Channel-Wise Latent Space Decorrelation for 3D Brain MRI Generation

Wonyoung Jang*, Junho Lee*, Youngyoon Choi,
Seunggeun Lee, and Joonseok Lee(✉)

Graduate School of Data Science, Seoul National University, Seoul, Korea
{jwy4888,joon2003,youngyoon911,lee7801,joonseok}@snu.ac.kr

Abstract. Latent generative models have emerged as a promising framework for high-fidelity 3D brain MRI generation. However, we identify that the high anatomical similarity across subjects causes the encoder to under-utilize the allocated latent capacity. It tends to produce redundant latent representations that reduce the effectiveness of downstream generative training. To address this, we propose DeCo-VAE (Decorrelation VAE), a VAE that improves latent space efficiency by a channel-wise decorrelation approach. Specifically, we introduce two novel regularization losses—a scale-invariant decorrelation loss and a variance-anchored decorrelation loss—to encourage the encoder to distribute information more uniformly across channels. Our analysis suggests that the proposed method reduces channel redundancy and improves latent space utilization. Experimental results show that our approach consistently improves both reconstruction fidelity and final 3D brain MRI generation quality, outperforming the state-of-the-art MAISI baseline. Code is available at https://github.com/Stomper10/decovae.

Keywords: Brain Imaging · Generative Models · Flow Matching · Representation Learning · Latent Space.

1 Introduction

The shortage of large-scale datasets remains a bottleneck in 3D brain magnetic resonance imaging (MRI) research, where data acquisition is costly and subject to stringent privacy regulations. Generative models have emerged as a compelling solution by synthesizing realistic 3D brain MRI volumes [1, 2, 10, 14, 17, 18, 22, 26, 34, 41, 43–45, 48–50]. Among these, latent generative models [6, 7, 12, 27, 36] have demonstrated exceptional performance, combining high-fidelity generation with practical computational efficiency. Notably, MAISI [18, 48] extended the latent generative model framework to 3D medical volumes. It compressed high-dimensional CT and MRI scans into a compact latent space using a variational autoencoder (VAE) [25] before training a diffusion or flow matching model [9, 21, 23, 28, 30, 33, 39]. This latent compression reduces computational costs while enabling high-resolution volumetric generation.

* These authors contributed equally to this work.

While this latent compression offers clear computational advantages, it introduces a fundamental trade-off: compressing the data into a significantly smaller latent space risks losing critical fine-grained details. This risk is especially critical in 3D medical imaging, where voxel-level preservation of subtle structures such as cortical boundaries is essential. Presumably for this reason, representative existing work like MAISI [18, 48] chose to conservatively compress the latent dimensionality (*e.g.*, reducing the spatial dimensions by 4 rather than 8) than as in natural image domain [6, 7, 12, 27, 36], prioritizing structural fidelity over compression efficiency. However, 3D brain MRIs are inherently similar across subjects, and this conservative strategy inadvertently allocates an excessively large latent space. We therefore hypothesize that a conservative compression rate, overlooking this inherent similarity, yields a latent space far larger than the intrinsic data complexity requires. Consequently, the encoder does not distribute information evenly across channels—it produces redundant latent representations instead of utilizing the full capacity. This under-utilization makes subject representations less distinguishable and makes it harder for the latent generative model to learn the target distribution.

To verify our hypothesis, we analyze the state-of-the-art MAISI-VAE [18, 48] latent space, revealing room for improvement. We evaluate it using a variety of quantitative metrics, most notably the effective rank [38], which quantifies the number of statistically independent dimensions actively used by the encoder. Our findings show that the encoder does not fully utilize its allocated capacity. Instead, the latent channels exhibit notable correlation. This indicates that the available latent space operates less independently than optimally possible, suggesting room for improvement through better latent utilization.

To address this inefficiency, we propose **DeCo-VAE** (Decorrelation VAE), a VAE that improves latent space utilization by encouraging a more uniform distribution of information across channels. We achieve this by introducing two novel regularization losses. First, we propose a **scale-invariant decorrelation loss ($\mathcal{L}_{\mathbf{SID}}$)**. It directly penalizes the normalized cross-correlation between channels (the Pearson correlation matrix). It provides a scale-invariant metric that purely measures linear dependence, regardless of channel magnitudes. Second, we propose a **variance-anchored decorrelation loss ($\mathcal{L}_{\mathbf{VAD}}$)**. This formulation combines a covariance loss to suppress inter-channel correlations (by penalizing off-diagonal entries) with a channel-wise variance loss. The variance loss explicitly maintains the variance of each channel to prevent scale collapse.

Experimental results demonstrate that both proposed regularization losses reduce redundancy, improve effective latent utilization, and enhance downstream 3D brain MRI generation quality. Compared to the MAISI [18, 48] baseline, our methods show consistent improvements. Specifically, $\mathcal{L}_{\text{SID}}$ yields stronger reconstruction fidelity, while $\mathcal{L}_{\text{VAD}}$ achieves superior downstream generative model performance. These findings support our hypothesis regarding latent under-utilization. They also suggest that our decorrelation framework is an effective remedy for improving 3D medical latent generative model performance.

2 Related Work

Generative Models for Medical Imaging. Generative modeling for brain MRI has evolved from adversarial frameworks to diffusion and flow matching-based approaches. Early works primarily relied on GANs [1, 15, 22, 26, 41, 44, 45, 50] to synthesize 3D volumes, focusing on structural consistency and realistic texture modeling. However, these approaches often suffer from training instability and limited diversity.

