The Learnability Gap in Medical Latent Diffusion

Mischa Dombrowski[1], Felix Nützel[1], and Bernhard Kainz[1,2]

[1] FAU Erlangen-Nürnberg, Erlangen, DE, mischa.dombrowski@fau.de
[2] Department of Computing, Imperial College London, London, UK

Abstract. Generative data augmentation with latent diffusion models is a promising strategy for addressing class imbalance in medical imaging, yet current approaches focus on perceptual fidelity and domain-specific autoencoder fine-tuning while neglecting a more fundamental bottleneck. We identify and formalize the *learnability gap*: large-scale pretrained autoencoders faithfully encode discriminative features for medical classification, as evidenced by near-lossless performance in reconstruction space, yet their latent representations are structured in ways that are difficult for classifiers to learn from. Across five autoencoder families and four medical benchmarks spanning chest radiography, dermatoscopy, computed tomography, and echocardiography, we show that this gap persists regardless of architecture, initialization strategy, or hyperparameter tuning, and that medical-domain fine-tuning of the autoencoder does not close it. To probe and partially narrow the gap, we introduce noise-conditioned latent classifiers with FiLM layers and image-space distillation, deploying them as diagnostic instruments and initial mitigations to evaluate latent space quality. These models offer 64× throughput and 120× memory gains over image-space models while serving as diagnostic tools for latent space quality. Our analysis provides a new framework for evaluating autoencoder latent spaces and identifies their structure, rather than their fidelity or domain specificity, as the primary obstacle to closing the performance gap between real and synthetic medical training data. Model and code made available at https://github.com/MischaD/LearnabilityGap.git.

Keywords: Image generation · Latent diffusion · Representation learning · Long-tail classification.

1 Introduction

Clinical datasets follow long-tailed distributions almost universally: common diagnoses dominate while rare but critical conditions populate the tail [15,16,23]. In chest radiography, "No Finding" can exceed 60% of studies while pneumomediastinum appears in fewer than 0.1% [17]. Similar imbalances arise in dermatoscopy [3,36], computed tomography [10], and congenital heart disease screening [38]. Models trained on such distributions systematically underdiagnose rare conditions [2,18], and acquiring additional labeled examples is expensive and privacy-constrained [33]. Latent diffusion models [34,30,20] offer a compelling

path: a hospital trains a generative model locally and shares synthetic samples on demand, a privacy-preserving alternative to federated learning [33,27]. The practical value of this pipeline depends on whether generated images carry the discriminative features downstream classifiers need, particularly for tail classes. Current work addresses this through perceptual fidelity (FID [12]) and domain-specific autoencoder fine-tuning [37], implicitly treating the latent space as a neutral intermediary. Recent studies on natural images challenge this assumption: latent spaces contain high-frequency artifacts [35], geometric equivariance [21,42], and benefit from alignment with pretrained representations [40,22,39] [41,9]. The emerging view is that latent space structure actively shapes what a generative model can learn [1,8], yet the implications for class-conditional generation in long-tailed medical settings remain unexplored. We make this question precise through a systematic study across five autoencoder families and four clinical benchmarks. Reconstruction-space classifiers match image-space performance (p=0.375, Wilcoxon signed-rank test), confirming that the autoencoders faithfully preserve discriminative information. Yet latent-space classifiers suffer a substantial drop despite extensive architecture search, distillation, and noise-conditional training ($p = 9.5 \times 10^{-7}$, Wilcoxon across all 20 AE×dataset pairs). We term this the *learnability gap* (Fig. 1): the information is present but structured in a way that resists direct learning. This gap propagates to generative augmentation: if the diffusion model cannot learn class-discriminative structure from the latents, its samples will lack the pathological markers classifiers rely on.

Our contributions are: **(1)** We identify and quantify the learnability gap across five autoencoder families and four medical benchmarks, showing it is systematic ($p = 9.5 \times 10^{-7}$) and independent of reconstruction fidelity, domain specificity, and initialization strategy. **(2)** We present noise-conditioned latent classifiers with FiLM layers and image-space distillation as diagnostic instruments and initial mitigations. Rather than treating these layers as our primary architectural contribution, we use them to probe the lower bound of generative utility while offering 64× throughput gains, making them practical for rejection sampling and quality filtering in generative pipelines. **(3)** We provide a controlled analysis isolating the sources of the gap through systematic ablations of fine-tuning, distillation, and noise conditioning, establishing latent space *structure* as the bottleneck for long-tail medical synthesis.

