DALL-E Brain: Generating 2D T1-w, T2-w, and FLAIR Brain MRI Images from Textual Prompts

Souhail El-Allaly
INSA Lyon
CREATIS UMR 5220, U1294
Villeurbanne, France
souhailallaly3@gmail.com

Thomas Grenier
INSA Lyon,
CREATIS UMR 5220, U1294
Villeurbanne, France
thomas.grenier@insa-lyon.fr

Chantal Revol-Muller
INSA Lyon,
CREATIS UMR 5220, U1294
Villeurbanne, France
chantal.muller@insa-lyon.fr

***Abstract*—Medical imaging studies leveraging deep learning often face data scarcity. This project explores the application of Latent Diffusion Models (LDMs) for generating synthetic yet realistic T1-weighted, T2-weighted, and FLAIR MRI brain images from textual prompts. To train our model, we constructed a dataset of over 40,000 MRI scans of healthy subjects, each paired with slice-specific textual descriptions. This dataset was derived from four publicly available sources: Kirby, IXI, OASIS, and IBSR. The IBSR dataset further enables the generation of brain structure segmentations, which are used to automatically create image-specific textual annotations. These annotations include legend descriptions of anatomical structures, demographic metadata such as age and sex, and imaging modality details specifying the scan plane and MRI sequence. To enhance textual conditioning, we modified an LDM to handle long prompts with a vocabulary specialized for the medical domain. Model performance is evaluated using MS-SSIM for the VAE component and FID for the diffusion-generated images. Experimental results demonstrate that using only textual descriptions, our method can generate realistic MRI scans, highlighting the potential of LDMs for medical imaging synthesis. The LDM code is available at https://gitlab.in2p3.fr/chantal.muller/dalle-brain.**

***Index Terms*—Latent Diffusion Models, Text-Guided Image Generation, MRI Slice Generation.**

I. Introduction

Recent advances in vision-language models, such as CLIP [1], have enabled unified text-image representations, leading to powerful text-to-image generative models like DALL-E and Latent Diffusion Models (LDMs) [2]. While these models excel at synthesizing realistic images, their application to medical imaging remains limited due to the need for anatomical precision and expert validation.

Several approaches have explored text-conditioned medical image synthesis, including transformer-based models [3], vision-language conditioning [4], and contrastive learning [5]. However, these methods often rely on limited paired text-image datasets, making text-driven medical image generation particularly challenging. Here, we propose a text-conditioned LDM trained from scratch for brain MRI synthesis to address this. By employing contrastive learning and transformer-based text encoders, the model can effectively capture semantic information, enabling the generation of realistic and anatomically coherent MRI slices.

Our approach generates 2D brain MRI slices paired with anatomical descriptions to ensure alignment with clinical structures. Section II describes the 2D image-text database, including dataset composition and textual annotation. Section III presents the image encoder-decoder, detailing the VAE-based compression of MRI slices. Section IV introduces the Text-Conditioned Latent Diffusion Model (LDM), covering its architecture, training, and text-conditioning modifications. Section V analyzes qualitative and quantitative results, evaluating reconstruction performance and the role of textual guidance. Finally, we discuss key findings and future research directions.

II. Building the 2D Image-Text Database

The development of a generative AI model requires a large and diverse training dataset. In this study, we first need brain MRI slices of healthy subjects paired with their anatomical descriptions. Publicly available MRI datasets with textual annotations are scarce.

However, such descriptions can be automatically generated if brain structure segmentation is performed beforehand. Labeled MRI images provide access to anatomical structures, and with voxel size information, their volumes can be computed. Only a few datasets of healthy subjects exist, and even fewer include brain segmentations.

To construct a comprehensive dataset, we start with the IBSR dataset [6], which consists of 18 T1-weighted MRI scans with corresponding brain segmentation maps. We further expand this dataset with unsegmented MRI scans from multiple sources covering different sequences: OASIS [7] (77 T1-w. MRI scans), IXI [8] (64 T2-w. MRI scans), and Kirby [9] (42 FLAIR MRI scans).

