Controllable Text-to-Image Synthesis for Multi-Modality MR Images

Kyuri Kim[1] Yoonho Na[1] Sung-Joon Ye[1] Jimin Lee[2]
Sung Soo Ahn[3] Ji Eun Park[4] Hwiyoung Kim[3]
[1]Seoul National University [2]Ulsan National Institute of Science & Technology
[3]Yonsei University College of Medicine [4]University of Ulsan College of Medicine
{kyurikim, yoonho94.na, sye}@snu.ac.kr jiminlee@unist.ac.kr
{sungsoo, hykim82}@yuhs.ac jieunp@gmail.com

Abstract

Generative modeling has seen significant advancements in recent years, especially in the realm of text-to-image synthesis. Despite this progress, the medical field has yet to fully leverage the capabilities of large-scale foundational models for synthetic data generation. This paper introduces a framework for text-conditional magnetic resonance (MR) imaging generation, addressing the complexities associated with multi-modality considerations. The framework comprises a pre-trained large language model, a diffusion-based prompt-conditional image generation architecture, and an additional denoising network for input structural binary masks. Experimental results demonstrate that the proposed framework is capable of generating realistic, high-resolution, and high-fidelity multi-modal MR images that align with medical language text prompts. Further, the study interprets the cross-attention maps of the generated results based on text-conditional statements. The contributions of this research lay a robust foundation for future studies in text-conditional medical image generation and hold significant promise for accelerating advancements in medical imaging research.

1. Introduction

In recent years, generative modeling has made rapid progress in the field of image synthesis. The latest method based on diffusion models has solved the problems prevalent in generative adversarial networks (GANs) [3], such as unstable learning, mode collapse, and gradient vanishing, enabling the generation of more realistic images [23]. These developments prompted various derived models such as denoising diffusion probabilistic models and score-based models. Models learn the process of gradually transforming data into noise and gradually removing this noise to return to the original data. This incremental process yields a training dynamic where, in contrast to GANs, the loss function tends to offer more stable gradients. In particular, diffusion models are easy to accommodate conditional generation during the training process, allowing users to have more detailed control over the image to be created. This control ability can guide the model to generate images with specific properties.

Practical methodologies show that controllable image generation techniques empowers the production of images that align with the intended tasks. The image manipulation method of latent space through inversion, which was first utilized in GAN-based models, has been extended and applied to diffusion models [24]. Recent methodologies involve manipulating local information in the image by extracting or transforming specific information based on words to serve as additional in-context guidance for the model or injecting it into a transformer layer [7]. In particular, the cross-attention mechanism of the transformer reports that multi head attention can associate a word with a specific region of the generated image, reporting the scalability of interpretability in large-scale text-to-image generation models. The controllability of these models establish the diffusion model as a large-scale basic generative model and form the basis for reporting its potential for extended interpretability in large-scale text-image generative models [19, 21, 22].

Along with these developments, the diversity of underlying models for data modalities has led to the emergence of multi-modal learning methods. The foundational model allows to composite images using a variety of data types, including text, video, audio, depth, and thermal. In medical imaging research, the use of diagnostic data containing diverse clinical information represents a promising direction, especially for text-based medical image generation [35, 37]. Pairing clinical diagnostic reports with medical images is especially helpful in maximizing synergy between modalities [14]. Clinical diagnostic reports facilitate the effective explanation of medical images. Consequently, the exploita-