Related work. Long-tailed medical classification has been extensively benchmarked [15,16,23], with LDAM [2], decoupled training [18], and focal loss [24] as standard remedies. Generative augmentation via class-balancing diffusion [32], guided adapters [28], synthetic pretraining [27], and diversity control [7] has shown promise. Latent diffusion [34,13,30,20] relies on autoencoder latent spaces whose structure significantly impacts generation quality through high-frequency artifacts [35], equivariance deficits [21,42], geometry [5], and spectral biases [8]. Representation alignment methods [40,22,39,41,9] improve latent spaces using pretrained encoders such as DINOv2 [29], and latent-space learnability has been studied for natural images [1] and memorization detection [6]. Domain-specific

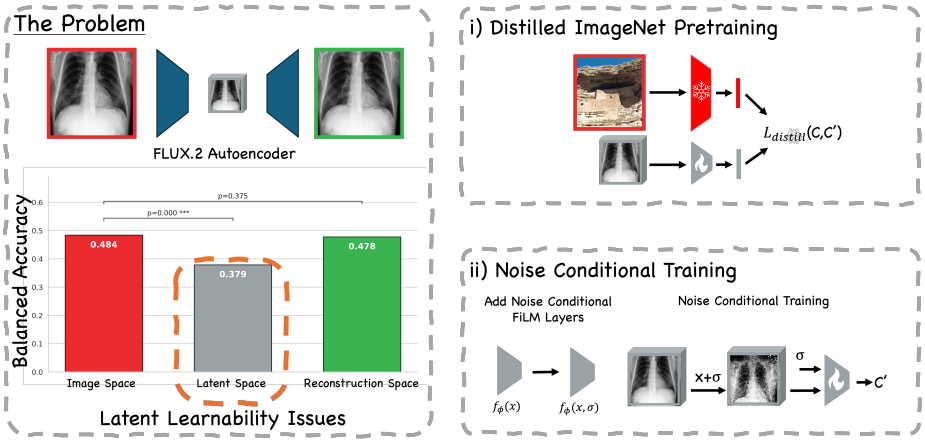

Fig. 1: Method overview. A frozen pretrained autoencoder maps images to latent space and back. The learnability gap: reconstruction-space classifiers match image space, but latent-space classifiers degrade substantially. A ConvNeXt-Tiny student classifier with FiLM-based noise conditioning is distilled from an image-space teacher.

autoencoders preserve clinical features [37]. Classifier guidance at inference [4,14] and autoguidance [19] improve sample quality but do not address the underlying representation. We differ by providing the first systematic diagnosis of the learnability gap in medical latent spaces and by developing noise-conditioned latent classifiers as both diagnostic and practical tools for generative pipelines. We show this gap persists in medical imaging and use a probing framework to isolate the diagnosis from the confounding sampling process, unlike prior natural-image studies [21,22].

2 Method

An overview is given in Fig. 1. We operate in the latent space of a frozen pretrained autoencoder. Let x_{img} denote an image, $x = E(x_{\text{img}})$ its latent encoding, and $\tilde{x}_{\text{img}} = G(x)$ the decoded reconstruction. All latents are channel-wise normalized over the training set.

Three evaluation spaces. To isolate the effect of latent encoding from information loss, we define three evaluation spaces: *image space* (classifiers trained on original images x_{img}), *latent space* (classifiers trained on latent codes x), and *reconstruction space* (classifiers trained on decoded reconstructions $\tilde{x}_{\text{img}} = G(E(x_{\text{img}}))$). If latent-space classifiers underperform reconstruction-space classifiers, the gap cannot be caused by missing information (since the decoder can recover it) but must instead reflect how the latent space *structures* that information. This three-way comparison is the core diagnostic tool of our analysis.

Noise-conditioned latent classifier. Our core component is a classifier $f_\phi(x_\sigma, \sigma)$ that predicts the class label c from noisy latents at any noise level σ, forcing recovery of semantic information even at low signal-to-noise ratios. We adopt a ConvNeXt-Tiny [25] backbone operating directly on the latent tensor, chosen for its strong performance in prior latent-space studies [6]. The noise level σ is embedded via Fourier features followed by an MLP with SiLU activations, and injected at the end of each ConvNeXt stage through Feature-wise Linear Modulation (FiLM) [31]: $\gamma, \beta = We(\sigma)$; $x' = x \odot (1 + \gamma) + \beta$. FiLM parameters are zero-initialized so the backbone initially acts as an identity mapping and gradually learns noise-dependent modulation. The classifier is trained with $\mathcal{L}_{\text{cls}} = \mathbb{E}_{x,\varepsilon,\sigma}[\text{CE}(f_\phi(x+\sigma\varepsilon,\, \sigma),\, c)]$.

Our motivation for noise conditioning is twofold. First, it improves classification by forcing the network to learn robust, noise-invariant features rather than overfitting to clean-latent statistics (Table 4). Second, noise-conditioned classifiers integrate naturally into diffusion model training as regularizers on the predicted clean latent $\hat{x}_0$ or as guidance models during sampling, providing a practical path toward narrowing the learnability gap at the generative level.