Diffusion-based methods [9, 21, 23, 30, 33, 39] have since become the dominant paradigm, demonstrating strong performance in modeling complex anatomical structures [2, 10, 14, 17, 18, 24, 34, 43, 48, 49]. This trend is further advanced by latent generative models [6, 7, 12, 27, 36], including flow matching-based approaches [2, 34, 43, 49], which reduce computational cost while preserving high-fidelity details. Recent frameworks such as MAISI [18, 48] enable large-scale and versatile synthesis for clinical applications.

Latent Generative Models. Latent generative models [6, 7, 12, 27, 36] perform diffusion or flow matching in a compressed latent space, enabling efficient high-resolution synthesis. This paradigm builds on the VAE [25], which models continuous latent distributions. Extensions such as VQ-VAE [31] and VQ-VAE-2 [35] introduce discrete and hierarchical representations to better capture complex structures. Modern approaches typically employ KL-regularized VAEs and incorporate perceptual and adversarial objectives from VQ-GAN [11] to improve reconstruction fidelity.

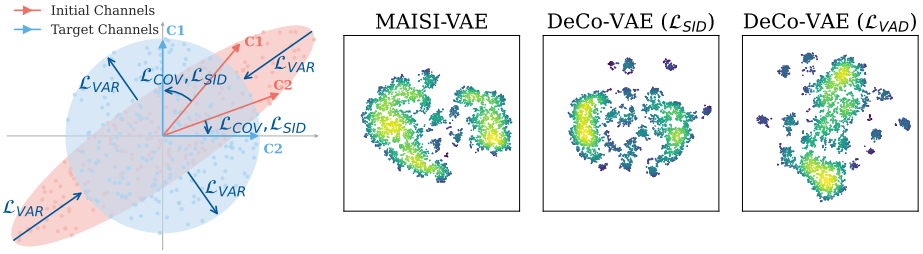
Latent Space Regularization. To ensure informative and non-degenerate latent representations, various regularization strategies have been explored. Early self-supervised methods [5, 16, 19, 42] relied on contrastive learning [8], which enforces instance discrimination but requires large batch sizes and carefully designed negative samples. To address these limitations, Barlow Twins [46] introduced a non-contrastive objective that encourages the cross-correlation matrix of latent features to approach the identity, reducing redundancy. However, decorrelation alone may lead to variance collapse. VICReg [4] resolves this by introducing an explicit variance constraint alongside covariance regularization.

3 Preliminary

We first review recent latent generative models underlying our proposed method.

3.1 3D Variational Autoencoder

We begin by formulating the standard 3D convolutional VAE used for latent space compression in medical latent generative models. The VAE framework consists of an encoder Enc_ϕ and a decoder Dec_θ. It takes each high-resolution volume $x^{(n)} \in \mathbb{R}^{1\times D\times H\times W}$ as input from a dataset $\mathcal{X} = \{x^{(n)}\}_{n=1}^{N}$ of N 3D volumes. Here, D, H, and W denote the spatial depth, height, and width.



(a) Conceptual illustration

(b) Empirical t-SNE visualizations

Fig. 1: **Overview of our latent space decorrelation.** (a) Conceptual illustration how the proposed optimization forces ($\mathcal{L}_{COV}, \mathcal{L}_{SID}$ and $\mathcal{L}_{VAD}$) reshape a skewed initial distribution into an independent target distribution. (b) Empirical 2D t-SNE visualizations demonstrating that our DeCo-VAE (middle and right) resolves the two tightly concentrated clusters observed in the baseline (left).

The encoder Enc_ϕ spatially compresses the high-dimensional input into a compact latent representation. It maps the input x to a low-dimensional latent space $z \in \mathbb{R}^{c \times d \times h \times w}$. The total dimensionality of this latent space is significantly smaller than the input space ($c \cdot d \cdot h \cdot w \ll 1 \cdot D \cdot H \cdot W$). This spatial downsampling ratio is defined as the compression factor f, where $f = D/d = H/h = W/w$. The encoder parameterizes the approximate posterior distribution $q_\phi(z|x) = \mathcal{N}(z; \mu, \text{diag}(\sigma^2))$. Using the reparameterization trick, a latent vector is sampled as $z = \mu + \sigma \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. Subsequently, the decoder Dec_θ reconstructs the volume $\hat{x} = \text{Dec}_\theta(z)$ from this latent code.

The VAE training objective balances voxel-wise reconstruction fidelity with latent space regularization. This composite objective is typically formulated as:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{Recon}} + \lambda_{\text{Adv}}\mathcal{L}_{\text{Adv}} + \lambda_{\text{Perc}}\mathcal{L}_{\text{Perc}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}}. \tag{1}$$

The term $\mathcal{L}_{\text{Recon}} = \|x - \hat{x}\|_1$ enforces voxel-wise fidelity. The adversarial loss $\mathcal{L}_{\text{Adv}}$ [15] is aggregated from a 2D PatchGAN discriminator Disc_{2D} (operating on 2D slices) and a 3D PatchGAN discriminator Disc_{3D} (operating on 3D volumetric patches). The perceptual loss $\mathcal{L}_{\text{Perc}}$ utilizes LPIPS [47] to capture feature-level similarities using a pre-trained network. Finally, the Kullback-Leibler divergence term $\mathcal{L}_{\text{KL}} = D_{KL}(q_\phi(z|x) \| \mathcal{N}(0, I))$ regularizes the latent distribution toward a standard normal prior.

