# GM-LDM: LATENT DIFFUSION MODEL FOR BRAIN BIOMARKER IDENTIFICATION THROUGH FUNCTIONAL DATA-DRIVEN GRAY MATTER SYNTHESIS

*Xu Hu*[*1], *Jingling Yang*[*1], *Sihan Jia*[2], *Yuda Bi*[2], *and Vince D Calhoun*[2]

*Anhui University, Hefei, China*[1]
*TReNDS Center (GSU, Gatech, Emory), Atlanta, GA, USA*[2]

## ABSTRACT

Deep learning-based generative models have been widely applied in medical imaging and multimodal scenarios, such as neural modality transformations and exploring relationships between structural and functional images. In this study, we propose a universal framework named GM-LDM that integrates the latent diffusion model into the medical imaging domain, with a particular focus on MRI-based brain imaging. Our framework primarily incorporates a 3D autoencoder trained on large-scale MRI datasets, using KL divergence loss to achieve statistical consistency. In addition, we use a Vision Transformer-based encoder-decoder architecture as a denoising network within the diffusion model. The framework also supports the integration of other conditions, such as functional MRI data, providing flexibility and adaptability to different imaging tasks.

***Index Terms*—** Magnetic resonance imaging, latent diffusion model, vision transformer, generative model

## 1. INTRODUCTION

Generative models, such as GANs [1] and diffusion models [2], have gained widespread use in medical imaging, showing promising potential in MRI-based neuroimaging [3]. These models can effectively fuse different modalities, such as combining T1- and T2-weighted images, and offer significant clinical applications, such as converting MRI to CT images, allowing patients to avoid radiation-based CT scans [4]. Among generative models, GANs stand out for their ability to synthesize high-quality images, adapt to various imaging tasks [5], and generate realistic images from noise, though they face challenges like mode collapse, unstable training, and requiring large labeled datasets [6]. Diffusion models, especially the latent diffusion model (LDM) [7], have addressed many of these issues by utilizing a pre-trained autoencoder to map images from data space to latent space, enhancing training and inference efficiency. However, while autoencoders have excelled in natural image domains due to abundant training data, in medical imaging, particularly MRI, obtaining large datasets for pre-training remains challenging, limiting the full potential of diffusion models in this field [8, 9, 10, 11].

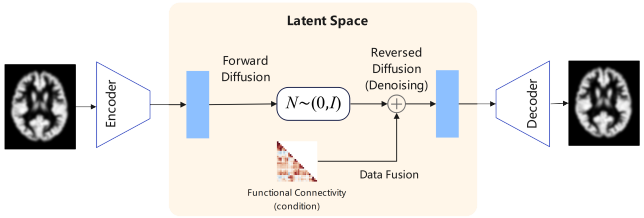


**Fig. 1**. The overall pipeline of our latent diffusion model guided with conditions

We made several key contributions to this study. (1) We developed a 3D gray matter generation framework based on LDM. This framework incorporates a 3D autoencoder pre-trained on the large-scale MRI dataset ABCD. The autoencoder efficiently and accurately extracts gray matter (GM) features and maps them into a reduced latent space. This autoencoder serves as a benchmark and basic model that can be applied to various downstream tasks. In addition, we introduced learnable interpolation layers at both ends of the autoencoder, which allows the model to adapt to medical images of any shape, making it more flexible for different imaging modalities. (2) We proposed a hybrid denoising framework based on CNN and vision transformer (ViT) [12]. The backbone of this model is a ViT encoder-decoder architecture, which receives state features extracted by a CNN at different stages. This allows effective transformation and guidance between the state and target generation modalities, improving the overall quality and control of the generated outputs. (3) In the results, we tested this model using functional brain network connectivity (FNC) as a condition for generating subject-wise 3D GM data. This experiment was designed to validate previously identified biomarkers associated with schizophrenia. The results show that the model can serve as a tool to help locate brain disease targets, highlighting its potential application in identifying disease-specific biomarkers for conditions such as schizophrenia.