Distillation from image space. Direct training of the latent classifier is limited by the learnability gap itself. We therefore distill from a strong image-space teacher T operating on reconstructions $\tilde{x}_{\text{img}}$. With teacher logits $z_T = T(\tilde{x}_{\text{img}})$ and student logits $z_S = f_\phi(x)$, we optimize:

$$\mathcal{L}_{\text{distill}} = \alpha\, \text{CE}(z_S, c) + (1 - \alpha)\, \|z_S - z_T\|_2^2, \qquad \alpha = 0.5, \tag{1}$$

combining ground-truth supervision with soft logit matching. This encourages the latent student to mimic the image-space decision boundary while retaining hard-label supervision. The combination of distillation and noise conditioning yields our strongest latent classifier, which we use both to quantify the learnability gap and as a practical component for downstream generative pipelines (*e.g.* rejection sampling, training-time regularization).

3 Experiments

Setup. We evaluate on four medical benchmarks: MIMIC-CXR long-tail (19 chest X-ray findings, N=111k) [17], ISIC 2019 (8 skin lesion classes, N=25k) [3,36], CT-RATE (13 CT findings, N=23k) [10], and Cardium (CHD vs. normal, N=6.6k) [38]. All datasets exhibit long-tailed distributions with imbalance ratios exceeding 50:1. We compare five autoencoder families: Stable Diffusion 1.4 (SD 1.4), Flux.1 dev, Flux.2 dev, MedVAE [37], and MedVAE fine-tuned on each target dataset (MedVAE-FT). Image-space and reconstruction-space classifiers use ImageNet-pretrained ResNet-50 [11] with LDAM loss [2] and deferred reweighting at epoch 10, trained for up to 60 epochs (patience 15). Latent classifiers use ConvNeXt-Tiny [25] with the input layer adapted to each autoencoder's channel count. All results report 5-fold cross-validation on identical folds and seeds; we report mean$_{\pm\text{std}}$. We evaluate balanced accuracy (bACC), area under the ROC curve (AUC), and Matthews correlation coefficient (MCC).

Table 1: The learnability gap across autoencoder families (pretrained init., 5-fold CV). Latent space (LS) classifiers underperform reconstruction space (RS), which matches image space. † Paired t-test $p<0.05$; ‡ $p<0.001$. Best LS per dataset in **bold**. Overall LS vs RS: Wilcoxon $p = 9.5 \times 10^{-7}$.