A. Generation of labeled MRI Images

To increase the number of available labeled MRI images while preserving anatomical consistency, we adopted a simple yet effective segmentation approach. While tools like SynthSeg and FreeSurfer are available, these offer fine-grained segmentation that exceeds our current needs, both in complexity and computational demand. Our objective is limited to identifying major brain structures corresponding to the IBSR labeling scheme, for which a lightweight method based on 3D affine registration proves sufficient and considerably faster.

1) Registration and Segmentation Process

We performed affine registration of the unlabeled MRI scans (OASIS, IXI, Kirby) onto the fixed IBSR labeled 3D images, using the ANTs registration framework [10]. Before registration, skull stripping was applied to ensure accurate multimodal alignment. The same affine transformation was applied to both the images and their corresponding segmentations. This process generated 18 × 77 new labeled T1-weighted MRI, 18 × 64 labeled T2-weighted MRI, and 18 × 42 labeled FLAIR MRI. The inverse transformation was also applied, registering IBSR onto OASIS, IXI, and Kirby, further increasing the dataset size (see Fig. 1).

2) Atlas Generation

To enhance anatomical consistency, we created three new atlases using a majority voting scheme among the 18 registered IBSR labels. This resulted in OASIS Atlas (77 labeled T1-w. MRI), IXI Atlas (64 labeled T2-w. MRI) and Kirby Atlas (42 labeled FLAIR MRI). By iteratively changing the reference fixed atlas and the moving databases, we further increased the number of labeled images in our dataset (see Fig. 2). In total, this procedure resulted in a dataset of 40,602 labeled 3D MRI scans, corresponding to approximately 25.6 million axial, sagittal, and coronal slices.

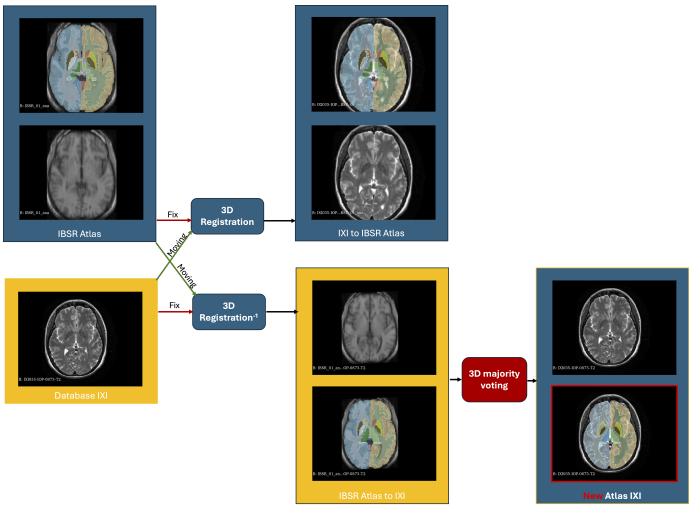


Figure 1: Creation of labeled MRI images from IBSR atlas and generation of new atlases.

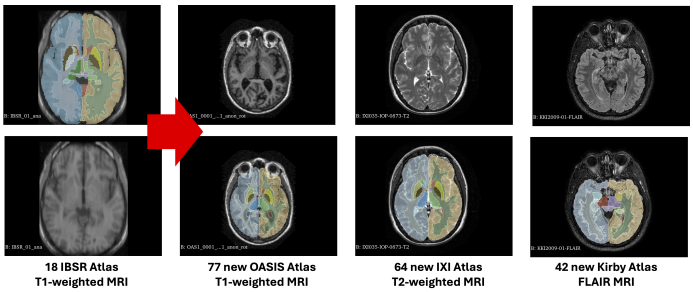


Figure 2: Generated atlases from IBSR registrations.