*These authors contributed equally

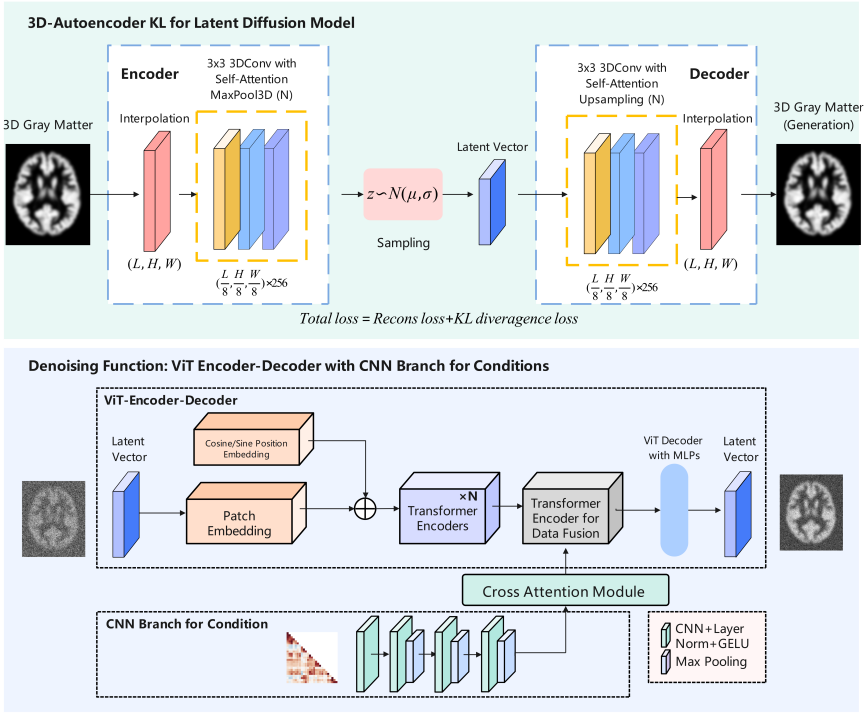


**Fig. 2**. The architectures of 3D-autoencoder with kl divergence loss and multimodal denoising network via ViT, CNN, and cross-attention module.

## 2. RELATED WORKS

Jiang et al. [13] introduce a conditioned LDM for multi-modal MRI synthesis, operating in latent space to reduce memory usage, with structural guidance via brain region masks to maintain anatomical details. Kim et al. [14] propose an adaptive LDM (ALDM) for 3D MRI translation, enabling multi-modal translations from a single source, outperforming other models. Pan et al. [15] present MRI-to-CT denoising diffusion (MC-DDPM), using a Swin-Vnet-based reverse diffusion process to generate high-quality synthetic CT images from MRI data. In reviewing recent studies, we find that while these models introduce innovative architectures, certain limitations remain. For example, the lack of a portable autoencoder for large-scale training and the limited exploration of links between generative models and biomarker discovery represent areas for further improvement.

## 3. METHODS

### 3.1. Latent Diffusion Models

Let $\mathbf{x}_0 \in \mathscr{X}$ represent the input data in the data space $\mathscr{X}$, and $\mathbf{z}_0 \in \mathscr{Z}$ denote its latent representation in the latent space $\mathscr{Z}$, obtained through an encoding function $E : \mathscr{X} \to \mathscr{Z}$. The latent diffusion process is defined as a sequence of noisy latent variables $\mathbf{z}_t$, $t = 0, 1, \ldots, T$, where noise is gradually added through a Markov process. The forward diffusion process can be expressed as:

$$q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathscr{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t \mathbf{I}),$$

where $\beta_t$ is the variance schedule that controls the amount of noise added at each step $t$. The **reverse process** aims to denoise $\mathbf{z}_T$ back to $\mathbf{z}_0$ by learning a parameterized model $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c})$, where $\mathbf{c}$ represents the conditional information. This denoising process is modeled as:

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{c}) = \mathscr{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, \mathbf{c}, t), \Sigma_\theta(\mathbf{z}_t, t)),$$

where $\mu_\theta$ and $\Sigma_\theta$ are learned mean and variance functions. After the reverse process, the final latent variable $\mathbf{z}_0$ is decoded back to the data space using the decoding function $D : \mathscr{Z} \to \mathscr{X}$, yielding the reconstructed data $\hat{\mathbf{x}}_0 = D(\mathbf{z}_0)$. Thus, the complete process showed in **Fig. 1** can be expressed as:

$$\mathbf{x}_0 \xrightarrow{E} \mathbf{z}_0 \xrightarrow{\text{diffusion}} \mathbf{z}_T \xrightarrow{\text{denoising } \mathbf{c}} \mathbf{z}_0 \xrightarrow{D} \hat{\mathbf{x}}_0(\mathbf{c}).$$

### 3.2. 3D Autoencoder

Let the input data be a 3D volume $\mathbf{x} \in \mathbb{R}^{L' \times W' \times H'}$, where $L', W', H'$ represent the initial spatial dimensions. The input is first passed through a learnable interpolation layer