		CT-RATE			ISIC-2019			MIMIC-LT			Cardium		
Method	Sp.	bACC	AUC	MCC	bACC	AUC	MCC	bACC	AUC	MCC	bACC	AUC	MCC
Image Space	–	$.255_{\pm.018}$	$.720_{\pm.017}$	$.280_{\pm.033}$	$.749_{\pm.021}$	$.947_{\pm.008}$	$.703_{\pm.027}$	$.263_{\pm.010}$	$.760_{\pm.015}$	$.279_{\pm.017}$	$.669_{\pm.028}$	$.743_{\pm.036}$	$.280_{\pm.061}$
SD 1.4	LS	$.156^{\ddagger}_{\pm.010}$	$.635_{\pm.020}$	$.147_{\pm.014}$	$.468^{\ddagger}_{\pm.012}$	$.835_{\pm.011}$	$.428_{\pm.020}$	$.135^{\ddagger}_{\pm.007}$	$.656_{\pm.013}$	$.191_{\pm.022}$	$.624_{\pm.014}$	$.676_{\pm.010}$	$.202_{\pm.012}$
	RS	$.246_{\pm.020}$	$.727_{\pm.018}$	$.257_{\pm.015}$	$.746_{\pm.009}$	$.943_{\pm.005}$	$.700_{\pm.009}$	$.261_{\pm.012}$	$.769_{\pm.008}$	$.285_{\pm.018}$	$.662_{\pm.032}$	$.723_{\pm.043}$	$.270_{\pm.077}$
Flux.1	LS	$.167^{\ddagger}_{\pm.009}$	$.639_{\pm.027}$	$.153_{\pm.015}$	$.496^{\ddagger}_{\pm.013}$	$.853_{\pm.006}$	$.466_{\pm.005}$	$.132^{\ddagger}_{\pm.009}$	$.662_{\pm.009}$	$.197_{\pm.011}$	$.627_{\pm.024}$	$.682_{\pm.026}$	$.209_{\pm.054}$
	RS	$.252_{\pm.020}$	$.725_{\pm.014}$	$.279_{\pm.017}$	$.744_{\pm.017}$	$.943_{\pm.004}$	$.697_{\pm.015}$	$.263_{\pm.025}$	$.773_{\pm.008}$	$.274_{\pm.023}$	$.667_{\pm.036}$	$.724_{\pm.055}$	$.287_{\pm.068}$
Flux.2	LS	$\mathbf{.169}^{\dagger}_{\pm.022}$	$.653_{\pm.019}$	$\mathbf{.174}_{\pm.008}$	$\mathbf{.552}^{\ddagger}_{\pm.013}$	$.884_{\pm.003}$	$\mathbf{.527}_{\pm.019}$	$\mathbf{.166}^{\ddagger}_{\pm.010}$	$.703_{\pm.005}$	$\mathbf{.221}_{\pm.034}$	$\mathbf{.639}_{\pm.026}$	$.690_{\pm.053}$	$\mathbf{.223}_{\pm.043}$
	RS	$.242_{\pm.017}$	$.726_{\pm.011}$	$.260_{\pm.008}$	$.755_{\pm.019}$	$.951_{\pm.005}$	$.714_{\pm.012}$	$.247_{\pm.023}$	$.767_{\pm.008}$	$.282_{\pm.014}$	$.666_{\pm.021}$	$.722_{\pm.036}$	$.268_{\pm.038}$
MedVAE	LS	$.164^{\dagger}_{\pm.014}$	$.641_{\pm.010}$	$.176_{\pm.007}$	$.539^{\ddagger}_{\pm.021}$	$.877_{\pm.008}$	$.509_{\pm.018}$	$.144^{\ddagger}_{\pm.006}$	$.673_{\pm.008}$	$.210_{\pm.013}$	$.633_{\pm.053}$	$.684_{\pm.068}$	$.227_{\pm.074}$
	RS	$.239_{\pm.015}$	$.719_{\pm.011}$	$.261_{\pm.017}$	$.725_{\pm.018}$	$.934_{\pm.007}$	$.667_{\pm.013}$	$.205_{\pm.011}$	$.730_{\pm.011}$	$.248_{\pm.029}$	$.644_{\pm.028}$	$.720_{\pm.051}$	$.235_{\pm.059}$
MedVAE-FT	LS	$.164^{\dagger}_{\pm.021}$	$.655_{\pm.010}$	$.168_{\pm.018}$	$.553^{\ddagger}_{\pm.015}$	$.882_{\pm.006}$	$.525_{\pm.013}$	$.157^{\dagger}_{\pm.006}$	$.682_{\pm.004}$	$.205_{\pm.018}$	$.623_{\pm.029}$	$.680_{\pm.048}$	$.203_{\pm.047}$
	RS	$.238_{\pm.019}$	$.717_{\pm.019}$	$.267_{\pm.011}$	$.750_{\pm.021}$	$.949_{\pm.008}$	$.709_{\pm.018}$	$.257_{\pm.027}$	$.767_{\pm.013}$	$.276_{\pm.017}$	$.661_{\pm.046}$	$.725_{\pm.048}$	$.265_{\pm.074}$

Table 2: Recon. quality. High fidelity does not predict latent-space learnability: Flux.2 leads in PSNR but shows the largest LS gap on ISIC.

	CT-RATE		ISIC		MIMIC		Cardium	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
SD 1.4	.924	34.2	.971	37.3	.983	35.2	.985	30.5
Flux.1	.946	40.8	.994	40.3	.997	40.5	.999	39.1
Flux.2	.941	40.6	**.994**	**42.3**	**.997**	**41.9**	**.999**	**40.8**
MedVAE	.938	38.2	.963	24.9	.711	23.7	.991	32.7
MedVAE-FT	.939	39.5	.991	41.1	.996	39.8	.996	35.2

Table 3: ImageNet, 24 h budget, single GH200. Latent training: 64× faster, 120× less memory.

Space	bACC	AUC	MCC	S/s	MB
Image	.459	–	–	50	20k
Latent	.630	.996	.629	3.2k	166
+ Distill.	.711	.993	.710	3.2k	166

The learnability gap is large and consistent. Table 1 presents the central finding. Across all five autoencoder families and four datasets, latent-space classifiers underperform their reconstruction-space counterparts. On ISIC-2019, the bACC gap exceeds 18 percentage points for every autoencoder (up to 27.8 for SD 1.4; $p<10^{-4}$, paired t-test). On MIMIC-LT the gap ranges from 6 to 13 points, consistently significant ($p<0.002$). A Wilcoxon signed-rank test across all 20 AE×dataset pairs confirms the gap is systematic ($p = 9.5 \times 10^{-7}$). Reconstruction-space performance is not significantly different from image space (p=0.375, Wilcoxon), confirming that discriminative information is faithfully preserved. Cardium is the only dataset where the gap does not reach significance, attributable to its binary nature and high cross-fold variance.

