail required for inversion. We show that this inheritance is unnecessary: semantic structure can be learned inside an autoencoder's own reconstructive representation. We introduce WithinVAE, a compact variational visual autoencoder with a sequential bottleneck, DINO/iBOT-style self-distillation from an exponential-moving-average copy of the autoencoder itself, and position-dependent token masking to achieve a 4-fold compression ratio, which is free of pretrained representation encoder or external semantic target. On ImageNet-1K at 256×256, WithinVAE improves all four reported reconstruction metrics over RPiAE including PSNR by 1.36 dB, while using half its latent scalar capacity. Under frozen evaluation, the same latents improve linear probing by 6.5 points over RPiAE and 5.1 points over our reconstruction-only baseline, attributing the semantic gains to self-distillation rather than architecture or external supervision. An auxiliary latent-to-latent model reproduces the ordered-compression trend, and a pixel-to-latent model transfers the objective to a pretrained generative space. These results separate the origin of semantic supervision from its use: a compact visual autoencoder can organize semantic structure without borrowing a recognition representation.

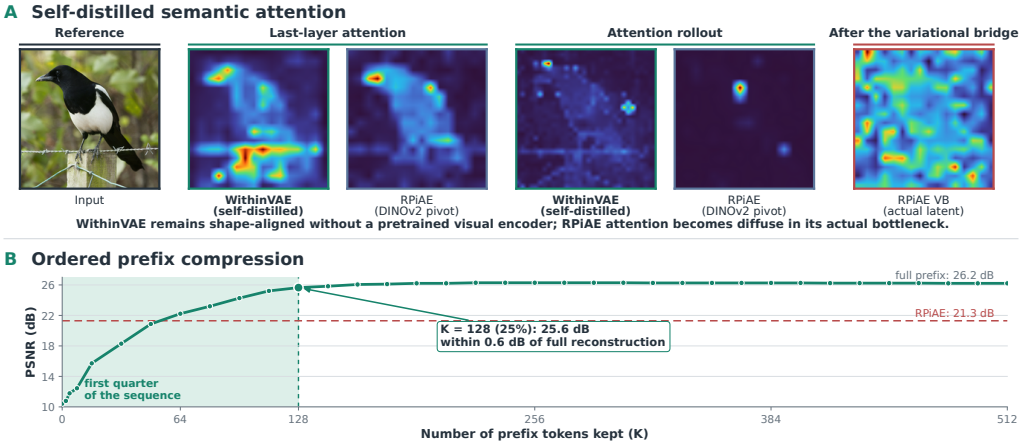

Figure 1: Two diagnostics of self-distilled semantic autoencoding. **A:** Under matched display settings, WithinVAE attends to object extent without any pretrained recognition encoder, while RPiAE totally fails in providing reasonable attention after borrowed DINOv2 (rollout and actual latent after variational-bridge). **B:** Prefix decoding of the 512×96 latent-to-latent variant front-loads reliability: $K = 128$ retains 25.6 dB PSNR, within 0.6 dB of the full sequence, and $K = 64$ (6,144 scalars, under half of RPiAE's budget) already exceeds RPiAE's 21.3 dB.

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.

1 INTRODUCTION

Latent generative models replace pixels with a lower-dimensional model-facing representation. The tokenizer that defines this interface determines reconstruction fidelity, latent dimensionality, and the geometry seen by every downstream generator. Ideally, its latents preserve visual detail, expose semantic structure, and remain compact. These requirements are not automatically compatible: stronger compression removes information, pixel losses favor local appearance, and recognition-oriented invariance can suppress precisely the variation needed for reversible generation (Shannon, 1959; Blau & Michaeli, 2019; Xi et al., 2026; Li, 2022; Zhang et al., 2025a).

Recent semantic tokenizers address this tension by transferring representations from pretrained vision encoders. Some align autoencoding or diffusion features to DINO- or CLIP-like targets (Yu et al., 2024b; Xu et al., 2025; Yao et al., 2025; Zhang et al., 2025b); others decode, compress, or adapt the representation encoder itself (Zheng et al., 2025; Singh et al., 2026; Gong et al., 2026; Gao et al., 2025; Chen et al., 2026a; Shi et al., 2026; An et al., 2026). These methods demonstrate the value of semantics, but they also couple the tokenizer to an understanding model whose invariances were learned for a different objective. Freezing the encoder restricts reconstruction adaptation, whereas unconstrained fine-tuning can erase the inherited geometry. RPiAE addresses this compromise with a frozen pivot and a variational bridge (Gong et al., 2026). We study the complementary question: *must the semantic target come from an external representation model at all?*

Our hypothesis is that view consistency and masked prediction can organize semantics within the reconstructive representation itself. One-dimensional tokenizers provide a sequence axis on which nonuniform corruption can order information (Yu et al., 2024a; Miwa et al., 2025), while EMA teacher–student objectives organize global and patch representations without labels (Caron et al., 2021; Zhou et al., 2022). Their combination need not make compression lossless or guarantee semantic ordering. It instead creates two coupled pressures: early tokens are more reliable for reconstruction, and the latent space must remain consistent across appearance-changing views.

We instantiate this idea in WithinVAE, a compact image-to-image variational autoencoder initialized from the generative One-D-Piece architecture (Miwa et al., 2025). It maps an image to a one-dimensional latent token sequence, applies a token-wise variational bottleneck, and reconstructs pixels. During training, token masking probability increases with sequence position. A separate EMA copy supplies global DINO-style targets from pooled compact tokens and local iBOT-style targets on masked image patches. "Self-distilled semantics" refers specifically to the origin of these targets: no DINO, DINOv2, or CLIP weights supervise training. It does *not* mean training every generative component from scratch.