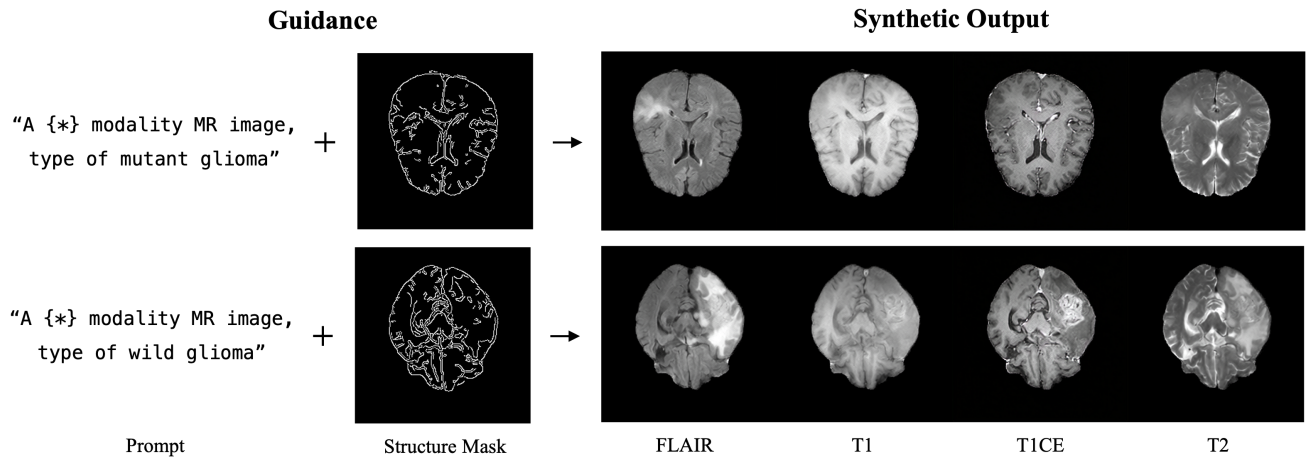

Figure 1. Application results of the framework. Suggested framework allows to synthesis multi-modal medical images control via prompt or structure mask guidance in diffusion models.

tion of multi-modal data, such as the integration of medical images and diagnostic reports, has attracted considerable interest in the area of text-to-image generation [18,25]. Nonetheless, research on generating medical imagery, such as computer tomography (CT) or magnetic resonance imaging (MRI) scans from medical prompts is scant. Challenges arise from the scarcity of matched diagnostic reports and medical images, which are essential for training. Moreover, the complexity of published reports, often containing a mixture of diagnostic information incomprehensible to general public, complicates their use as training data. Additionally, processing this data requires significant computing resources.

Building on these ideas, the present study introduces a framework for generating MR images using text prompts and structural masks as conditions. This framework comprises a pre-trained large language model, a diffusion-based prompt-conditional image generation architecture, and an additional induced noise removal network for input structural binary masks. Experimental results demonstrate the framework's ability to generate realistic, high-resolution, and high-fidelity multi-modal MR images that align with medical language text prompts. Additionally, the study aims to interpret the generated results through pixel-level cross-attention maps for attribution. This approach separates medical imaging data into structure and style, transforming relatively simple structural information into a guiding structural mask while enabling the inclusion of detailed texture information through text. The main contributions of the study can be summarized as follows:

- A controllable text-to-MR image framework is introduced, capable of generating medical images adapted to diverse conditions.
- A strategy based on medical prompts and structural mask for generating multi-modal MR images, confirming the ability to produce sequences of different modalities.
- Analysis of activation areas through the visualization of attention maps corresponding to text-conditional statements in the synthesized results.

This study purpose the generation of a synthetic dataset, specifically focusing on brain tumors, by leveraging medical images and corresponding clinical information. The aim is to facilitate learning in the realm of text-conditional medical image generation—a field that has not yet seen extensive research. Through this approach, the study establishes a critical foundation for future research into text-driven medical image generation.

2. Related Work

Diffusion-based Generative Models. The diffusion probabilistic model, initially unveiled in [26], marked a transformative moment in the field of image generation techniques. Since its introduction, the architecture has undergone significant enhancements, notably through pioneering training and sampling methodologies like the denoising diffusion probabilistic model (DDPM) [9], the denoising diffusion implicit model (DDIM) [27], and score-based diffusion [28]. The U-net architecture serves as the foundational neural network for these methods [5]. However, the evaluation and optimization of these models in pixel space come with inherent challenges, such as slow inference speeds and elevated computational costs. To address these issues, the latent diffusion model (LDM) [22] was proposed, inspired by the concept of latent images [6]. This

methodology was later extended to include Stable Diffusion, and ongoing research is delving into advanced sampling strategies [17] as well as hierarchical approaches [10].