$\mathscr{I}_1 : \mathbb{R}^{L' \times W' \times H'} \to \mathbb{R}^{L \times W \times H}$, transforming the input shape to $(L, W, H)$. Next, the encoder network $E$ maps the interpolated input to a feature space, producing an encoded representation $\mathbf{z} \in \mathbb{R}^{256 \times \frac{L}{8} \times \frac{W}{8} \times \frac{H}{8}}$. This encoding can be treated as the parameters of a Gaussian distribution $\mathscr{N}(\mu, \sigma)$, where:

$$\mathbf{z} \sim \mathscr{N}(\mu_E(\mathbf{x}), \sigma_E(\mathbf{x})),$$

with $\mu_E(\mathbf{x})$ and $\sigma_E(\mathbf{x})$ being the mean and variance learned by the encoder. To obtain the latent vector $\mathbf{z}_0 \in \mathbb{R}^{256 \times \frac{L}{8} \times \frac{W}{8} \times \frac{H}{8}}$, we sample from this distribution:

$$\mathbf{z}_0 = \mu_E(\mathbf{x}) + \sigma_E(\mathbf{x}) \cdot \varepsilon, \quad \varepsilon \sim \mathscr{N}(0, I).$$

The decoder network $D$ takes the latent vector $\mathbf{z}_0$ and decodes it back to the shape $\mathbb{R}^{256 \times \frac{L}{8} \times \frac{W}{8} \times \frac{H}{8}}$, which is then upsampled to $(L, W, H)$ using an upsampling function $\mathscr{U} : \mathbb{R}^{256 \times \frac{L}{8} \times \frac{W}{8} \times \frac{H}{8}} \to \mathbb{R}^{L \times W \times H}$. Finally, the output is passed through a second learnable interpolation layer $\mathscr{I}_2 : \mathbb{R}^{L \times W \times H} \to \mathbb{R}^{L' \times W' \times H'}$, returning the output to the original input shape:

$$\hat{\mathbf{x}} = \mathscr{I}_2(D(\mathbf{z}_0)) \in \mathbb{R}^{L' \times W' \times H'}.$$

### 3.3. Loss Functions

**KL Divergence Loss** is used to measure the difference between the learned latent distribution $q(\mathbf{z}|\mathbf{x})$ and the prior distribution $p(\mathbf{z})$, typically chosen as a standard normal distribution $\mathscr{N}(0, I)$. The KL divergence between these two distributions is defined as:

$$\mathscr{L}_{\text{KL}} = D_{\text{KL}}(q(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[ \log \frac{q(\mathbf{z}|\mathbf{x})}{p(\mathbf{z})} \right].$$

For a Gaussian posterior $q(\mathbf{z}|\mathbf{x}) = \mathscr{N}(\mu_E(\mathbf{x}), \sigma_E(\mathbf{x}))$ and a prior $p(\mathbf{z}) = \mathscr{N}(0, I)$, the KL divergence loss can be computed analytically as:

$$\mathscr{L}_{\text{KL}} = -\frac{1}{2} \sum_{i=1}^{d} \left( 1 + \log \sigma_E^2(\mathbf{x}_i) - \mu_E^2(\mathbf{x}_i) - \sigma_E^2(\mathbf{x}_i) \right),$$

where $d$ is the dimensionality of the latent space. This loss ensures that the learned latent space remains close to the prior distribution. The **Reconstruction Loss** measures how well the VAE can rebuild the input $\mathbf{x}$ from its latent representation:

$$\mathscr{L}_{\text{recon}} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{x}_i')^2,$$

where $N$ is the number of data points. So, The **total loss** is the sum of these two with a hyperparameter $\alpha \in [0, 1]$:

$$\mathscr{L}_{\text{total}} = \alpha \mathscr{L}_{\text{KL}} + (1 - \alpha) \mathscr{L}_{\text{recon}}.$$

### 3.4. Denoising Network

The denoising network is applied to the latent vector $\mathbf{z}_T$ and follows a ViT encoder-decoder architecture. The encoder consists of multiple transformer layers, where each layer applies a self-attention mechanism followed by a feed-forward network.

The decoder reconstructs from the latent vector $\mathbf{z}_0$ using a transformer layer for data fusion, which incorporates cross-attention (CA) mechanisms to integrate conditional information. Then it is followed by a series of paralleled MLPs for reconstruction [16]. The cross-attention layer can be formulated as:

$$\text{CA}(\mathbf{Q}, \mathbf{K}_{\text{cond}}, \mathbf{V}_{\text{cond}}) = \text{softmax} \left( \frac{\mathbf{Q}\mathbf{K}_{\text{cond}}^T}{\sqrt{d_k}} \right) \mathbf{V}_{\text{cond}},$$

where $\mathbf{K}_{\text{cond}}$ and $\mathbf{V}_{\text{cond}}$ are derived from the condition $\mathbf{c}$, extracted using a CNN. The **Fig.2** shows architectures of our 3D autoencoderKL and denoising network.