Reconstruction fidelity does not predict learnability. Table 2 reports reconstruction quality for each autoencoder. Flux.2 achieves the highest PSNR (41.4 dB macro average) yet also exhibits a large LS gap, while MedVAE sometimes has lower fidelity (*e.g.* 23.7 dB on MIMIC) but a comparable classification gap. Fig. 3 confirms this visually: pixel-wise differences between originals and reconstructions are minimal across autoencoders, yet the classification gap persists. Fig. 2 makes this explicit: across all 20 AE×dataset combinations, PSNR

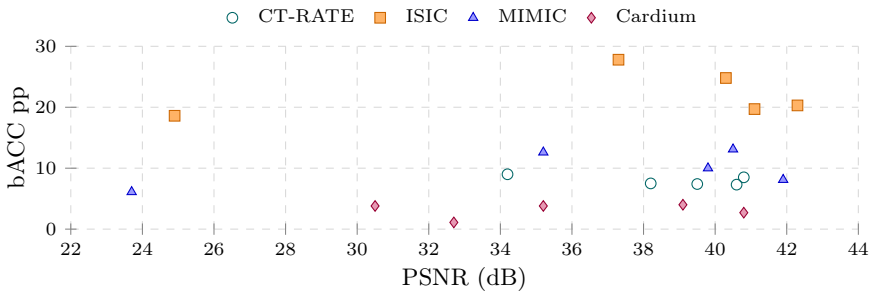

Fig. 2: PSNR vs. learnability gap. Each point is one AE×dataset pair. Higher fidelity does not reduce the gap; dataset difficulty dominates.

Table 4: Latent classifier ablation on Flux.2 (5-fold CV). Each component helps incrementally, but the gap to RS persists, confirming its structural nature. ⋆ Significant vs. naïve (p<0.05, paired t-test).

	CT-RATE			ISIC-2019			MIMIC-LT			Cardium			Macro
Method	bACC	AUC	MCC	bACC	AUC	MCC	bACC	AUC	MCC	bACC	AUC	MCC	AUC
RS baseline	$.242_{\pm.017}$	$.726_{\pm.011}$	$.260_{\pm.008}$	$.755_{\pm.019}$	$.951_{\pm.005}$	$.714_{\pm.012}$	$.247_{\pm.023}$	$.767_{\pm.008}$	$.282_{\pm.014}$	$.666_{\pm.021}$	$.722_{\pm.036}$	$.268_{\pm.038}$	.791
Naïve LS	$.169_{\pm.022}$	$.653_{\pm.019}$	$.174_{\pm.008}$	$.552_{\pm.013}$	$.884_{\pm.003}$	$.527_{\pm.019}$	$.166_{\pm.010}$	$.703_{\pm.005}$	$.221_{\pm.034}$	$.639_{\pm.026}$	$.690_{\pm.053}$	$.223_{\pm.043}$	.731
+ Distill.	$.174_{\pm.007}$	$.664_{\pm.018}$	$.202_{\pm.015}$	$.581_{\pm.064}$	$.898_{\pm.021}$	$.565_{\pm.045}$	$.152_{\pm.005}$	$.697_{\pm.017}$	$.228_{\pm.029}$	$.653_{\pm.024}$	$.728_{\pm.028}$	$.260_{\pm.040}$	.747
+ HP opt.⋆	$.205_{\pm.015}$	$.695_{\pm.018}$	$.209_{\pm.018}$	$.643_{\pm.021}$	$.901_{\pm.012}$	$.584_{\pm.039}$	$.202_{\pm.010}$	$.740_{\pm.007}$	$.265_{\pm.022}$	$.567_{\pm.093}$	$.573_{\pm.159}$	$.109_{\pm.151}$	.727
+ Noise cond.	$.174_{\pm.028}$	$.669_{\pm.027}$	$.203_{\pm.028}$	$.591_{\pm.043}$	$.896_{\pm.008}$	$.563_{\pm.020}$	$.177_{\pm.011}$	$.720_{\pm.014}$	$.258_{\pm.012}$	$.663_{\pm.020}$	$.733_{\pm.009}$	$.289_{\pm.021}$	.754

shows no predictive relationship with gap magnitude. This rules out information loss as the cause and implicates the *structure* of the latent space.

Medical fine-tuning does not close the gap. Comparing MedVAE with MedVAE-FT in Table 1 reveals that fine-tuning the autoencoder on in-domain data improves reconstruction quality substantially (PSNR from 29.9 to 38.9 dB on average, Table 2) but does not meaningfully improve latent-space classification: the macro-average LS bACC changes from .370 to .374, with significance on only one of four datasets (MIMIC-LT, p=0.04; paired t-test). Meanwhile, RS benefits more from fine-tuning (MIMIC p=0.008, ISIC p=0.02), widening the gap rather than closing it. This confirms that the learnability gap is structural, not a domain mismatch, and that investing compute in autoencoder fine-tuning is not an effective remedy.