To augment the control over the image synthesis process, a range of mechanisms have been introduced. Contemporary image diffusion models frequently incorporate additional conditions, often employing text-to-image methods [2]. These methods typically encode text inputs into latent vectors using pretrained language models like contrastive language-image pre-Training (CLIP) [20]. Alternative conditioning methods, such as edge masks, semantic maps, and depth maps, are also in use [38]. Concurrently, research is being conducted on the exploration and editing of diffusion models, including techniques for model inversion [24] and attention map manipulation [7]. These methods offer users an intuitive editing experience by allowing them to alter both local and global image details through simple text prompt modifications.

Multimodal training in Medical artificial intelligence(AI).

As the need for medical artificial intelligence continues to grow, acquiring high-quality medical data has become an important area of research. This includes not only traditional diagnostic imaging, but also various data types such as audio recordings, signal data, and text reports [1, 11, 34, 39]. The emergence of high-performance, large-scale text-based models has sparked interest in leveraging diagnostic reports, especially for machine learning applications. Publicly available clinical report datasets, such as MIMIC [12], often serve as the foundation for this research and are incorporated into studies of genomics [29] or multi-view imaging [40]. These various forms of data are also increasingly being used to generate synthetic data [32, 36].

However, there are challenges associated with developing models that can generate medical images from text. These challenges include extensive data sets encompassing imaging data and corresponding radiological reports, specialized skill sets required for accurate image annotation, and strict regulations regarding data sharing. Additionally, with publicly available datasets such as MIMIC, most studies are often limited to chest X-rays (CXRs) [11, 30, 34, 40]. In particular, reports within these datasets are standardized with repeated use of specific terminology and often use ambiguous language that can apply to both normal and abnormal findings, making it difficult to capture the essence of the disease. To tackle these challenges, our approach involves constructing a paired dataset of images and prompts, selectively incorporating essential information specific to brain MR images. We then synthesize multi-modal sequences that are crucial for diagnosis. Our method employs multiple guidance mechanisms for the synthesis process and optimizes the model without the need for constructing parallel layers.

3. Methods

The proposed framework is designed to execute image generation based on text prompts while incorporating structural instructions to ensure fidelity to given conditions and realism of the generated output. The subsequent sections provide a detailed analysis of this framework. The discussion begins with an introduction to the diffusion models in Section 3.1. This is followed by an exploration of how structural guidance is applied to multi-modal MR image generation in Section 3.2. Section 3.3 provides an overview of the overall framework, and Section 3.4 describes the interpretation of the guidance conditions using cross-attention.

3.1. Preliminaries

In a diffusion model, two distinct processes are at play: (1) a forward process that incrementally introduces minor Gaussian noise to the sample over a series of T steps, and (2) a complementary backward process equipped with learnable parameters designed to restore the original input images by identifying and removing the added noise. This study employs the LDM as its foundational framework to demonstrate the capability of generating multi-modal medical images with controllable features. The LDM integrates a U-Net architecture for its denoising component, which is structured into three main parts: an encoder, a middle block, and a decoder. Each of these main parts contains tweleve corresponding blocks. The implementation of skip connections enables the decoder to directly leverage features from the encoder, thereby reducing information loss.

In the LDM, the cross-attention layers perform a dual role. These layers are essential for capturing the semantic information of the input text descriptions, and play a pivotal role in aligning the visual content with the textual context at the stage where noise are predicted. This process is facilitated by the cross-attention layers, which integrates the embeddings of visual and textual information corresponding to each text token. Formally, let $\phi(z_t)$ represent the incoming noise features and $\psi(p)$ denote the text token embeddings generated by the language encoder. The query (Q), key (K), and value (V) in the cross-attention mechanism can be formulated as follows:

$$\begin{aligned} Q &= W_q(\phi(z_t)), \\ K &= W_k(\psi(p)), \\ V &= W_v(\psi(p)), \end{aligned} \tag{1}$$

where W_q, W_k, and W_v are linear projection matrices.