The experiments establish three results. First, at half of RPiAE's latent scalar capacity, WithinVAE improves all four reported reconstruction metrics simultaneously (rFID, PSNR, LPIPS, SSIM), and its 128-token prefix remains stronger than RPiAE in PSNR and SSIM at one quarter of that capacity. Second, frozen linear-probe and nearest-neighbor evaluations attribute the semantic gains to the self-distillation objective itself: it improves linear probing by 5.072 points over our reconstruction-only baseline and by 6.482 points over RPiAE, with simultaneous nearest-neighbor gains ruling out collapse. Third, the representation is directly modelable by a latent diffusion transformer—a 675M LightningDiT trained on WithinVAE latents already exceeds RPiAE on recall without guidance—although the remaining generation metrics trail RPiAE, which we analyze as a generator–autoencoder compatibility question rather than a failure of the representation. Our claims are accordingly improved reconstruction at lower capacity, capacity-efficient prefix decoding, and frozen semantic quality independent of external recognition features; state-of-the-art generation is not among them.

Our contributions are:

- We formulate *self-distilled semantic autoencoding*, in which semantic supervision is produced online by the autoencoder's own EMA copy rather than transferred from a pretrained recognition encoder.
- We instantiate it as a compact variational autoencoder coupling position-dependent token masking with global–local EMA self-distillation, and show it improves all four reconstruction metrics over RPiAE at half the scalar capacity (+1.36 dB PSNR on 256×32 model), while its 128-token prefix still exceeds RPiAE in PSNR and SSIM at one quarter.

- We attribute the semantic gains through controlled component ablations (+5.1 linear-probe points over the reconstruction-only baseline, with reconstruction improving at the same time), and verify that the objective transfers across pixel and pretrained generative-latent interfaces, including prefix decoding that retains 25.6 dB PSNR at $4\times$ fewer tokens without retraining.

2 Related Work

Visual tokenization and compact sequences. VQ-VAE and VQGAN established discrete image codes for generative modeling (van den Oord et al., 2017; Esser et al., 2021), while latent diffusion popularized continuous spatial latents (Rombach et al., 2022). Modern tokenizers improve codebooks, architectures, and image–video unification (Yu et al., 2022; Luo et al., 2024; Weber et al., 2024; Wang et al., 2024). A complementary line removes the fixed two-dimensional grid. SEED learns compact tokens for multimodal models (Ge et al., 2024), TiTok compresses images with latent queries (Yu et al., 2024a), and One-D-Piece orders a one-dimensional sequence with Tail Token Drop (Miwa et al., 2025). Rateless autoencoding similarly removes later dimensions stochastically (Koike-Akino & Wang, 2020; Chen et al., 2026b). We retain the query-based sequence topology, but use continuous variational tokens and a position-dependent reliability profile. Our method achieves excellent compression as reported in token count, width, and total scalar count, while maintaining reconstruction quality.

Representation priors for generation. Self-supervised and vision–language encoders provide useful semantic geometries (Radford et al., 2021; Oquab et al., 2024). REPA and REPA-E align diffusion features without replacing the tokenizer (Yu et al., 2024b; Leng et al., 2025); ReaLS, VA-VAE, and PS-VAE align the autoencoding latent (Xu et al., 2025; Yao et al., 2025; Zhang et al., 2025b). RAE decodes frozen representation features, FAE compresses them, SVG separates semantic and residual features, and AlignTok progressively adapts a foundation encoder (Zheng et al., 2025; Singh et al., 2026; Gao et al., 2025; Shi et al., 2026; Chen et al., 2026a; Shao & Li, 2026). RPiAE fine-tunes a representation-initialized encoder under a frozen pivot before learning a compact variational bridge (Gong et al., 2026). WithinVAE differs in where semantic supervision originates: its targets are produced online by the autoencoder's EMA copy.

Self-supervision inside an autoencoder. Masked autoencoding learns by reconstructing hidden content (He et al., 2022); DINO uses centered and sharpened EMA targets across views (Caron et al., 2021); iBOT extends this objective to masked patches (Zhou et al., 2022). MAETok combines masked reconstruction with external feature targets for diffusion-oriented tokens (Chen et al., 2025). We borrow the optimization forms of DINO and iBOT, not their pretrained representations. The global branch directly supervises pooled compact tokens, whereas the local branch regularizes the shared image-patch encoder. This distinction matters when local masked prediction can shape the bottleneck through the shared encoder, but is not itself a loss on individual compact tokens.

3 Method

3.1 WithinVAE: compact variational autoencoder

Figure 2 gives the complete training system. A single encoder and bottleneck feed two paths that meet only through shared weights: an upper reconstruction path that must invert the image from an ordered, partially corrupted token sequence, and a lower semantic path that matches student outputs against an EMA copy of the same autoencoder. The subsections below specify these paths in turn.