B. Generation of Textual Descriptions

The IBSR atlas is associated with a lookup table mapping integer labels in the segmented image to the corresponding anatomical structure names. For each registered MRI volume, metadata such as voxel size, MRI sequence type, subject gender, and age were available for both the moving and fixed images. Using these segmentations, we automatically generated structured descriptions for each MRI slice by exporting anatomical structures and their volumes (in mm^3) into CSV format. Finally, natural language descriptions were generated using template-based text synthesis. To introduce prompt variability, 20 sentence templates were randomly selected to describe each slice (see Fig. 3).

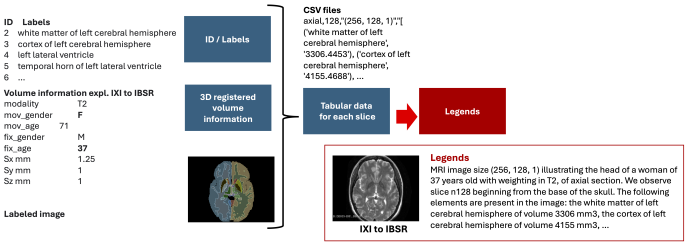


Figure 3: Example of a generated slice-text description pair.

C. Resulting Database

The final dataset generated through our processing pipeline comprises 40,600 labeled MRI scans from healthy subjects, covering multiple MRI modalities. Each MRI volume is paired with anatomical text descriptions, ensuring that each slice retains meaningful structural annotations. This results in an extensive dataset containing approximately 26 million text-image pairs, providing a diverse and well-annotated resource for training generative AI models in medical imaging.

III. Image Encoder/Decoder

A. Variational AutoEncoder (VAE)

A VAE is an encoder-decoder architecture designed to learn a compact latent representation of input data. The encoder maps an image into a lower-dimensional latent space by estimating a probability distribution with a mean vector μ and a standard deviation vector σ. The decoder reconstructs an image from a latent sample, ensuring that the representation captures essential data characteristics [11] (see Fig. 4).

In our generative model, the VAE acts as a compression mechanism, encoding MRI slices into latent space for the diffusion process (see Fig. 6). After denoising, the VAE decoder reconstructs the image back to the pixel domain.

B. Loss

Given an input image $\mathbf{x} \in \mathbb{R}^{H\times W\times 1}$, the encoder maps it into a latent representation with reduced spatial dimensions $\mathbf{z} \in \mathbb{R}^{h\times w\times c}$, $\mathbf{z} = E(\mathbf{x})$. The decoder then reconstructs the image from its latent representation, $\tilde{\mathbf{x}} = D(\mathbf{z}) = D(E(\mathbf{x}))$. The VAE training is driven by the loss function (Eq. 1):

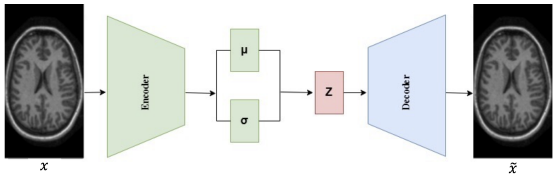


Figure 4: Variational Autoencoder.

$$\mathcal{L}_{VAE} = \min_{E,D} \max_{\psi} \Big(\mathcal{L}_{\text{rec}} \left(\mathbf{x}, D\left(E\left(\mathbf{x}\right)\right)\right) - \mathcal{L}_{\text{adv}} \left(D\left(E\left(\mathbf{x}\right)\right)\right) + \log D_{\psi}\left(\mathbf{x}\right) + \mathcal{L}_{\text{reg}} \left(\mathbf{x}; E, D\right) \Big) \quad (1)$$

where

- $\mathcal{L}_{\text{rec}}$: perceptual reconstruction loss measuring difference between original and reconstructed images;
- $-\mathcal{L}_{\text{adv}} + \log D_{\psi}(\mathbf{x})$: adversarial loss discouraging trivial solutions;
- $\mathcal{L}_{\text{reg}}$ regularization loss minimizing KL divergence between latent space and a standard Gaussian distribution.