## 4. EXPERIMENTS AND RESULTS

### 4.1. Datasets

We used the large-scale ABCD dataset (n=11220) to train our 3D autoencoderKL. The ABCD dataset contains comprehensive neuroimaging and cognitive data from adolescents, providing a rich source for evaluating brain structure and function through MRI scans. For brain diseases analysis, We employed two schizophrenia-related datasets, combining data from three international studies (fBIRN, MPRC, COBRE) and several hospitals in China. The dataset included 1,642 participants, with 803 healthy controls and 839 individuals with schizophrenia. Resting-state fMRI data were collected using 3.0 Tesla scanners across multiple sites, with EPI sequences (TR/TE 2000/30 ms, voxel sizes from $3 \times 3 \times 3$ mm to $3.75 \times 3.75 \times 4.5$ mm).

### 4.2. Model Design, Training, and Validation

Our 3D autoencoderKL, designed with an encoder-decoder architecture that integrates multiple convolutional layers with attention mechanisms, was extensively pre-trained and fine-tuned using the ABCD dataset, where we employed fp16 precision for faster training, a multisteplr learning rate scheduler, and AdamW optimizer (learning rate: 2e-4, batch size: 16), deploying parallel training on four H100 GPUs. Similarly, our GM-LDM model focused on training the denoising network to generate subject-specific GM images conditioned on FNCs, employing the same configuration as the autoencoderKL but with a lower learning rate of 3e-5; this model also utilized four H100 GPUs and underwent 5-fold cross-validation for enhanced robustness and generalization. To assess model effectiveness, we further conducted baseline and

**Table 1**. Model performance details for baselines, comparisons, and our preferred model.

| Name | Model | Pre-trained | Condition | Pearson Corr |
|---|---|---|---|---|
| Baseline-1 | Diffusion | No | Random-vector | 0.78 |
| Baseline-2 | LDM | No | Random-vector | 0.79 |
| Comparison-1 | LDM | Yes | Random-vector | 0.86 |
| Comparison-2 | LDM | No | FNC | 0.83 |
| GM-LDM | LDM | Yes | FNC | 0.89 |

comparative experiments, such as training the autoencoderKL without ABCD pre-training and employing a traditional diffusion model, evaluating the similarity between generated and real GM using Pearson correlation as the primary metric due to the grayscale and structural nature of GM images. In LDM training, we also tested FNC as a conditioning factor against random noise of equivalent dimensions to determine FNC's effectiveness in guiding the generation process.

### 4.3. Basic Results

The results of our experiments are summarized in **Table 1**, where we compare the performance of our proposed model with baseline and comparison models. Our GM-LDM model achieved the highest Pearson correlation (0.89) when conditioned on schizophrenia-related FNC, outperforming models that only conditioned on random-vectors or without autoencoder pre-training. This confirms that conditioning on FNC data improves the ability of the model to generate GM images that more closely resemble actual brain structures associated with schizophrenia. For baseline models, Baseline-1 and Baseline-2 achieved Pearson correlations of 0.78 and 0.79, respectively, confirming the limited effectiveness of models trained without FNC or pre-trained data. Comparison-1, which included random vector pre-training but no FNC conditioning, achieved a higher correlation (0.86), demonstrating the benefit of pre-training but also highlighting the importance of including relevant FNC data. Comparison-2 further demonstrated this effect by using FNC conditioning without pre-training, achieving a correlation of 0.83. These results highlight the importance of pre-training with a large dataset (such as ABCD) and FNC conditioning in capturing schizophrenia-related gray matter patterns.

### 4.4. Biomarkers Discovery

Through our comparison of GM generated under the guidance of FNC from subjects with schizophrenia versus GM generated with random vector guidance, we observed that certain brain regions were more prominent in the FNC-guided GM. The **Fig.3** shows the GM we generated and the brain regions that related to schizophrenia. By analyzing these regions, we found similarities with biomarkers associated with schizophrenia that were identified in our previous research [3], where GM was used as a conditional input to generate FNC matrices for schizophrenia subjects, and attention weights were applied to create GM saliency maps. These saliency maps highlighted critical brain regions, offering insights into the spatial importance within GM structures that contribute to FNC patterns in schizophrenia. In this study, we not only identified a strong link between the cerebellum and schizophrenia but also observed significant associations between schizophrenia and specific basal ganglia structures, such as the Caudate and Putamen, which have traditionally been implicated in cognition and emotion. Our model demonstrates potential for further application in identifying regions of interest (ROIs) associated with various brain disorders.