Given an image $\boldsymbol{x} \in [0,1]^{3\times H\times W}$, an encoder E_θ aggregates image patches with N learned queries, a bottleneck B_ψ produces continuous tokens $\boldsymbol{z} \in \mathbb{R}^{N\times d}$, and a decoder D_ϕ reconstructs $\hat{\boldsymbol{x}}$. Our main model uses $N = 256$, $d = 32$, and ViT-L encoder and decoder backbones. This is the reconstruction path along the top of Figure 2. A diagonal posterior is predicted independently at each token:

$$(\boldsymbol{\mu}_i, \log \boldsymbol{\sigma}_i^2) = B_\psi^{\text{enc}}(E_\theta(\boldsymbol{x})_i), \qquad \boldsymbol{z}_i = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{1}$$

The posterior mean is used at evaluation. Training combines pixel L1, ConvNeXt perceptual, KL, and generator-adversarial terms with a LeCam-regularized discriminator. Adding semantic self-distillation

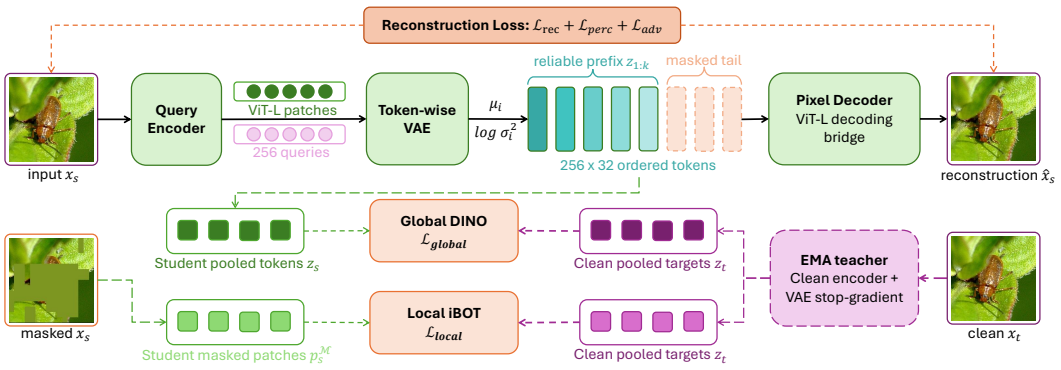

Figure 2: WithinVAE training. The reconstruction path maps an image x_s through a query encoder (ViT-L patches and 256 latent queries) and a token-wise VAE to a 256×32 ordered token sequence; position-dependent masking keeps a reliable prefix $z_{1:k}$ and corrupts the masked tail before the ViT-L pixel decoder produces $\hat{x}_s$, supervised by a reconstruction loss $\mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{perc}} + \mathcal{L}_{\text{adv}}$. The semantic branches below match mean-pooled compact tokens via global DINO ($\mathcal{L}_{\text{global}}$) and student masked patches via local iBOT ($\mathcal{L}_{\text{local}}$) against targets produced by an EMA copy of the encoder and VAE that observes a clean view x_t.

losses gives

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{g}}\mathcal{L}_{\text{global}} + \lambda_{\text{l}}\mathcal{L}_{\text{local}}. \tag{2}$$

The KL term is averaged over tokens and summed over latent dimensions. Shape-compatible encoder, decoder, latent-query, and pixel-decoder weights are initialized from One-D-Piece; the 32-dimensional boundary projections and the variational bottleneck are random. Thus our comparison isolates external *semantic* supervision, not generative pretraining.

3.2 Position-dependent token reliability

At training time, reconstruction uses independently masked tokens

$$p_i = p_{\max}\frac{i-1}{N-1}, \qquad m_i \sim \text{Bernoulli}(p_i), \qquad \tilde{\boldsymbol{z}}_i = (1-m_i)\boldsymbol{z}_i, \tag{3}$$

with $p_{\max} = 0.5$. The increasing corruption probability makes early tokens statistically more reliable than late tokens, biasing the encoder to place broadly useful information near the front of the sequence. This is the split drawn between the reliable prefix $z_{1:k}$ and the corrupted tail in Figure 2: only the surviving sequence reaches the pixel decoder. We test this predicted reliability ordering directly by decoding variable-length prefixes,

$$\hat{\boldsymbol{x}}_K = D_\phi(\boldsymbol{z}_{1:K}), \qquad 1 \le K \le N, \tag{4}$$

using the decoder's variable-length cross-attention. No retraining is performed for a new K.

3.3 Global and local semantic self-distillation

Let f_θ denote the student encoder, bottleneck, mean pool, and prototype head, and $f_{\bar{\theta}}$ its stop-gradient EMA copy, shown as the teacher branch at the lower right of Figure 2. Two views $\boldsymbol{x}_s$ and $\boldsymbol{x}_t$ of the same image produce

$$\boldsymbol{p}_s = \text{softmax}(f_\theta(\boldsymbol{x}_s)/\tau_s), \qquad \boldsymbol{p}_t = \text{softmax}((f_{\bar{\theta}}(\boldsymbol{x}_t) - \boldsymbol{c})/\tau_t), \quad \mathcal{L}_{\text{global}} = -\sum_j \boldsymbol{p}_{t,j}\log \boldsymbol{p}_{s,j}. \tag{5}$$

The head predicts 8,192 normalized prototypes from the mean of the same compact tokens passed to the reconstruction decoder before token masking. Centering, teacher sharpening, and an EMA momentum scheduled from 0.996 toward 1 discourage collapse.

For local structure, block masks cover 10–50% of the encoder's 16×16 image-patch grid. The student observes a mean-filled masked view and the EMA teacher the corresponding clean view. For masked patch set $\mathcal{M}$,

$$\mathcal{L}_{\text{local}} = -\frac{1}{|\mathcal{M}|} \sum_{u \in \mathcal{M}} \sum_{j} \boldsymbol{p}_{t,j}^{(u)} \log \boldsymbol{p}_{s,j}^{(u)}. \tag{6}$$

The local branch reuses the encoder and prototype head but bypasses the compact-token mean pool, which is why Figure 2 draws $\mathcal{L}_{\text{local}}$ from the student masked patches rather than from the token sequence. Its role is to preserve spatially grounded patch information in the shared encoder while the global branch organizes the compact bottleneck.