C. Training Setup

The VAE was trained using 40,000 axial MRI slice, divided into training set with 24,000 images, validation set with 6,000 images and test set with 10,000 images.

Hyper parameters: epochs=100; batch size=26; GPU=32GB; image size= $(256, 256)$; latent dimension= $(64, 64, 1)$ and total training time= $\sim$ 40 hours.

D. Evaluation of VAE

This section evaluates the VAE's performance in encoding and decoding MRI slices, ensuring that reconstructed images retain key anatomical features. We assess the model both qualitatively and quantitatively by comparing the reconstructed images $\tilde{\mathbf{x}} = D(E(\mathbf{x}))$ with the original inputs $\mathbf{x}$.

1) Qualitative results

Fig. 5 presents a visual comparison between the original MRI slices (top row) and their VAE-reconstructed versions (bottom row). The high visual similarity confirms that the VAE successfully preserves essential anatomical structures.

2) Quantitative results

To objectively assess reconstruction fidelity, we use the Multi-Scale Structural Similarity Index Measure (MS-SSIM) [12], which evaluates image similarity across multiple resolutions, incorporating luminance, contrast, and structural components.

$$\text{MS-SSIM}(x, y) = [l_M(x, y)]^{\alpha_M} \cdot \prod_{j=1}^{M} [C_j(x, y)]^{\beta_j} \cdot [S_j(x, y)]^{\gamma_j} \quad (2)$$

where

- x, y : two images being compared.

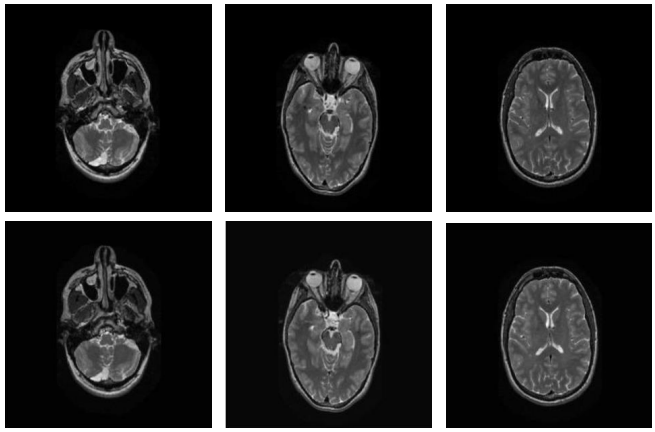


Figure 5: Comparison between original MRI slices (top) and their VAE reconstructions (bottom).

- M : number of scales used in the multi-scale SSIM computation.
- $l_M(x, y)$: luminance comparison function at the coarsest scale.
- $C_j(x, y)$: contrast comparison function at scale j.
- $S_j(x, y)$: structural similarity function at scale j.
- $\alpha_M, \beta_j, \gamma_j$: exponents controlling the relative importance of luminance, contrast, and structure at different scales.

Table I presents MS-SSIM scores for 10,000 randomly selected MRI slices from both the training and test datasets. The high similarity scores (~ 0.98) confirm that the VAE accurately reconstructs input slices while preserving anatomical details.

Table I: MS-SSIM scores for VAE reconstruction.

Dataset	Train (10,000 slices)	Test (10,000 slices)
MS-SSIM [0-1] ↑	0.984	0.977

IV. Text-Conditioned Latent Diffusion Model

We build upon the MONAI 2D Latent Diffusion Model (LDM) tutorial, which does not incorporate text conditioning [13], [14]. This baseline model focuses solely on image-based generation. To extend its capabilities, we modify the original code by integrating text conditioning.