Latent classifier ablation. Table 4 ablates strategies for training latent-space classifiers on Flux.2. Naïve training already exposes the full learnability gap. Systematic hyperparameter optimization yields a distilled variant reaching a macro bACC of .404 (p<0.03 on 3 of 4 datasets), though it remains unstable on the data-limited Cardium task ($bACC = .567_{\pm.093}$). Introducing noise conditioning via FiLM layers slightly reduces macro bACC to .401 but substantially improves discriminative robustness: mean AUC rises from .727 to .754 and MCC from .292 to .328. This improvement is most notable on Cardium (MCC .223 → .289), confirming that noise-aware training learns more robust decision boundaries. Despite these interventions, a gap of 4–11 bACC points relative to reconstruction space

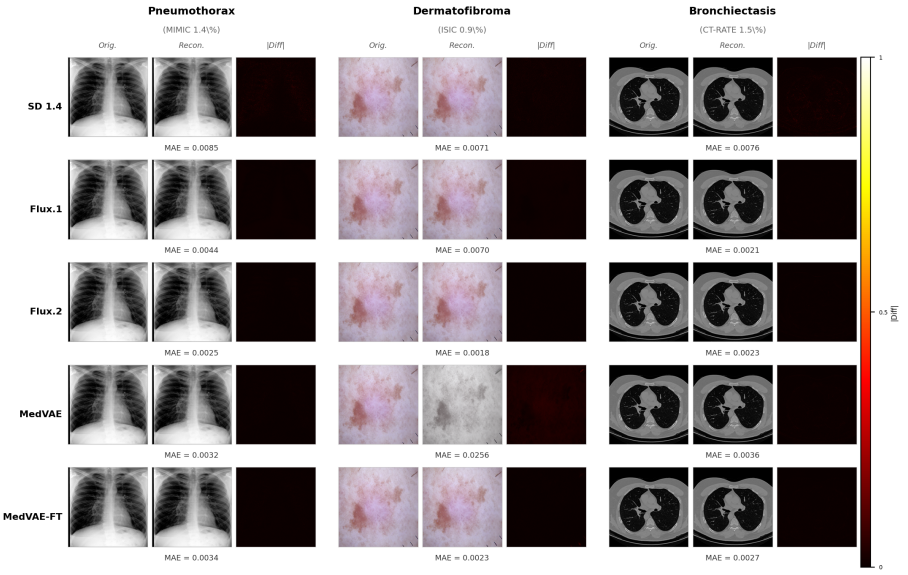

Fig. 3: Reconstruction samples and pixel-wise absolute difference for long-tail classes across CT-RATE, MIMIC-LT, and ISIC. Reconstructions are near-identical to originals, confirming that the autoencoder preserves discriminative information despite the large classification gap in latent space.

(RS) persists. This suggests the learnability gap is a structural property of the latent space rather than an artifact of suboptimal training.

Latent classifiers are fast. Table 3 demonstrates that latent classifiers, despite the learnability gap, offer compelling practical advantages. On ImageNet with a fixed 24-hour budget on a single NVIDIA GH200, latent ConvNeXt-Tiny reaches bACC .630 vs. .459 in image space, a 17-point gain from operating in the compressed domain. With distillation it reaches .711. Throughput increases 64× (3.2k vs. 50 samples/sec) and memory drops 120× (166 MB vs. 20 GB) compared to image space classifiers which have to decode the latents first. This makes latent classifiers practical as fast quality filters for generated images and as components in training-time regularization schemes for diffusion models.

Why does the gap exist? Our experiments systematically rule out several potential explanations. Information loss is excluded by the RS≈IS equivalence (p=0.375). Domain mismatch is excluded by the failure of medical fine-tuning to close the gap despite improving reconstruction from 29.9 to 38.9 dB PSNR. Insufficient classifier capacity is unlikely: ConvNeXt-Tiny achieves strong RS and image-space performance, and the gap persists or widens with stronger classifiers and extensive hyperparameter tuning. We hypothesize that autoencoders trained for pixel-wise reconstruction distribute class-discriminative features across high-frequency spatial patterns and inter-channel correlations that convolutional classifiers struggle to exploit, while the decoder inverts these pat-

terns precisely because it was jointly trained with the encoder. We see this CNN-unfriendly latent format as a highly plausible structural explanation for this gap. This interpretation is consistent with recent findings on spectral biases [8] and high-frequency artifacts [35] in autoencoder latent spaces. Our probing framework isolates the gap from standard generative training confounders, while our autoencoder finetuning results counter the assumption that in-domain autoencoders resolve latent degradation. The gap is largest on ISIC-2019 (18-28 points), where fine-grained texture differences between lesion types may be particularly susceptible to such encoding artifacts. These findings suggest that latent space restructuring, *e.g.* through class-conditional objectives, equivariance-aware architectures [21], or alignment with discriminative encoders [40], which is a more promising direction than improving reconstruction fidelity.