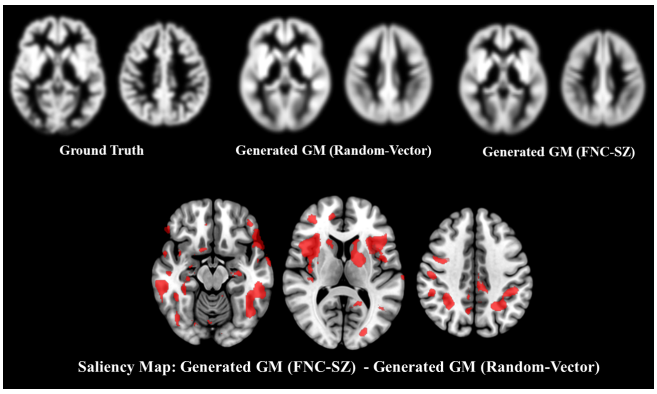


**Fig. 3**. This figure shows GM examples generated from both random vectors and schizophrenia-related FNC data, emphasizing differences between the two GMs. These variations are linked to brain regions strongly associated with schizophrenia and align closely with our previous research on schizophrenia biomarkers, supporting the validity of our findings.

## 5. CONCLUSIONS

Our model, based on a basic model pre-trained on a large dataset, has achieved strong performance in research focused on brain disorders using smaller, disease-specific datasets. In the future, our model will be widely applied to explore and validate biomarkers associated with various brain disorders. These novel or potential biomarkers hold great promise for advancing our understanding of brain disorders and the structural characteristics unique to specific individuals, providing valuable insights into the pathology of neurological diseases.

## 6. REFERENCES


[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[3] Yuda Bi, Anees Abrol, Sihan Jia, Jing Sui, and Vince D. Calhoun, "Gray matters: Vit-gan framework for identifying schizophrenia biomarkers linking structural mri and functional network connectivity," *NeuroImage*, vol. 297, pp. 120674, 2024.

[4] Yan Zhuang, Tejas Sudharshan Mathai, Pritam Mukherjee, and Ronald M Summers, "Segmentation of pelvic structures in t2 mri via mr-to-ct synthesis," *Computerized Medical Imaging and Graphics*, vol. 112, pp. 102335, 2024.

[5] Kazunari Wada, Katsuyuki Suzuki, and Kazuo Yonekura, "Physics-guided training of gan to improve accuracy in airfoil design synthesis," *Computer Methods in Applied Mechanics and Engineering*, vol. 421, pp. 116746, 2024.

[6] André Ferreira, Jianning Li, Kelsey L Pomykala, Jens Kleesiek, Victor Alves, and Jan Egger, "Gan-based generation of realistic 3d volumetric data: A systematic review and taxonomy," *Medical image analysis*, p. 103100, 2024.

[7] 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, pp. 10684–10695.

[8] Juanhua Zhang, Ruodan Yan, Alessandro Perelli, Xi Chen, and Chao Li, "Phy-diff: Physics-guided hourglass diffusion model for diffusion mri synthesis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 345–355.

[9] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso, "Brain imaging generation with latent diffusion models," in *MICCAI Workshop on Deep Generative Models*. Springer, 2022, pp. 117–126.

[10] Yuheng Fan, Hanxi Liao, Shiqi Huang, Yimin Luo, Huazhu Fu, and Haikun Qi, "A survey of emerging applications of diffusion probabilistic models in mri," *Meta-Radiology*, p. 100082, 2024.

[11] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li, "A survey on generative diffusion models," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[12] Alexey Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] Lan Jiang, Ye Mao, Xiangfeng Wang, Xi Chen, and Chao Li, "Cola-diff: Conditional latent diffusion model for multi-modal mri synthesis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 398–408.

[14] Jonghun Kim and Hyunjin Park, "Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7604–7613.

[15] Shaoyan Pan, Elham Abouei, Jacob Wynne, Chih-Wei Chang, Tonghe Wang, Richard LJ Qiu, Yuheng Li, Junbo Peng, Justin Roper, Pretesh Patel, et al., "Synthetic ct generation from mri using 3d transformer-based denoising diffusion model," *Medical Physics*, vol. 51, no. 4, pp. 2538–2548, 2024.

[16] Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu, "Vitgan: Training gans with vision transformers," *arXiv preprint arXiv:2107.04589*, 2021.