A. Architecture

LDMs operate in a compressed latent space rather than processing high-dimensional pixel data. Our architecture (Fig. 6) includes: (i) a VAE encoding MRI slices into a latent representation, (ii) a text encoder converting descriptions into embeddings, and (iii) a Diffusion Model refining noisy representations, which the VAE decodes into MRI images.

B. Text Embedding Extraction with a tokenizer

To generate text-conditioned MRIs, we use a WordPiece tokenizer and a transformer pre-trained on biomedical literature. The input text T_{orig} is tokenized into contextualized

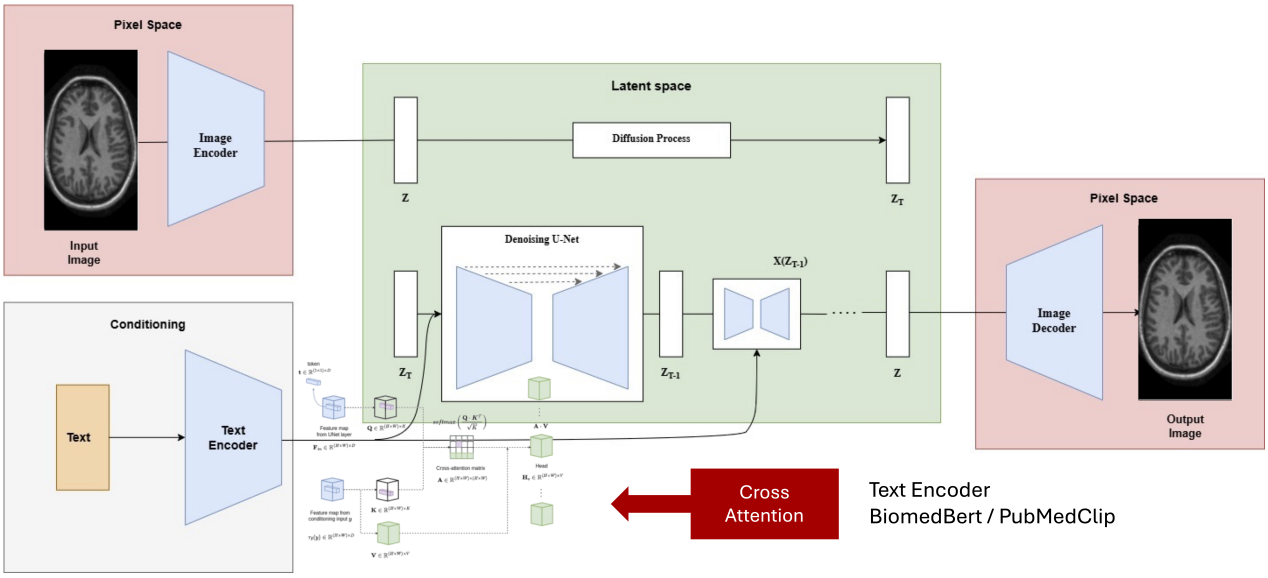


Figure 6: LDM architecture conditioned by textual cross-attention.

embeddings, which condition the diffusion model, whether from a contrastively trained model (PubMedCLIP [15]) or a language model (BiomedBERT [16]).

$$\text{Embeddings} = \text{tokenizer}(T_{\text{orig}}) \in \mathbb{R}^{T \times d_t} \tag{3}$$

where:

- T : number of tokens.
- $d_t = 768$: text embedding dimension.

Here, our batch of embeddings $\in \mathbb{R}^{(B,T,768)}$ represents per-token contextual embeddings (with B the batch size).

C. Cross-Attention Mechanism

To align textual and visual representations, we integrate cross-attention into the LDM. Queries $Q \in \mathbb{R}^{(B,N_{\text{spatial}},C_{\text{feat}})}$ come from U-Net latent features ($N_{\text{spatial}} = h' \times w'$, $C_{\text{feat}} = 768$), while keys and values $K, V \in \mathbb{R}^{(B,T,768)}$ originate from tokenizer embeddings. The attention mechanism follows the query-key-value formulation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_t}}\right) V. \tag{4}$$

This mechanism injects semantic content from the text into the latent image features, guiding the denoising process toward text-consistent representations.