Discussion. Our results establish that the bottleneck in latent diffusion for medical long-tail synthesis is not the autoencoder's representational capacity but the *learnability* of its latent space. The gap is robust across five autoencoder families, four clinical modalities, and extensive hyperparameter tuning ($p = 9.5 \times 10^{-7}$, Wilcoxon). This finding has direct implications for the field: efforts to improve synthetic medical data should prioritize latent space restructuring over reconstruction fidelity or domain-specific fine-tuning. Our noise-conditioned latent classifiers, while not fully closing the gap, provide both a diagnostic framework for evaluating latent space quality and practical components for generative pipelines through their 64× throughput advantage. A limitation is that we diagnose the gap but do not resolve it at the representation level. Removing this gap and validating with expert radiologist evaluation are natural next steps.

Broader impact. Our findings have practical implications for the deployment of generative models in clinical settings. The learnability gap means that current latent diffusion pipelines may systematically underrepresent rare pathologies in synthetic training data, potentially perpetuating the very diagnostic biases they are designed to mitigate. By identifying latent space structure as the bottleneck, we redirect optimization efforts away from expensive autoencoder fine-tuning toward more targeted interventions. Our noise-conditioned latent classifiers can serve as lightweight quality filters in privacy-preserving synthetic data sharing pipelines, enabling hospitals to verify the clinical plausibility of generated images without exposing real patient data. Addressing the learnability gap could accelerate the adoption of generative augmentation for rare disease diagnosis, where data scarcity is most acute and the clinical need is greatest.

4 Conclusion

We introduced the *learnability gap*: a systematic discrepancy between the discriminative information preserved in autoencoder latent spaces and the ability of downstream models to access it. Through a comprehensive exploration spanning five autoencoder families, four long-tailed medical benchmarks, and multiple classifier training paradigms, we showed that this gap, *i.e.* not reconstruction fidelity or domain specificity, is the primary bottleneck limiting the utility of

latent representations for medical data augmentation. The gap is highly significant ($p = 9.5 \times 10^{-7}$) and is not remedied by medical-domain autoencoder fine-tuning. We developed noise-conditioned latent classifiers with FiLM layers and image-space distillation that partially narrow the gap and offer dramatic efficiency gains (64× throughput, 120× memory reduction), making them practical for quality filtering and rejection sampling in generative pipelines. We provide three actionable insights: (i) large-scale pretrained autoencoders are sufficient for medical synthesis without domain fine-tuning, as the bottleneck lies in latent space structure rather than domain specificity; (ii) latent-space classifiers are fast and useful but require careful training via distillation and noise conditioning; and (iii) the persistent learnability gap identifies latent space restructuring as the critical open problem for advancing synthetic medical data augmentation.

Acknowledgments. We acknowledge HPC resources from NHR@FAU (projects b143dc, b180dc), funded by federal and Bavarian state authorities and Gerhard Wellein's and his team's HPC approach. NHR@FAU hardware is partially funded by DFG 440719683. Additional support was received from ERC projects MIA-NORMAL 101083647, CHARMS 101246053, and ORACLE 101326871, DFG 513220538 and 512819079, and the state of Bavaria (HTA and the Bavarian Foundation Model Initiative). We further acknowledge resources provided by the Isambard-AI National AI Research Resource (AIRR), operated by the University of Bristol and funded by DSIT via UKRI and STFC [ST/AIRR/I-A-I/1023] [26]. We used coding agents and LLMs from Anthropic, OpenAI, Google, and Mistral AI, as writing aid, coding, experiment orchestration, and cluster monitoring.

Disclosure of Interests. The authors have no relevant competing interests.

References

1. Black Forest Labs: FLUX.2: Analyzing and enhancing the latent space of FLUX – representation comparison (2025), https://bfl.ai/research/representation-comparison
2. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. In: NeurIPS. vol. 32 (2019)
3. Codella, N.C., Gutman, D., Celebi, M.E., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). arXiv:1710.05006 (2018)
4. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. In: NeurIPS. pp. 8780–8794 (2021)
5. Dillon, B.M., Plehn, T., Sauer, C., Sorrenson, P.: Better latent spaces for better autoencoders. SciPost Physics **11**(3), 061 (2021)
6. Dombrowski, M., Nützel, F., Kainz, B.: LCMem: A universal model for robust image memorization detection. arXiv:2512.14421 (2025)
7. Dombrowski, M., Zhang, W., Cechnicka, S., Reynaud, H., Kainz, B.: Image generation diversity issues and how to tame them. In: CVPR. pp. 3029–3039 (2025)
8. Falck, F., Pandeva, T., Zahirnia, K., et al.: A Fourier space perspective on diffusion models. arXiv:2505.11278 (2025)