To further enhance reconstruction quality, a post-hoc decoder fine-tuning (DFT) step is frequently employed. After the initial end-to-end VAE training, the encoder Enc_ϕ is frozen to preserve the learned latent structure. The decoder Dec_θ is then fine-tuned with a stronger emphasis on the perceptual loss ($\mathcal{L}_{\text{Perc}}$). This stage helps the decoder better recover high-frequency anatomical details while preserving the learned latent structure.

Table 1: **Quantitative comparison of latent channel capacity.** DeCo-VAE ($\mathcal{L}_{\text{SID}}$) reduces cosine similarity while preserving the natural data magnitude. DeCo-VAE ($\mathcal{L}_{\text{VAD}}$) both reduces inter-channel correlation and increases the effective rank. Arrows indicate the preferred direction.

Model	Cosine Sim. ↓	Effective Rank ↑
MAISI-VAE	0.2292	3.22
DeCo-VAE ($\mathcal{L}_{\text{SID}}$)	0.1148	3.37
DeCo-VAE ($\mathcal{L}_{\text{VAD}}$)	0.0971	3.97

3.2 Flow Matching Generative Model

After training the VAE, we extract the latent representation and then train a 3D convolutional U-Net [37] model using conditional flow matching (CFM) [28]. Let $x_0 \sim p_0 = \mathcal{N}(0, I)$ be standard noise and $x_1 \sim q(z)$ be a latent vector from the optimized VAE. We define a probability path p_t interpolating between x_0 and x_1 for time $t \in [0, 1]$. The objective is to learn a vector field $v_t(x)$ that generates this path:

$$\mathcal{L}_{FM} = \mathbb{E}_{t,x_0,x_1}\left[\|v_t(\psi_t(x_0, x_1)) - u_t(x_0, x_1)\|^2\right], \tag{2}$$

where $\psi_t(x_0, x_1) = (1 - t)x_0 + tx_1$ is the linear interpolation and $u_t(x_0, x_1) = x_1 - x_0$ is the target velocity.

4 Proposed Method

4.1 Motivation: Underutilized Latent Channels

In 3D medical imaging, voxel-level preservation of fine-grained anatomical structures such as cortical boundaries is far more critical than in the natural image domain, leading prior work [18, 48] to adopt a more conservative compression factor f than natural images. However, since 3D brain MRIs share high structural similarity across subjects, this choice yields a latent space that far exceeds the intrinsic complexity of the data. Consequently, rather than distributing information uniformly across all available channels, we discover that the encoder tends to rely on a small subset of dominant channels. This leaves the remaining capacity unused and results in highly correlated feature representations, as illustrated by the red-colored initial distribution in Fig. 1a.

To gauge how much the latent channels are utilized, we employ two voxel-wise metrics. Channel cosine similarity calculates the average absolute cosine similarity between all pairs of latent channels, measuring information overlap. Effective rank is computed using the entropy of the normalized singular values from the channel covariance matrix, quantifying the true dimensionality utilized by the network. As shown in Table 1, the baseline MAISI-VAE [18, 48] exhibits an average cosine similarity of 0.2292 and an effective rank of 3.22 relative to the allocated channel capacity. This indicates that the encoder does not fully utilize

the space independently, instead encoding overlapping information across multiple channels. As depicted in Fig. 1b (left), 2D t-SNE visualizations of latent code distribution further reveal that the baseline MAISI-VAE exhibits two dominant large clusters tightly concentrated in the center of the embedding space. This suggests that the encoder collapses into a small number of redundant representations. This redundancy is not merely an inefficiency—when multiple channels capture overlapping features, the latent representations of different subjects become less distinguishable. This makes it harder for the downstream generative model to learn the target data distribution, suggesting a framework for explicit latent space decorrelation.

4.2 Channel-Wise Latent Space Decorrelation

To address the observed inter-channel redundancy and its adverse effect on downstream generation, we propose regularization losses. These losses encourage the encoder to distribute information more uniformly across all c channels, promoting a more effective utilization of the allocated latent capacity.

Scale-Invariant Decorrelation. To explicitly suppress inter-channel redundancy, we first propose a scale-invariant decorrelation loss ($\mathcal{L}_{\text{SID}}$). At a glance, we might be tempted to simply penalize the raw covariances. However, this allows the network to trivially satisfy the loss by shrinking the variance of weaker channels toward zero—rather than producing more distinguishable latent representations. To prevent this scale collapse, $\mathcal{L}_{\text{SID}}$ directly penalizes the Pearson correlation matrix R. Since R normalizes covariance by the standard deviations of the channels, it measures pure linear dependence regardless of scale:

$$\mathcal{L}_{\text{SID}} = \frac{1}{c^2} \sum_{i \neq j} R_{ij}^2. \tag{3}$$

Conceptually, as depicted by the $\mathcal{L}_{\text{SID}}$ force in Fig. 1a, this loss straightens the skewed initial channels without artificially inflating their variances. As shown in Table 1, $\mathcal{L}_{\text{SID}}$ halves the average cosine similarity to 0.1148, largely reducing structural redundancies. Meanwhile, the effective rank remains stable at 3.37, suggesting that the data's inherent scale is preserved without directly constraining channel magnitudes. As observed in Fig. 1b (middle), $\mathcal{L}_{\text{SID}}$ pulls apart two dominant large clusters into several distinct, small clusters and spreads them further apart, while maintaining the overall scale and footprint of the baseline distribution. This separation suggests that the latent representations become naturally more distinguishable from one another.