The U-Net is configured to apply the cross-attention mechanism across its last two levels.

D. Comparison of Text Encoders

We evaluate two text encoders for this purpose: PubMedCLIP and BiomedBERT. PubMedCLIP is constrained to 77 tokens, which limits its capacity for handling long-form anatomical descriptions. In contrast, BiomedBERT can process up to 512 tokens, providing a richer and more detailed text representation. Furthermore, BiomedBERT achieves comparable performance with half the training epochs required by PubMedCLIP. Based on these advantages, we select BiomedBERT as our primary text encoder.

E. Loss Function

The LDM is trained by minimizing the following objective function (Eq. 5):

$$\mathcal{L}_{LDM} = \mathbb{E}_{\epsilon(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right] \tag{5}$$

where ϵ represents the added noise and $\epsilon_\theta(z_t, t, \tau_\theta(y))$ the noise predicted by the U-Net at time step t.

F. Training Setup

The LDM was trained on 40,000 axial MRI slices, with the following dataset split: 24,000 images for training, 6,000 images for validation and 10,000 images for test.

Hyper parameters: Pre-trained VAE + LDM conditioned using BiomedBERT. 100 epochs, batch size 40, GPU 32GB, image size: (256, 256), latent dimension: (64, 64, 1), training time: 14d 22h 9m (∼358 hours).

V. Results and Discussion

A. Evaluation Metric

The generative quality is assessed using the Fréchet Inception Distance (FID) (Eq. 6), which quantifies the feature distribution distance between generated and real MRI slices in an InceptionV3-based embedding space [17].

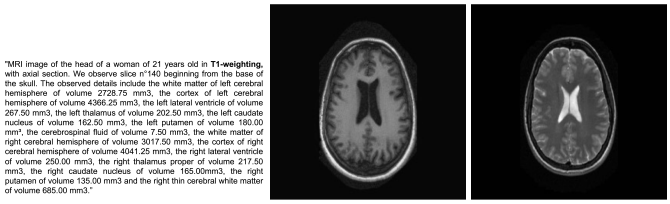
$$\text{FID} = \|\mu_x - \mu_y\|_2^2 + \text{Tr}\left(\Sigma_x + \Sigma_y - 2\left(\Sigma_x\Sigma_y\right)^{\frac{1}{2}}\right) \tag{6}$$

where:

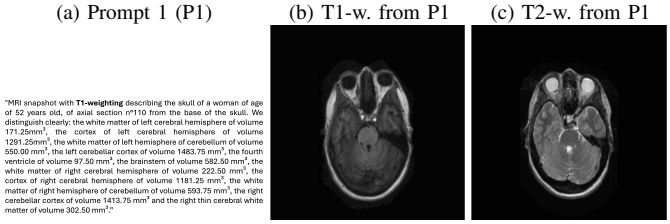
- μ_x and μ_y : mean vectors of the feature distributions for generated and real images, respectively.
- Σ_x and Σ_y : covariance matrices of the feature distributions for generated and real images.

B. Generated Image Evaluation

We evaluate 1,000 synthetic MRI slices generated from training/validation dataset descriptions and 1,000 synthetic slices from test descriptions against 10,000 real slices from the respective datasets. Fig. 7 illustrates 2D axial slices generated using our text-conditioned LDM. The first row corresponds to prompt 1, and the second row to prompt 2. Based on input prompts, the model accurately distinguishes MRI sequence types (T1-w vs. T2-w).



(a) Prompt 1 (P1) (b) T1-w. from P1 (c) T2-w. from P1



(d) Prompt 2 (P2) (e) T1-w. from P2 (f) T2-w. from P2

Figure 7: T1-w. and T2-w. MRI slices generated by prompt-conditioned LDM.