3.4 Complementary roles of the objectives

The progressive mask and the two semantic losses in Equation 2 act at different interfaces. Position-dependent corruption changes which compact tokens are reliable to the pixel decoder, but supplies no semantic label and does not by itself require view invariance. Global self-distillation instead compares mean-pooled compact sequences across two augmented views. It directly regularizes the representation exposed to a downstream model, but pooling can discard spatial evidence. Local masked prediction supplies that missing pressure at image-patch resolution through the shared encoder. It therefore affects compact tokens only indirectly and should not be interpreted as assigning a semantic label to every latent position.

The EMA teacher is also materially different from a fixed representation pivot. Its targets co-evolve with the reconstructive model, while stop-gradient, centering, teacher sharpening, and delayed loss warm-up stabilize the teacher–student loop. Reconstruction prevents a purely invariant solution from ignoring image detail; self-distillation discourages the reconstruction space from becoming useful only to its paired decoder. This coupling motivates the central empirical questions: whether semantic gains survive frozen evaluation, whether prefix reliability appears despite independent rather than contiguous masking, and whether the resulting interface remains usable by a separately trained generator.

4 Experiments

4.1 Implementation details

We organize the evaluation around four questions. First, does semantic self-distillation preserve reconstruction while reducing the model-facing capacity? Second, are any semantic gains visible to frozen, decoder-independent probes rather than only to the training heads? Third, does the position-dependent reliability profile support prefix decoding? Finally, can the same representation be modeled by a diffusion transformer and can the objective transfer across reconstruction interfaces? Each claim below is tied to the corresponding metric; no single score is used as evidence for all four questions.

Data and optimization. We train on ImageNet-1K at 256×256 with random crops and horizontal flips. Reconstruction uses L1, ConvNeXt-S perceptual, KL, and adversarial weights of 1, 0.1, 3.75×10^{-5}, and 0.01; global and local distillation each use 0.05. KL, adversarial, and semantic terms use scheduled warm-up, and the teacher temperature increases from 0.04 to 0.07. AdamW uses a per-GPU batch size of 16 with four-step gradient accumulation, learning rate and weight decay 10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, gradient clipping at 1.0, and bfloat16.

Architecture and initialization. The 16×16 image-patch grid and 256 learned latent queries are processed by a ViT-L encoder. A token-wise MLP of hidden width 2,048 predicts the 32-dimensional posterior parameters, and a ViT-L decoder maps the sampled sequence back to pixel patches. This yields a nonspatial model interface: the decoder addresses a sequence of compact tokens rather than a fixed $h \times w$ latent grid. At evaluation we use posterior means, so stochastic sampling affects training regularization but not the reported reconstructions or frozen features.

Shape-compatible encoder, decoder, query, and pixel-head parameters are initialized from One-D-Piece. Boundary projections introduced by the narrower latent width, the variational bottleneck, and the 8,192-prototype semantic head are new. Pretrained parameters use a $0.1\times$ learning-rate multiplier,

Table 1: ImageNet-1K reconstruction at 256×256. WithinVAE improves all four metrics over RPiAE at half its scalar count, without any external representation model. Baselines are reported values from their source papers (consolidated in RPiAE Table 3); WithinVAE uses our 50K-image evaluation. Ext. RM indicates an external pretrained representation model used for semantic initialization or supervision. Scalar count is an interface-capacity proxy, not a bitrate.

Model	Ext. RM	Latent	Scalars	rFID↓	PSNR↑	LPIPS↓	SSIM↑
SD-VAE (Rombach et al., 2022)	No	$32\times32\times4$	4,096	0.61	26.9	0.130	0.736
VA-VAE (Yao et al., 2025)	Yes	$16\times16\times32$	8,192	0.27	27.7	0.097	0.779
RAE-S (Zheng et al., 2025)	Yes	256×384	98,304	0.64	18.9	0.252	0.489
RAE-B (Zheng et al., 2025)	Yes	256×768	196,608	0.57	18.8	0.256	0.483
FAE-d32 (Gao et al., 2025)	Yes	$16\times16\times32$	8,192	0.68	–	–	–
RPiAE (Gong et al., 2026)	Yes	256×64	16,384	0.50	21.3	0.216	0.525
WithinVAE	No	256×32	8,192	**0.46**	**22.66**	**0.159**	**0.593**
WithinVAE prefix	No	128×32	**4,096**	0.47	22.33	0.163	0.581

while new parameters use the full rate. The EMA teacher duplicates the autoencoder-side encoder and semantic pathway; it neither loads nor queries an external recognition network. Consequently, the experiment removes recognition-feature supervision while retaining a generative architectural initialization, a distinction maintained throughout our claims.

Reconstruction and capacity. All WithinVAE reconstruction metrics use all 50,000 ImageNet validation images under deterministic preprocessing and lossless image export. We report rFID, PSNR, LPIPS-Alex, and SSIM. For continuous latents, "scalars" means token count times channel width. It is a transparent proxy for the interface presented to a downstream model, not an entropy-coded bitrate. Baseline values are taken from the cited papers; our local RPiAE reconstruction gives 20.75 dB PSNR, consistent with rather than stronger than its reported 21.3 dB.

Frozen semantic evaluation. We mean-pool compact tokens and freeze each autoencoder. Linear probing trains one classifier with SGD for 100 epochs. For 20-NN, L2-normalized features use exact cosine neighbors and temperature-weighted voting. Same-image view consistency is 100 times cosine similarity between two independently cropped and flipped validation views, reported as mean $\pm$ standard deviation. Consistency is interpreted only alongside the non-collapsed linear and nearest-neighbor results.