Variance-Anchored Decorrelation. To suppress inter-channel redundancy, we additionally propose the variance-anchored decorrelation loss ($\mathcal{L}_{\text{VAD}}$). We first formulate a covariance loss to directly penalize the off-diagonal elements of the latent covariance matrix Σ:

$$\mathcal{L}_{\text{COV}} = \frac{1}{c^2} \sum_{i \neq j} \Sigma_{ij}^2. \tag{4}$$

However, as we noted in the scale-invariant decorrelation loss section, minimizing $\mathcal{L}_{\text{COV}}$ alone risks scale collapse. Inspired by VICReg [4], we address this by introducing a channel-wise variance loss, $\mathcal{L}_{\text{VAR}}$, as a stabilizing counterforce. It explicitly maintains the variance of each channel to remain above a target threshold γ, preventing the scale from shrinking to zero:

$$\mathcal{L}_{\text{VAR}} = \frac{1}{c}\sum_{i=1}^{c} \max(0,\, \gamma - \sqrt{\Sigma_{ii}})^2. \tag{5}$$

Combining these two terms yields the variance-anchored decorrelation loss:

$$\mathcal{L}_{\text{VAD}} = \mathcal{L}_{\text{COV}} + \mathcal{L}_{\text{VAR}}. \tag{6}$$

Conceptually illustrated as navy-colored arrows in Fig. 1a, this combined objective enforces strict orthogonalization: $\mathcal{L}_{\text{COV}}$ straightens the axes to ensure statistical independence, while $\mathcal{L}_{\text{VAR}}$ anchors the structural scale to a predefined target. As shown in Table 1, $\mathcal{L}_{\text{VAD}}$ achieves the lowest cosine similarity of 0.0971, effectively reducing inter-channel correlations. Furthermore, because all channels are driven toward the same variance target, the effective rank increases to 3.97, approaching the maximum possible effective rank for a 4-channel latent space. This capacity utilization is reflected in Fig. 1b (right). It fragments the two dense initial clusters into finer, widely dispersed small clusters, achieving a highly space-filling distribution. This well-separated latent geometry provides a better-conditioned latent representation for downstream generative modeling.

Overall Objective. The final training objective integrates our proposed decorrelation loss with the standard VAE loss. By incorporating either $\mathcal{L}_{\text{SID}}$ or $\mathcal{L}_{\text{VAD}}$ into Eq. 1, we formulate the complete loss function:

$$\mathcal{L} = \mathcal{L}_{\text{VAE}} + \lambda_{\text{DeCo}}\mathcal{L}_{\text{DeCo}}, \tag{7}$$

where $\mathcal{L}_{\text{DeCo}} \in \{\mathcal{L}_{\text{SID}}, \mathcal{L}_{\text{VAD}}\}$, and λ_{DeCo} controls the relative importance of the two losses. This combined objective ensures that all c latent channels remain statistically independent and fully utilized.

5 Experiments

5.1 Experimental Setup

Dataset. We utilize a large-scale dataset of T1-weighted 3D brain MRI volumes from the UK Biobank [3, 40] (Field ID 20252). To ensure a healthy cohort, we exclude subjects with known structural brain pathologies based on the ICD-10 [32] diagnostic codes. The final curated dataset consists of 25,252 healthy samples. All volumes are inherently pre-registered to the standard MNI-152 space [13]. To manage the high dimensionality of the 3D data and optimize computational efficiency, the original $182 \times 218 \times 182$ volumes are resized to a fixed input resolution of $128 \times 256 \times 128$.

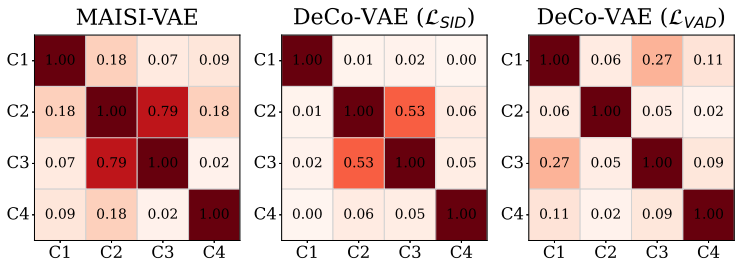


Fig. 2: **Channel-wise absolute cosine similarity matrices of the latent representations.** Both proposed methods suppress high off-diagonal correlations compared to the baseline (left). Notably, DeCo-VAE ($\mathcal{L}_{SID}$) (middle) retains meaningful inherent correlations, whereas DeCo-VAE ($\mathcal{L}_{VAD}$) (right) substantially reduced inter-channel correlation.

Training Configuration. The VAE maps the input to a latent space of size $4\times32\times64\times32$. For training, we use 96^3 patches with a batch size of 2 per GPU for 100,000 iterations. λ_{DeCo} is set to 10 for L_{SID}, and 1 for L_{VAD}. Subsequently, the CFM model is trained on the latent representations for 100,000 iterations with a batch size of 4. Finally, we perform decoder fine-tuning for 10,000 iterations to further enhance the visual fidelity. All experiments are conducted on a computing cluster comprising 2 nodes with a total of 8 NVIDIA RTX 3090 GPUs. Details are in our GitHub repository (https://github.com/Stomper10/decovae).

Evaluation Metrics. The VAE reconstruction fidelity is measured by peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and LPIPS [47]. Generative performance is evaluated using the reconstructed Fréchet Inception Distance (rFID) and standard generation FID (gFID) [20]. To accurately capture domain-specific anatomical features, both rFID and gFID are computed on 2D slices across three anatomical planes using a RadImageNet [29] pre-trained feature extractor.

5.2 Latent Channel Analysis

Following the formulation in Section 4.1, our latent channel analysis suggests that the baseline MAISI-VAE [18, 48] under-utilizes its 4-channel capacity and suffers from structural redundancy. Fig. 2 (left) visually supports this overlapping information, showing high off-diagonal correlations, such as 0.78 between C2 and C3. Consequently, Fig. 3 shows the representation exhibits skewed variance, leaving PC4 with a mere 4.5% of the total variance.