C. Quantitative Results

Table II presents FID scores comparing 1,000 synthetic MRI slices to 10,000 real slices. The model achieves low FID scores, indicating high fidelity and structural consistency.

Table II: FID scores for LDM-generated MRI slices.

	Train/Val (1,000 vs 10,000)	Test (1,000 vs 10,000)
FID ↓	**16.9**	**17.9**

The FID scores indicate strong generalization. The first column reports results for prompts seen during training, while the second column evaluates unseen prompts. Despite this, the model maintains low FID scores (16.9 for known prompts, 17.9 for unseen prompts), demonstrating its ability to synthesize realistic MRI slices from novel textual inputs.

D. Discussion

Our results confirm that the text-conditioned LDM effectively synthesizes realistic MRI slices while maintaining anatomical coherence. The low FID scores suggest the model learns meaningful representations, producing images that closely resemble real MRI scans. Qualitative evaluation (Fig. 7) further highlights its capability to generate diverse MRI slices based on textual descriptions. The model accurately captures anatomical structures described in the prompts and differentiates MRI sequence types, validating the effectiveness of the conditioning mechanism. While these results are promising, limitations remain. FID measures statistical similarity but does not fully assess clinical relevance. Future work includes expert radiologist evaluation to validate anatomical and pathological consistency.

VI. Conclusion

We introduced a text-conditioned LDM for synthetic brain MRI generation, leveraging anatomical descriptions for conditioning. By integrating BiomedBERT embeddings and a pre-trained VAE, our approach produces high-fidelity MRI slices, validated through qualitative and quantitative evaluations. Results show that text-driven diffusion models generate realistic, anatomically coherent medical images, addressing data scarcity in medical imaging. Clinical validation remains crucial to assess diagnostic utility. Future work includes expert radiologist evaluation for anatomical accuracy, extending the model to 3D MRI generation, and integrating pathological conditions to simulate disease progression.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[3] A. G. Barreto, J. M. de Oliveira, F. N. B. Gois, P. C. Cortez, and V. H. C. de Albuquerque, "A new generative model for textual descriptions of medical images using transformers enhanced with convolutional neural networks," *Bioengineering*, vol. 10, no. 9, p. 1098, 2023.

[4] X. Xing, J. Ning, Y. Nan, and G. Yang, "Deep generative models unveil patterns in medical images through vision-language conditioning," *arXiv preprint arXiv:2410.13823*, 2024.

[5] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text," *arXiv preprint arXiv:2210.10163*, 2022.

[6] "Nitrc ibsr," https://www.nitrc.org/projects/ibsr.

[7] "Oasis," https://www.oasis-brains.org/data.

[8] "Ixi dataset," https://brain-development.org/ixi-dataset.

[9] "Nitrc kirby21," https://www.nitrc.org/frs/?group_id=313.

[10] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, "A reproducible evaluation of ants similarity metric performance in brain image registration," *NeuroImage*, vol. 54, no. 3, pp. 2033–2044, 2011.

[11] D. P. Kingma, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[12] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.

[13] "Tutorial 2d ldm," https://github.com/Project-MONAI/tutorials/tree/main/generation/2d_ldm.

[14] W. H. Pinaya, M. S. Graham, E. Kerfoot, P.-D. Tudosiu, J. Dafflon, V. Fernandez, P. Sanchez, J. Wolleb, P. F. Da Costa, A. Patel *et al.*, "Generative ai for medical imaging: extending the monai framework," *arXiv preprint arXiv:2307.15208*, 2023.

[15] "Pubmedclip," https://huggingface.co/flaviagiammarino/pubmed-clip-vit-base-patch32.

[16] "Biomedbert," https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext.

[17] S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, and S. Kumar, "Rethinking fid: Towards a better evaluation metric for image generation," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 9307–9315.