Training duration and compute. The main autoencoder is trained for 10 epochs, and the downstream LightningDiT is trained for 80 epochs. All reported experiments are conducted on eight NVIDIA H100 GPUs. Reconstruction, frozen-semantic, prefix, and attention evaluations use the trained autoencoder and its EMA teacher; generation samples classes uniformly and decodes generated latents with the corresponding decoder.

4.2 Quantitative results

Table 1 reports the main comparison. Against RPiAE—the closest baseline pursuing the same compact, one-dimensional, semantic, variational interface—WithinVAE halves scalar capacity from 16,384 to 8,192 while improving all four reconstruction metrics simultaneously: rFID 0.46 vs. 0.50, PSNR +1.36 dB, LPIPS -0.057, and SSIM $+0.068$. The 128-token prefix uses only 4,096 scalars, one quarter of RPiAE's budget, and still exceeds RPiAE's full-sequence result in PSNR and SSIM. For context, SD-VAE and VA-VAE use spatial ($h \times w$ grid) latents and reach higher PSNR at comparable budgets, but they are two-dimensional interfaces and VA-VAE depends on a pretrained representation model; the comparison that controls for interface type and supervision origin is the one against RPiAE.

The trend is not limited to one prefix. Figure 1B reports the denser sweep for the auxiliary latent-to-latent model: its 64×96 prefix uses 6,144 scalars and reaches 22.23 dB PSNR, while 128×96 reaches 25.63 dB, within 0.64 dB of the full sequence. The progressive mask therefore yields useful prefix inference despite never training with contiguous tail removal.

Table 2: Reconstruction and frozen-latent semantics on ImageNet-1K. The full objective improves both fidelity and linear-probe accuracy over every partial variant. Circled indices ①–⑥ label rows for reference in the text. RPiAE reconstruction values are published; its semantic columns use our frozen-latent protocol. LP and 20-NN are top-1 accuracy (%). View is same-image cross-view cosine multiplied by 100. Dashes denote evaluations not run for intermediate ablations.

	Variant	PSNR↑	LPIPS↓	LP↑	20-NN↑	View↑
①	RPiAE	21.30	0.216	65.404	66.856	98.339 ± 1.466
②	WithinVAE reconstruction only	21.90	0.175	66.814	–	–
③	+ progressive mask	21.69	0.180	67.812	69.435	98.029 ± 1.441
④	+ mask + global	21.47	0.186	69.794	–	–
⑤	+ mask + local	22.13	0.162	70.312	–	–
⑥	**WithinVAE full**	**22.66**	**0.159**	**71.886**	**72.394**	$\mathbf{99.125 \pm 0.732}$

4.3 Ablation study

Table 2 separates architectural and semantic effects; circled indices refer to its rows. The compact architecture accounts for part of the frozen-probe gap on its own: ② exceeds RPiAE ① by 1.410 linear-probe points (66.814% vs. 65.404%) with no semantic objective. Each component then adds further—progressive masking ③ adds 0.998 over ②, the global objective ④ adds 1.982 over ③, and the local objective ⑤ adds 2.500—and the two semantic branches are not redundant, since combining them in ⑥ gains 1.574 over ⑤ and 2.092 over ④ to reach 71.886%, i.e. 5.072 above ② and 6.482 above ①.

The intermediate rows show that fidelity and semantics improve together only for the complete objective. Taken alone, masking and the global branch cost reconstruction: ③ and ④ fall 0.21 dB and 0.43 dB below ② in PSNR. The local branch ⑤ recovers fidelity, and only ⑥ is best on both axes at once (22.66 dB, 0.159 LPIPS; $+0.76$ dB and -0.016 over ②) while adding the 5.072-point probe gain. Semantic organization is therefore not purchased with reconstruction quality, even though two of the three components would individually suggest such a trade-off. Collapse is also ruled out: ⑥ improves 20-NN by 5.538 points over ① and 2.959 over ③, so its 99.125 view score is accompanied by—not substituted for—class-discriminative structure.

Figure 1A provides a controlled attention diagnostic. The displays hold the input crop, preprocessing, per-map normalization, and color map fixed, but the attention operators are necessarily architecture-specific. The WithinVAE map averages latent-query-to-patch attention across heads at encoder layer 10 (zero-based) of the EMA teacher. The RPiAE map averages its variational-bridge encoder attention over heads and bridge queries. Figure 3A extends the comparison to seven matched validation images. We do not use attention as proof of causal token contribution; the quantitative semantic claim rests on the frozen evaluations above.

4.4 Downstream generation

We train a 675M-parameter LightningDiT (Yao et al., 2025) for 80 epochs directly on WithinVAE latents; Table 3 confirms that the representation is immediately modelable by a diffusion transformer. Without guidance, WithinVAE already exceeds RPiAE on recall (0.66 vs. 0.60), indicating that the compact sequence preserves sample diversity. The remaining generation metrics trail RPiAE (e.g., gFID 9.12 vs. 2.25 unguided), and guidance narrows but does not close the gap: better autoencoder reconstruction and frozen semantics do not automatically yield a better generator. We view this as a latent-normalization and generator–autoencoder compatibility question rather than a failure of the representation, and as the clearest direction for future work.