Our proposed DeCo-VAE ($\mathcal{L}_{\text{SID}}$) substantially reduces this redundancy by directly penalizing scale-invariant correlation. Fig. 2 (middle) reveals that while major overlaps are suppressed, it retains inherent data correlations (*e.g.*, a 0.54 correlation between C2 and C3) rather than forcing a complete orthogonalization. Because it does not explicitly enforce a variance target, the resulting variance distribution preserves the natural data magnitude, leaving PC4 at 5.1%

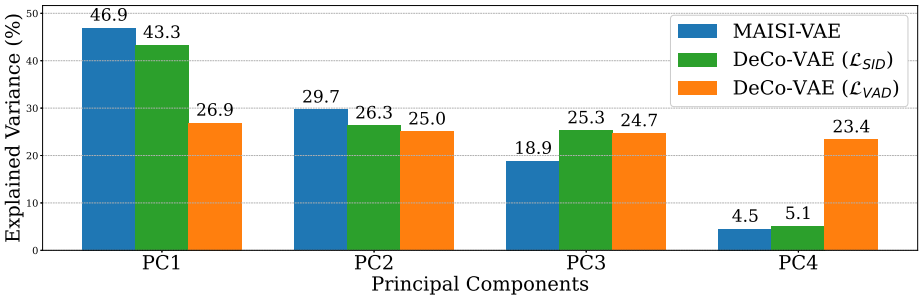


Fig. 3: **Comparison of the explained variance ratio (%) across the four principal components.** While the baseline exhibits skewed variance across components, DeCo-VAE ($\mathcal{L}_{\text{SID}}$) preserves the natural variance hierarchy, and DeCo-VAE ($\mathcal{L}_{\text{VAD}}$) achieves a balanced variance distribution.

Table 2: **Quantitative evaluation of VAE reconstruction performance.** DeCo-VAE ($\mathcal{L}_{\text{SID}}$) achieves the highest fidelity across all metrics by reducing structural redundancy while preserving the natural magnitude of the data. rFID score is computed on 2D slices along three anatomical planes and their average. Arrows indicate the preferred direction. Bold values indicate the **best** results and second-best are underlined.

VAE Model	PSNR ↑	SSIM ↑	LPIPS ↓	rFID (Avg.) ↓
MAISI-VAE	32.80	0.971	0.027	6.426
DeCo-VAE ($\mathcal{L}_{SID}$)	**33.77**	**0.981**	**0.022**	**5.822**
DeCo-VAE ($\mathcal{L}_{VAD}$)	33.64	0.979	0.023	5.946

as shown in Fig. 3. By encouraging less redundant feature organization while preserving this relative variance structure more naturally, it provides a more favorable latent code for the decoder to achieve high-fidelity reconstruction.

Conversely, DeCo-VAE ($\mathcal{L}_{\text{VAD}}$) anchors the variance to a predefined target to increase latent capacity utilization. It achieves substantially reduced inter-channel correlation as shown in Fig. 2 (right). Furthermore, because all channels are forced to the same variance target, the variance is now evenly distributed from PC1 (26.9%) to PC4 (23.4%) as seen in Fig. 3. This balanced variance distribution leads to a better-utilized latent space, providing a more favorable target distribution for the downstream generative model to learn.

5.3 Reconstruction and Generation Performance

As predicted by our latent channel analysis, DeCo-VAE ($\mathcal{L}_{SID}$) improves empirical VAE reconstruction performance. By reducing structural redundancy while preserving the data's natural magnitude, it provides a stable representation for the decoder. Table 2 shows that it achieves the highest PSNR (33.77) and SSIM

Table 3: **Quantitative evaluation of downstream generative performance.** DeCo-VAE ($\mathcal{L}_{VAD}$) achieves the best gFID scores across all anatomical planes, demonstrating that its latent space is structurally closer to the assumed latent prior. gFID scores are computed on 2D slices along three anatomical planes. All models use 30 inference steps. Arrows indicate the preferred direction. Bold values indicate the **best** results and second-best are underlined.

VAE Model	FID (Axial)↓	FID (Coronal)↓	FID (Sagittal)↓	FID (Avg.)↓
MAISI-VAE	6.566	7.808	5.424	6.599
DeCo-VAE ($\mathcal{L}_{SID}$)	<u>6.102</u>	<u>7.498</u>	<u>4.431</u>	<u>6.010</u>
DeCo-VAE ($\mathcal{L}_{VAD}$)	**6.014**	**7.485**	**4.413**	**5.971**

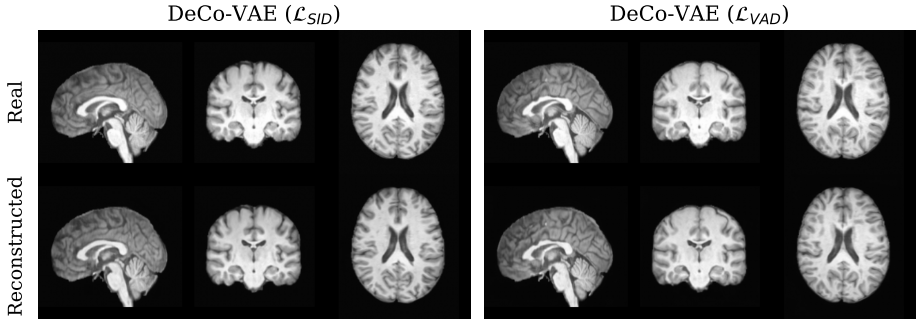


Fig. 4: **Qualitative comparison of reconstruction.** Real 3D MRI volumes and VAE reconstructed samples are visually compared. Reconstructed slices closely match the real anatomical structures across three anatomical planes.