9. Gui, M., Schusterbauer, J., Phan, T., et al.: Adapting self-supervised representations as a latent space for efficient generation. arXiv:2510.14630 (2025)
10. Hamamci, I.E., Er, S., Wang, C., et al.: Generalist foundation models from a multimodal dataset for 3d computed tomography. Nature Biomedical Engineering (2026). https://doi.org/10.1038/s41551-025-01599-y
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR'16. pp. 770–778 (2016)
12. Heusel, M., Ramsauer, H., Unterthiner, T., et al.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: NeurIPS (2017)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
14. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
15. Holste, G., Wang, S., Jiang, Z., et al.: Long-tailed classification of thorax diseases on chest X-ray: A new benchmark study. In: DALI'22, LNCS, vol. 13567, pp. 22–32. Springer (2022). https://doi.org/10.1007/978-3-031-17027-0_3
16. Holste, G., Zhou, Y., Wang, S., et al.: Towards long-tailed, multi-label disease classification from chest X-ray: Overview of the CXR-LT challenge. Medical Image Analysis **97**, 103224 (2024). https://doi.org/10.1016/j.media.2024.103224
17. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1), 317 (2019). https://doi.org/10.1038/s41597-019-0322-0
18. Kang, B., Xie, S., Rohrbach, M., et al.: Decoupling representation and classifier for long-tailed recognition. In: ICLR'20, arXiv:1910.09217 (2020)
19. Karras, T., Aittala, M., Kynkäänniemi, T., Lehtinen, J., Aila, T., Laine, S.: Guiding a diffusion model with a bad version of itself. NeurIPS'24 **37**, 52996–53021 (2024)
20. Karras, T., Aittala, M., Lehtinen, J., et al.: Analyzing and improving the training dynamics of diffusion models. In: CVPR. pp. 24174–24184 (2024). https://doi.org/10.1109/CVPR52733.2024.02282
21. Kouzelis, T., Kakogeorgiou, I., Gidaris, S., et al.: EQ-VAE: Equivariance regularized latent space for improved generative image modeling. arXiv:2502.09509 (2025)
22. Leng, X., Singh, J., Hou, Y., et al.: REPA-E: Unlocking VAE for end-to-end tuning with latent diffusion transformers. arXiv:2504.10483 (2025)
23. Lin, M., Holste, G., Wang, S., et al.: CXR-LT 2024: A MICCAI challenge on long-tailed, multi-label, and zero-shot disease classification from chest X-ray. arXiv:2506.07984 (2025)
24. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV'17. pp. 2980–2988 (2017)
25. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: CVPR'22. pp. 11976–11986 (2022)
26. McIntosh-Smith, S., Alam, S.R., Woods, C.: Isambard-ai: a leadership class supercomputer optimised specifically for artificial intelligence. arXiv.2410.11199 (2024)
27. Moroianu, S.L., Bluethgen, C., Chambon, P.o.: Improving performance, robustness, and fairness of radiographic AI models with finely-controllable synthetic data. arXiv:2508.16783 (2025)
28. Nützel, F., Dombrowski, M., Kainz, B.: GRASP: Guided residual adapters with sample-wise partitioning. arXiv:2512.01675 (2025)
29. Oquab, M., Darcet, T., Moutakanni, T., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

30. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV, arXiv:2212.09748. pp. 4195–4205 (2023)
31. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: FiLM: Visual reasoning with a general conditioning layer. arXiv:1709.07871 (2017)
32. Qin, Y., Zheng, H., Yao, J., et al.: Class-balancing diffusion models. arXiv:2305.00562 (2023)
33. Rieke, N., Hancox, J., Li, W., et al.: The future of digital health with federated learning. npj Digital Medicine **3**(1), 119 (2020). https://doi.org/10.1038/s41746-020-00323-1
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
35. Skorokhodov, I., Girish, S., Hu, B., et al.: Improving the diffusability of autoencoders. arXiv preprint arXiv:2502.14831 (2025)
36. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**, 180161 (2018). https://doi.org/10.1038/sdata.2018.161
37. Varma, M., Kumar, A., Van der Sluijs, R., et al.: MedVAE: Efficient automated interpretation of medical images with large-scale generalizable autoencoders (2025)
38. Vega, D., Ceballos, H.V., Vera, J.S., et al.: CARDIUM: Congenital anomaly recognition with diagnostic images and unified medical records. arXiv:2510.15208 (2025)
39. Yao, J., Yang, B., Wang, X.: Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models (2025), https://arxiv.org/abs/2501.01423
40. Yu, S., Kwak, S., Jang, H., et al.: Representation alignment for generation: Training diffusion transformers is easier than you think. arXiv:2410.06940 (2025)
41. Zheng, B., Ma, N., Tong, S., Xie, S.: Diffusion transformers with representation autoencoders. arXiv preprint arXiv:2510.11690 (2025)
42. Zhou, Y., Xiao, Z., Yang, S., Pan, X.: Alias-free latent diffusion models: Improving fractional shift equivariance of diffusion latent space. arXiv preprint arXiv:2503.09419 (2025)