4.5 Qualitative results

Figure 3 compares the main model against RPiAE on seven matched validation images. In the upper panel—including a thin safety pin, multiple cranes, and a small insect—WithinVAE attention concentrates on the foreground more clearly than the RPiAE bridge maps, which are more fragmented. Both remain architecture-specific, coarse diagnostics that can extend into correlated background regions; neither is interpreted as segmentation or causal evidence. The lower panel shows recognizable

Table 3: Class-conditional ImageNet generation with a 675M LightningDiT trained for 80 epochs. RPiAE values are reported by Gong et al. (2026); WithinVAE uses a 50K-sample evaluation. Guidance settings follow each implementation and are not asserted to be identical beyond the reported on/off condition.

Model	Guidance	gFID↓	IS↑	Prec.↑	Rec.↑
RPiAE	Off	2.25	208.7	0.81	0.60
RPiAE	On	1.51	225.9	0.79	0.65
WithinVAE	Off	9.12	124.3	0.69	**0.66**
WithinVAE	On	3.67	169.2	0.76	0.58

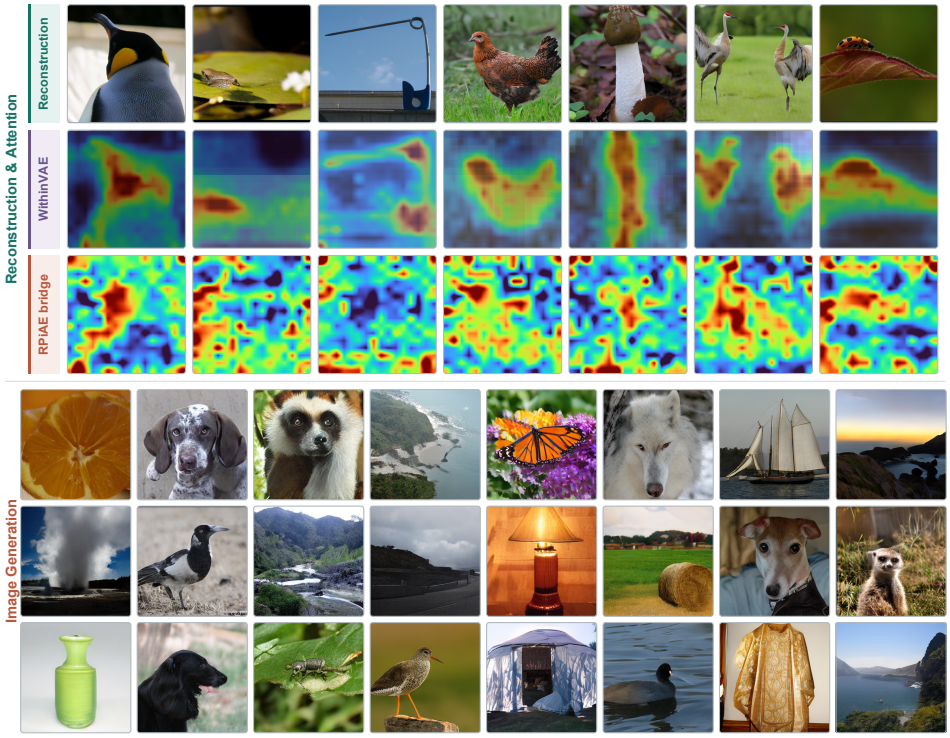

Figure 3: Qualitative results. **Up:** Matched validation examples showing WithinVAE reconstruction, WithinVAE head-averaged attention, and RPiAE variational-bridge attention averaged over heads and bridge queries. The displays share the input crop, per-map normalization, and color map, but the attention operators are architecture-specific. **Down:** Class-conditional generation samples from the 80-epoch LightningDiT. Table 3 provides the distribution-level comparison.

object and scene structure across animal, object, and landscape classes, but a selected montage cannot expose coverage failures or the frequency of malformed samples, so the 50K-sample metrics in Table 3 remain the basis for comparing generators.

4.6 Transfer across reconstruction interfaces

We test the training principle in two auxiliary systems: pixels compressed to a pretrained Qwen-Image latent, and Qwen-Image latents compressed and reconstructed in their own space (Wu et al., 2025). Table 4 reports end-to-end RGB metrics after the Qwen-Image decoder. These models do not share the main architecture and their full 512×96 bottlenecks are larger, so they test transfer of the objective rather than superiority of one architecture or budget. The transfer succeeds on the key phenotype: the latent-to-latent model preserves the prefix trend, losing only 0.64 dB from 512 to 128 tokens—evidence that ordered reliability follows from the objective rather than from one architecture.

Table 4: End-to-end reconstruction across interfaces at 256×256. Qwen rows use a pretrained generative autoencoder; metrics are computed in RGB after decoding. Scalars expose the unequal model-facing capacities.

Interface	Bottleneck	Scalars	rFID↓	PSNR↑	LPIPS↓	SSIM↑
Pixel → pixel	256×32	8,192	0.46	22.66	0.159	0.593
Pixel → pixel, prefix	128×32	4,096	0.47	22.33	0.163	0.581
Pixel → Qwen latent	512×96	49,152	0.43	23.91	0.144	0.614
Qwen latent → Qwen latent	512×96	49,152	0.40	26.27	0.131	0.714
prefix	128×96	12,288	0.42	25.63	0.136	0.698
prefix	64×96	6,144	0.48	22.23	0.160	0.579

The two interfaces answer different questions. Pixel-to-Qwen-latent reconstruction must learn both compression and the geometry expected by a fixed generative decoder, so its 23.91 dB PSNR cannot be attributed solely to the self-distillation objective. Qwen-latent-to-Qwen-latent reconstruction removes that first mapping and reaches 26.27 dB, but exposes 49,152 scalars at full length; its prefix curve is the informative result, since a $4\times$ sequence reduction moves rFID only from 0.40 to 0.42 and PSNR from 26.27 to 25.63 dB without retraining. At 64 tokens the sharper PSNR and SSIM drop shows that ordered reliability is useful but not lossless.