(0.981), alongside the lowest LPIPS (0.022) and rFID (5.822). The qualitative results in Fig. 4 (left) further support this, showing that the reconstructed slices closely match the real anatomical structures.

For downstream generation, the latent space of DeCo-VAE ($\mathcal{L}_{VAD}$) demonstrates a distinct advantage. Because $\mathcal{L}_{VAD}$ forces the latent channels to be less correlated and more uniformly distributed, the resulting data distribution better matches the independence and scale assumptions of the prior. This structural alignment may reduce the difficulty of the mapping task, appearing to facilitate the generative model in learning the target distribution. Table 3 shows that it lowers the average gFID from 6.599 to 5.971, achieving consistent improvements across all anatomical planes (axial, coronal, and sagittal). As shown in Fig. 5, this quantitative improvement translates directly into enhanced visual fidelity. First, our DeCo-VAE ($\mathcal{L}_{\text{VAD}}$) resolves the *structural blurring* typically observed in the baseline. For instance, it recovers the detailed, branching architecture of the cerebellum in the sagittal plane. Second, our method mitigates *vertical grid-like artifacts* that compromise structural integrity. In the coronal plane, for example, it repairs artificial disconnections within continuous white matter, en-

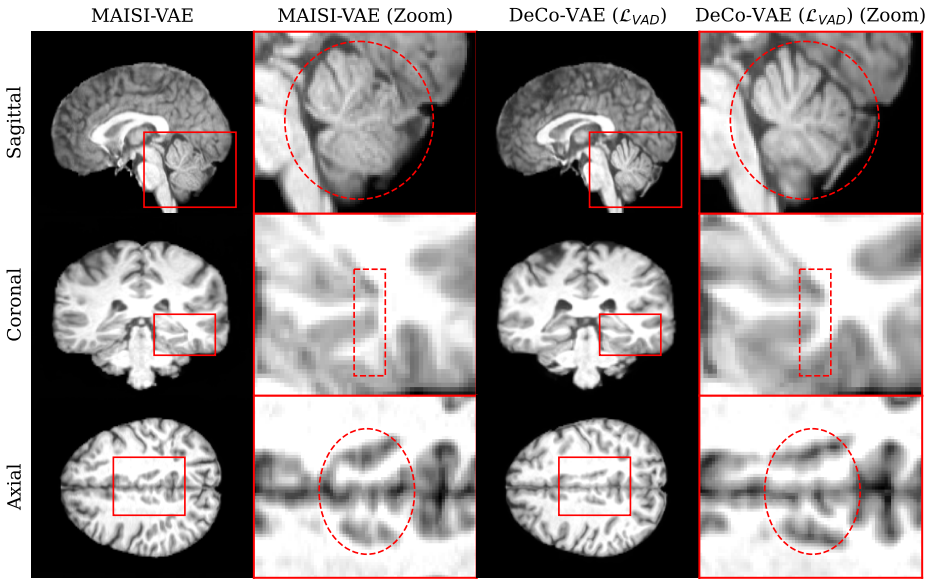


Fig. 5: **Qualitative comparison of generation.** DeCo-VAE ($\mathcal{L}_{\text{VAD}}$) resolves the structural blurring in the cerebellum (sagittal) and mitigates the vertical grid-like artifacts that harm structural continuity (coronal and axial). Both volumes are generated from the same seed.

suring smooth tissue transitions. Similarly, in the axial plane, it prevents these grid artifacts from cluttering the dark cerebrospinal fluid (CSF) spaces, maintaining clear boundaries where anatomical separation is required.

These improvements in FID represent a meaningful advance for 3D brain imaging, where preserving voxel-level quality and fine anatomical boundaries is essential. While generated images lack ground-truth reference volumes to compute direct voxel-wise metrics, the reconstruction fidelity provides a useful indirect indicator of this voxel-level quality. As visually supported in Fig. 4 and Fig. 5, both the reconstructed and generated outputs from our proposed methods visually retain sharp, well-defined cortical structures, suggesting that the generative pipeline preserves fine anatomical structure.

5.4 Ablation Study

Decoder Fine-Tuning. To improve visual fidelity and capture high-frequency details, we perform VAE decoder fine-tuning (DFT) with a stronger emphasis on the perceptual loss while freezing the encoder. As shown in Table 4, applying DFT is not universally effective for all models. For the baseline $\mathcal{L}_{\text{VAE}}$, DFT provides negligible improvements in generation performance (gFID from 6.599 to 6.594). However, our regularized spaces provide a stable and disentangled foundation that effectively benefits from further enhancement. This behavior is

Table 4: **Ablation study on decoder fine-tuning.** To investigate the effect of decoder fine-tuning (DFT), we compared the performance before and after its application. While applying DFT improves both reconstruction and generation performance in all cases, it does not overturn the pre-existing competitive advantages. FID scores are computed on 2D slices along three anatomical planes. All models use 30 inference steps. Arrows indicate the preferred direction. Bold values indicate the **best** results and second-best are underlined.