3.2. Adding Structural Guidance for Multi-Modal generation

To incorporate structural guidance into the model, auxiliary neural networks are employed, with the weights for

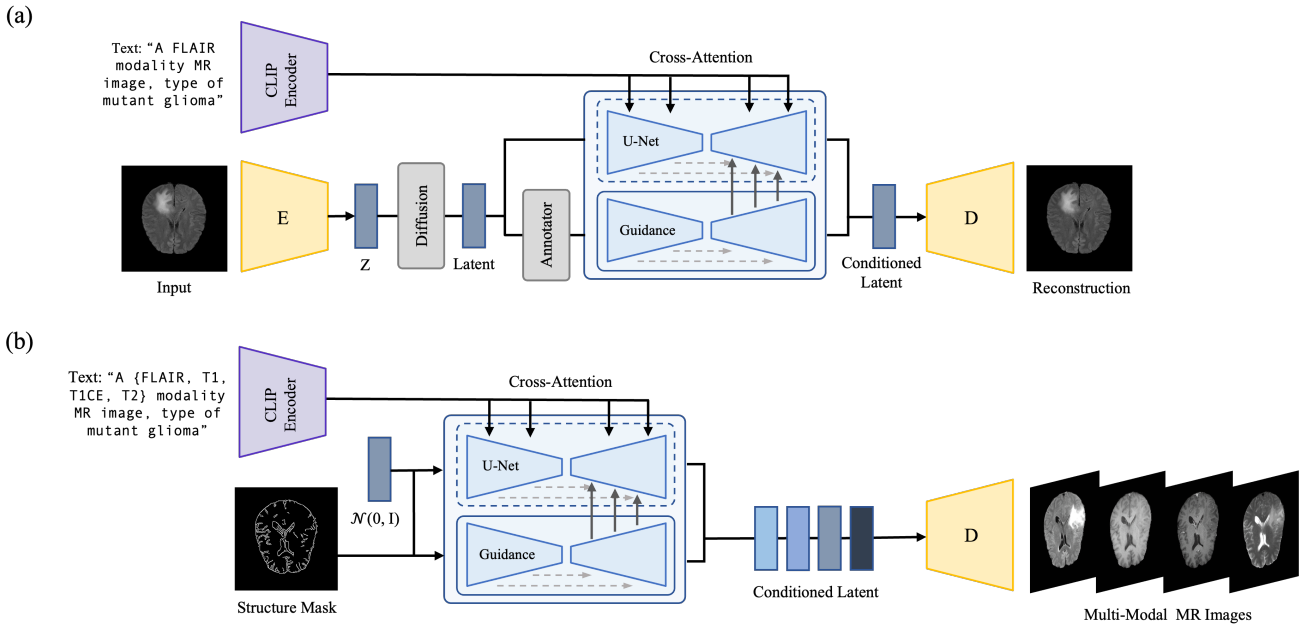

Figure 2. Overall pipeline of the suggested framework. (a) Model training involves inputs such as a medical image, a text prompt, and a structural mask provided by an annotator. These elements are integrated in the reverse diffusion process of the U-Net for guidance. (b) For multi-modal image synthesis, the model is capable of generating images based on the structural mask and text corresponding for each MR modality.

the encoder and middle block copied from the LDM. This approach allows for the integration of structural guidance information during the decoding phase, thereby influencing the overall behavior of the neural network. In practical terms, the guidance network clones the parameters θ from the primary U-Net stream, resulting in a copied set θ_c. These cloned parameters can then be updated with a structural mask condition c_m. The encoders and middle layers of the replicated guidance networks are activated, allowing for model parameter updates. Features extracted from the convolution blocks are then channeled into the decoder, which includes a zero convolution $z(\cdot,\cdot)$ and a 1x1 convolution layer initialized to zero. The input feature map I is updated via the zero convolution as follows:

$$z(I,(W,B)) = B + \sum I \cdot W, \tag{2}$$

where W and B represent the weight and bias, respectively.

Among the various input conditions that the model receives, the convolution layer—predominantly influenced by local information and structural patterns—is particularly sensitive to the input binary mask. To prioritize global context learning, a zero convolution is inserted before and after θ_c. Gradient updates are selectively applied only to the first middle block of θ, rather than to every block in the decoder. To focus on global context learning, a zero convolution is inserted before and after θ_c, with gradient updates applied only to the first middle block of θ, rather than every block of the decoder. The structure can be formulated as:

$$y_{\text{cond}} = f(x,\theta) + z(f(x, z(x,\theta_{z1}),\theta_c),\theta_{z2}), \tag{3}$$

where y_{cond} returns the conditioned latent variable, serving as the output of the denoising network.