4.7 Analysis of Experimental Results

The ablations, the transfer results, and the generation result answer three different questions, and only together do they identify what is responsible for the gains. The generator result is the necessary negative control: frozen classification and reconstruction quality do not guarantee an easier diffusion distribution. The central claim therefore concerns the source of semantic supervision—a pretrained recognition encoder is unnecessary for the measured semantic organization. Table 2 makes this attribution causal in direction, since removing the self-distillation objective removes the semantic gain while keeping it improves fidelity at the same time, and the cross-interface experiments extend it to objective transfer, since the same recipe trains successfully on both auxiliary interfaces and the latent-to-latent model reproduces the prefix-reliability trend. Appendix A states the inferential strength and evaluation scope of each comparison.

4.8 Limitations

We further discuss limitations for concrete openings. Because capacity is counted in continuous scalars rather than entropy-coded bits, quantizing the ordered sequence would turn prefix decoding into a genuine variable-rate codec and place semantic structure on a rate–distortion frontier. Because our semantic evidence is global and classification-based, causal token interventions and dense prediction would establish what individual positions encode. Most importantly, better reconstruction and frozen semantics did not produce better generation, which isolates latent distribution geometry rather than autoencoder quality as the binding constraint on compact tokenizers, and raises the question of how much semantic structure an EMA teacher can organize before an external target becomes necessary.

5 Conclusion

In this paper we present WithinVAE, a compact visual autoencoder that learns its semantic targets from the same generative system that must reconstruct the image. We find that a pretrained recognition encoder is unnecessary for semantic organization: an EMA copy of the autoencoder itself supplies targets that raise frozen linear probing by 5.1 points over a reconstruction-only baseline while reconstruction improves at the same time. We also find that position-dependent token corruption orders information along the sequence, so a single trained model decodes variable-length prefixes without retraining. Overall, WithinVAE improves all four reconstruction metrics over the closest representation-pivoted baseline at half its scalar capacity, and its 128-token prefix still exceeds that baseline in PSNR and SSIM at one quarter. We hope our work provides useful insights for building semantic tokenizers without external supervision.

AI DISCLOSURE

Among the research subtasks whose disclosure is required, we used generative AI tools only to clean and reformat datasets: specifically, to write and revise data-preprocessing and dataset-adaptation code for the ImageNet-1K pipeline and for the auxiliary Qwen-Image latent interfaces, including resize, crop, normalization, sharding, and caching utilities. We did not use generative AI tools to develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on our research methodology or experiments, implement the method itself, interpret results, or perform translation. The remaining required-disclosure subtasks are not applicable to this work: we generate no synthetic datasets (all data are real images from ImageNet-1K), the paper states no mathematical claims and contains no proofs, and we conduct no qualitative or thematic data analysis.

Additionally, among the subtasks whose disclosure is recommended, we used generative AI tools to improve the readability of the manuscript and to proofread it, to draw initial sketches of the method figure, to help identify relevant literature, and to propose candidate keywords for the paper.

We have reviewed all AI-assisted work. All preprocessing and dataset-adaptation code was read line by line, executed, and validated by the authors against expected tensor shapes, value ranges, and sample counts; the reported reconstruction metrics were then computed by our own evaluation scripts on the full 50,000-image validation set. The AI-generated figure sketches were used only as layout drafts and were subsequently redrawn manually by the authors, so every element in the final figures is author-produced. All AI-suggested references were retrieved and read in their original form before citation, and any suggestion we could not verify was discarded rather than cited. All AI-edited prose was checked sentence by sentence against the underlying experimental records to confirm that no claim, number, or qualification was altered, strengthened, or introduced. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work trains visual autoencoders on ImageNet-1K, a standard public benchmark used here under its existing terms for non-commercial research; it involves no human subjects, no personally identifying annotation, and no new data collection or release. We are nonetheless aware of two concerns that follow from the artifact rather than the experiments. First, ImageNet is known to carry geographic, categorical, and representational imbalance, and our frozen-probe accuracies inherit that distribution: they measure semantic organization relative to ImageNet classes and should not be read as evidence of uniform performance across populations or visual domains. Second, WithinVAE is a tokenizer for latent generative modeling, so improvements to compact reconstruction and to the latent interface are, in principle, transferable to synthetic image generation and therefore to misuse such as non-consensual or deceptive imagery. We report no capability aimed at photorealistic synthesis of real individuals, and our generation results trail the baseline we compare against, but we consider provenance and detection safeguards a legitimate obligation for downstream deployment of any tokenizer of this kind. We declare no competing interests or sponsorship that could have influenced the reported results.

REPRODUCIBILITY STATEMENT

We have aimed to make every reported number reconstructible from the paper itself. The method is specified completely in Section 3: Equations 1–6 define the posterior, the position-dependent masking schedule, prefix decoding, and both self-distillation losses, and Algorithms 1–3 in Appendix B.3 give the per-step training procedure for the main model and both auxiliary interfaces, including the stop-gradient and EMA-update ordering. Section 4 states the data, augmentation, loss weights, optimizer, schedules, architecture, initialization source, and compute budget, and Appendix B repeats these settings with the learning-rate multipliers and the independent implementation details of the auxiliary paths. Appendix B.2 and Table 1 therein fix the frozen-evaluation protocol exactly—feature extraction, linear-probe and 20-NN hyperparameters, and the paired-view procedure—so that our RPiAE comparison can be rerun under identical conditions.