	Reconstruction				**Generation** (gFID ↓)			
Loss	PSNR ↑	SSIM ↑	LPIPS ↓	Avg. rFID ↓	Axial	Coronal	Sagittal	Avg.
$\mathcal{L}_{\text{VAE}}$	32.80	0.971	0.027	6.426	6.566	7.808	5.424	6.599
$\mathcal{L}_{\text{VAE}}$ + DFT	32.99	0.969	0.025	6.243	6.577	7.850	5.356	6.594
$\mathcal{L}_{\text{SID}}$	33.47	0.980	0.024	6.240	6.380	7.719	4.983	6.361
$\mathcal{L}_{\text{SID}}$ + DFT	**33.77**	**0.981**	**0.022**	**5.822**	6.102	7.498	4.431	6.010
$\mathcal{L}_{\text{VAD}}$	33.17	0.978	0.025	6.251	6.270	7.651	4.886	6.269
$\mathcal{L}_{\text{VAD}}$ + DFT	33.64	0.979	0.023	5.946	**6.014**	**7.485**	**4.413**	**5.971**

Table 5: **Ablation study on covariance and variance regularizer.** All methods share the same MAISI-VAE architecture with different regularizations added on top of $\mathcal{L}_{\text{VAE}}$. An explicit variance constraint is beneficial to improve both reconstructive and generative capabilities. FID scores are computed on 2D slices along three anatomical planes. All models use 30 inference steps. Arrows indicate the preferred direction. Bold values indicate the **best** results.

Loss	**Reconstruction**				**Generation** (gFID ↓)			
	PSNR ↑	SSIM ↑	LPIPS ↓	Avg. rFID ↓	Axial	Coronal	Sagittal	Avg.
$\mathcal{L}_{VAE}$	32.99	0.969	0.025	6.243	6.577	7.850	5.356	6.594
$+\mathcal{L}_{COV}$	**33.77**	0.963	0.024	6.015	6.334	7.644	4.789	6.256
$+\mathcal{L}_{\text{COV}} + \mathcal{L}_{\text{VAR}}$	33.64	**0.979**	**0.023**	**5.946**	**6.014**	**7.485**	**4.413**	**5.971**

consistent with the view that our proposed regularizers provide a better, more distinguishable latent distribution not only for the downstream generative model but also for the decoder itself. When fine-tuned, $\mathcal{L}_{\text{SID}}$+DFT achieves the highest reconstruction scores, reaching a PSNR of 33.77 and an rFID of 5.822. For the generative pipeline, $\mathcal{L}_{\text{VAD}}$+DFT yields the lowest average gFID of 5.971 in 3D brain MRI generation.

Necessity of Variance Constraints. We ablate our regularizers by replacing them with a naive covariance loss ($+\mathcal{L}_{\text{COV}}$). This approach penalizes off-diagonal elements of the covariance matrix but lacks an explicit variance constraint (such as scale-invariant normalization or variance anchoring). This allows the network to take an optimization shortcut, minimizing the loss by shrinking the variance of weaker channels toward zero. Due to this dimensional collapse, the performance gains are restricted, and structural coherence fails to improve over the baseline (with SSIM remaining at 0.963 compared to 0.969) as shown in Table 5. Incorporating the variance loss ($+\mathcal{L}_{\text{COV}} + \mathcal{L}_{\text{VAR}}$) stabilizes the latent space and prevents this shortcut. This combination restores the SSIM to 0.979 and achieves

the lowest rFID (5.946) and gFID (5.971). This demonstrates that while a covariance loss provides initial benefits, an explicit variance constraint is beneficial to improve both reconstructive and generative capabilities.

6 Conclusion

In this paper, we identified latent space under-utilization as a bottleneck in our setting for 3D medical latent generative models. The inherent structural similarity of 3D brain MRIs, combined with conservative compression rates, yields redundant representations that make it harder for downstream generative models to effectively learn the data distribution.

To address this, we proposed DeCo-VAE, introducing two complementary channel-wise regularizers. The scale-invariant decorrelation loss ($\mathcal{L}_{\text{SID}}$) reduces redundancy while preserving the relative variance structure more naturally, supporting high-fidelity reconstruction. Alternatively, the variance-anchored decorrelation loss ($\mathcal{L}_{\text{VAD}}$) enforces low correlation and uniform variance, substantially increasing latent capacity utilization to improve downstream generation quality.

Our experiments support our central hypothesis, demonstrating the distinct advantages of each loss. By achieving consistent performance gains without architectural modifications, we show that latent space utilization is an important, underexamined aspect of latent generative model performance. We hope DeCo-VAE highlights the importance of latent space design and serves as a practical foundation for more expressive and efficient 3D medical image synthesis.

7 Limitation

While our experiments demonstrate the effectiveness of channel-wise latent decorrelation, a few aspects fall outside the scope of this study. First, our evaluation relies on a relatively limited set of comparison models, and a broader benchmark would further contextualize our gains. Second, we center on latent utilization, reconstruction, and generative quality, while assessing the utility of the generated data on downstream tasks such as segmentation or classification remains a valuable next step. Third, we intentionally focus on healthy, T1-weighted brain MRI, where the anatomical similarity that motivates our method is most pronounced; extending to other modalities, pathological cohorts, and more heterogeneous populations is a natural direction for future work.

Acknowledgments. This work was supported by Korea Health Technology R&D Project through the Korea Health Industry Development Institute (RS-2025-02307233), Samsung Electronics, Youlchon Foundation, National Research Foundation of Korea (NRF) grants (RS-2021-NR05515, RS-2024-00336576, RS-2023-0022663), and the Institute for Information & Communication Technology Planning & Evaluation (IITP) grants (RS-2022-II220264, RS-2024-00353131) funded by the Korean government. This research has been conducted using the UK Biobank Resource under application number 45227.

References

1. Ahmad, W., Ali, H., Shah, Z., Azmat, S.: A new generative adversarial network for medical images super resolution. Scientific Reports **12**(1), 9533 (2022)
2. Ahn, S., Park, W., Cho, J., Park, J.: Volumetric conditioning module to control pretrained diffusion models for 3d medical images. In: WACV (2025)
3. Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., et al.: Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. Neuroimage **166**, 400–424 (2018)
4. Bardes, A., Ponce, J., LeCun, Y.: Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In: ICLR (2022)
5. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS (2020)
6. Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., Li, Z.: Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In: ECCV (2024)
7. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: ICLR (2024)
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS (2021)
10. Dorjsembe, Z., Pao, H., Odonchimed, S., Xiao, F.: Conditional diffusion models for semantic 3d brain mri synthesis. IEEE Journal of Biomedical and Health Informatics (2024)
11. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR (2021)
12. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
13. Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L., Group, B.D.C., et al.: Unbiased average age-appropriate atlases for pediatric studies. Neuroimage **54**(1), 313–327 (2011)
14. Gong, K., Johnson, K., El Fakhri, G., Li, Q., Pan, T.: Pet image denoising based on denoising diffusion probabilistic model. European Journal of Nuclear Medicine and Molecular Imaging **51**(2), 358–368 (2024)
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
16. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent: A new approach to self-supervised learning. In: NeurIPS (2020)
17. Güngör, A., Dar, S.U., Öztürk, Ş., Korkmaz, Y., Bedel, H.A., Elmas, G., Ozbey, M., Çukur, T.: Adaptive diffusion priors for accelerated mri reconstruction. Medical image analysis **88**, 102872 (2023)
18. Guo, P., Zhao, C., Yang, D., Xu, Z., Nath, V., Tang, Y., Simon, B., Belue, M., Harmon, S., Turkbey, B., et al.: Maisi: Medical ai for synthetic imaging. In: WACV (2025)

19. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
20. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
22. Hong, S., Marinescu, R., Dalca, A.V., Bonkhoff, A.K., Bretzner, M., Rost, N.S., Golland, P.: 3d-stylegan: A style-based generative adversarial network for generative modeling of three-dimensional medical images. In: MICCAI Workshop (2021)
23. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. In: NeurIPS (2022)
24. Khader, F., Müller-Franzes, G., Tayebi Arasteh, S., Han, T., Haarburger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baeßler, B., Foersch, S., et al.: Denoising diffusion probabilistic models for 3d medical image generation. Scientific reports **13**(1), 7303 (2023)
25. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
26. Kwon, G., Han, C., Kim, D.: Generation of 3d brain mri using auto-encoding generative adversarial networks. In: MICCAI (2019)
27. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
28. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2023)
29. Mei, X., Liu, Z., Robson, P.M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K.E., Yang, T., et al.: Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. Radiology: Artificial Intelligence **4**(5), e210315 (2022)
30. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML (2021)
31. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. In: NeurIPS (2017)
32. Organization, W.H.: International Statistical Classification of Diseases and related health problems: Alphabetical index, vol. 3. World Health Organization (2004)
33. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
34. Pinaya, W.H., Tudosiu, P.D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J.: Brain imaging generation with latent diffusion models. In: MICCAI Workshop (2022)
35. Razavi, A., van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. In: NeurIPS (2019)
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
37. Ronneberger, O., Fischer, P., Brox, T.: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
38. Roy, O., Vetterli, M.: The effective rank: A measure of effective dimensionality. In: 2007 15th European signal processing conference. pp. 606–610. IEEE (2007)
39. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2020)

40. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al.: Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS medicine **12**(3), e1001779 (2015)
41. Sun, L., Chen, J., Xu, Y., Gong, M., Yu, K., Batmanghelich, K.: Hierarchical amortized gan for 3d high resolution medical image synthesis. IEEE journal of biomedical and health informatics **26**(8), 3966–3975 (2022)
42. Tian, Y., Krishnan, D., Isola, P.: What makes for good views for contrastive learning? In: NeurIPS (2020)
43. Wang, H., Liu, Z., Sun, K., Wang, X., Shen, D., Cui, Z.: 3d meddiffusion: A 3d medical latent diffusion model for controllable and high-quality medical image generation. IEEE Transactions on Medical Imaging (2025)
44. Yang, H., Sun, J., Carass, A., Zhao, C., Lee, J., Prince, J.L., Xu, Z.: Unsupervised mr-to-ct synthesis using structure-constrained cyclegan. IEEE transactions on medical imaging **39**(12), 4249–4261 (2020)
45. Yurt, M., Dar, S.U., Erdem, A., Erdem, E., Oguz, K.K., Çukur, T.: mustgan: multi-stream generative adversarial networks for mr image synthesis. Medical image analysis **70**, 101944 (2021)
46. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Jegelka, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: ICML (2021)
47. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
48. Zhao, C., Guo, P., Yang, D., He, Y., Tang, Y., Simon, B., Belue, M., Harmon, S., Turkbey, B., Xu, D.: Maisi-v2: Accelerated 3d high-resolution medical image synthesis with rectified flow and region-specific contrastive loss. In: AAAI (2026)
49. Zhu, L., Xue, Z., Jin, Z., Liu, X., He, J., Liu, Z., Yu, L.: Make-a-volume: Leveraging latent diffusion models for cross-modality 3d brain mri synthesis. In: MICCAI (2023)
50. Zuo, L., Dewey, B.E., Carass, A., He, Y., Shao, M., Reinhold, J.C., Prince, J.L.: Synthesizing realistic brain mr images with noise control. In: International Workshop on Simulation and Synthesis in Medical Imaging. Springer (2020)