Two limits on reproducibility are worth stating plainly. Reconstruction uses all 50,000 ImageNet validation images with deterministic preprocessing and lossless export, and generation uses 50,000 class-uniform samples, so our own numbers are not sensitive to sample selection; but the baseline reconstruction values in Table 1 are taken from published papers and therefore remain subject to exporter and metric-library differences, which is why we do not treat small gaps between them as a ranking. Appendix A states the inferential strength and evaluation scope of each comparison for this reason. All reported runs use a single node of eight NVIDIA H100 GPUs, with the autoencoder trained for 10 epochs and the downstream LightningDiT for 80; results at other budgets are not claimed. We will release code, configuration files, and trained checkpoints for the main model and both auxiliary interfaces.

REFERENCES

Hongjun An, Wenhan Hu, Sida Huang, Siqi Huang, Ruanjun Li, Yuanzhi Liang, Jiawei Shao, Yiliang Song, Zihan Wang, Cheng Yuan, et al. Ai flow: Perspectives, scenarios, and approaches. *Vicinagearth*, 3(1):1, 2026.

Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, 2019.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Bowei Chen, Sai Bi, Hao Tan, He Zhang, Tianyuan Zhang, Zhengqi Li, Yuanjun Xiong, Jianming Zhang, and Kai Zhang. Aligntok: Aligning visual foundation encoders to tokenizers for diffusion models. In *International Conference on Learning Representations*, 2026a.

Hao Chen, Yujin Han, Fangyi Chen, Xiang Li, Yidong Wang, Jindong Wang, Ze Wang, Zicheng Liu, Difan Zou, and Bhiksha Raj. Masked autoencoders are effective tokenizers for diffusion models. *arXiv preprint arXiv:2502.03444*, 2025.

Xiangyu Chen, Jixiang Luo, Jingyu Xu, Fangqiu Yi, Chi Zhang, and Xuelong Li. Generative video compression: towards 0.01% compression rate for video transmission. *Vicinagearth*, 3(1):7, 2026b.

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

Yuan Gao, Chen Chen, Tianrong Chen, and Jiatao Gu. One layer is enough: Adapting pretrained visual encoders for image generation. *arXiv preprint arXiv:2512.07829*, 2025.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. In *International Conference on Learning Representations*, 2024.

Yue Gong, Hongyu Li, Shanyuan Liu, Bo Cheng, Yuhang Ma, Liebucha Wu, Xiaoyu Wu, Manyuan Zhang, Dawei Leng, Yuhui Yin, and Lijun Zhang. RPiAE: A representation-pivoted autoencoder enhancing both image generation and editing. *arXiv preprint arXiv:2603.19206*, 2026.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Toshiaki Koike-Akino and Ye Wang. Stochastic bottleneck: Rateless auto-encoder for flexible dimensionality reduction. In *IEEE International Symposium on Information Theory*, 2020.

Xin Leng, Jaskirat Singh, Yutong Hou, Zirui 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*, 2025.

Xuelong Li. Positive-incentive noise. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):8708–8714, 2022.

Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410*, 2024.

Keita Miwa, Kento Sasaki, Hidehisa Arai, Tsubasa Takahashi, and Yu Yamaguchi. One-d-piece: Image tokenizer meets quality-controllable compression. *arXiv preprint arXiv:2501.10064*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. 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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

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*, 2022.

Claude E. Shannon. Coding theorems for a discrete source with a fidelity criterion. In *IRE National Convention Record*, volume 4, pp. 142–163. Institute of Radio Engineers, 1959.

Jiawei Shao and Xuelong Li. AI flow at the network edge. *IEEE Network*, 40(1):330–336, January 2026. doi: 10.1109/MNET.2025.3541208.

Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder. In *International Conference on Learning Representations*, 2026.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. *arXiv preprint arXiv:2605.18324*, 2026.

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

Junke Wang, Yi Jiang, Zehuan Yuan, Binyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. *arXiv preprint arXiv:2406.09399*, 2024.

Mark Weber, Lijun Yu, Qihang Yu, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. Maskbit: Embedding-free image generation via bit tokens. *arXiv preprint arXiv:2409.16211*, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Dianbing Xi, Zhaoqi Zhou, Youming Wang, Jiepeng Wang, and Chi Zhang. A survey: advances in multi-modal visual understanding and generation. *Vicinagearth*, 3(1):8, 2026.

Wanghan Xu, Xiaoyu Yue, Zidong Wang, Yao Teng, Wenlong Zhang, Xihui Liu, Luping Zhou, Wanli Ouyang, and Lei Bai. Exploring representation-aligned latent space for better generation. *arXiv preprint arXiv:2502.00359*, 2025.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. *arXiv preprint arXiv:2501.01423*, 2025.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2022.

Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. In *Advances in Neural Information Processing Systems*, 2024a.

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. *arXiv preprint arXiv:2410.06940*, 2024b.

Hongyuan Zhang, Sida Huang, Yubin Guo, and Xuelong Li. Variational positive-incentive noise: How noise benefits models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47 (9):8313–8320, 2025a.

Shilong Zhang, Hao Zhang, Zhaoyang Zhang, Chongjian Ge, Siting Xue, Shilong Liu, Mengyao Ren, Seonghyeon Y. Kim, Yutao Zhou, Qing Liu, et al. Both semantics and reconstruction matter: Making representation encoders ready for text-to-image generation and editing. *arXiv preprint arXiv:2512.17909*, 2025b.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022.