3.3. Overall Framework

By enhancing the capabilities of the LDM, the proposed framework is designed to facilitate the synthesis of multi-modal medical images. Through the use of user-friendly prompts, it is capable of generating images that capture the specific modality and attributes required, while maintaining a uniform organ structure. The framework integrates CLIP for a pre-trained large language model along with a diffusion-based, prompt-conditional image generation architecture and a auto encoder (AE) [15]. Additionally, the diffusion architecture incorporates a guidance denoising network specifically for input structure binary masks.

In practical terms, the model is trained to progressively denoise images for synthesis within the perceptual latent space. The final objective L for the entire model is defined as:

$$L = \mathbb{E}_{z_0,t,c_t,c_m,\epsilon\sim N(0,1)}\left[\|\epsilon - \epsilon_\theta(z_t,t,c_t,c_m)\|_2^2\right], \tag{4}$$

where t is the time step, z_t is the noisy image, c_t is the text condition, c_m is the structure mask, and ϵ_θ is the training network.

During the sampling process, latent vectors are initialized to random noise and fed a text prompt. These vectors are then iteratively denoised using the U-Net and finally decoded into an image using the decoder of the AE. By fixing the structural input conditions, random noise and the model's initial seed while adjusting the prompt conditions, the model can generate a variety of conditional latent variables. This enables the creation of the sequence of multi-modal images. The overall pipeline is illustrated in Figure 2.

3.4. Interpretation of guidance conditions using cross attention

As previously outlined in Section 3.1, the input prompt undergoes a denoising process facilitated by a cross-attention layer. In this configuration, each text token from the input prompt generates a spatial attention map within the cross-attention layer, serving to fuse visual and textual embeddings. Building upon prior research [7], an in-depth analysis is conducted to explore the relationship between the spatial layout of images and individual words in the prompt. The focus is specifically on the cross-attention layer within the text-conditioned model. A proposal is made to generate two-dimensional attention maps based on text tokens, aiming to identify which attributes should be emphasized during the synthesis process. The attention map M is obtained as follows:

$$M = \textit{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right), \tag{5}$$

where d represents the dimensions of the latent projection. Intuitively, this equation captures the similarity between Q and K.

Attention maps generated from both the U-Net architecture and guidance networks exhibit varying scales across different layers. To address this issue, the attention score arrays are upscaled to a fixed input image size (h, w) through linear interpolation. These arrays are then aggregated across layers, denoising time steps, and heads, as given by:

$$H_k[h, w] = \sum_{i,j,l} M^{up}_{t_j,k,l}[h, w] + M^{down}_{t_j,k,l}[h, w] \tag{6}$$

where k is the k^{th} word from the prompt, i^{th} represents the up and down sampling layers, and l^{th} head. The aggregated feature maps are subsequently normalized and visualized as heat maps H_k, each corresponding to individual words in the prompt.

4. Experiments

In this section, we assess the performance of the proposed approach on generating synthetic multi-modal medical images.

4.1. Dataset

To evaluate the performance of the proposed method, a brain MR imaging dataset of glioma patients was sourced from Seoul Asan Medical Center. This dataset features four distinct imaging modalities—FLAIR, T1, T1Gd, and T2—and comprises a cohort of 484 patients (n=2k). During the preprocessing stage, each MR image slice was normalized to fit a value range between 0 and 1. Additionally, both the MR images and their corresponding annotated tumor masks were resized from an original resolution of 224×224 pixels to a more standardized 256×256-pixel format. Only slices displaying visible tumors were selected for inclusion in the research experiments.

Accompanying each slice in the dataset is a diagnostic report that specifies the presence or absence of isocitrate dehydrogenase (IDH) mutations. These mutations are classified as either mutant or wild type, a categorization confirmed by two certified radiologists. For the text prompts used in the study, a standardized sentence structure was adopted: "A modality MR image, type of glioma." The placeholders in this sentence were filled with specific details related to the MR modality and IDH type for each individual slice.

4.2. Implementation Details

The network underwent training with a batch size of 8 and a learning rate of 1E-6. During the inference phase, DDIM was adopted for the sampling process, employing a sequence of 100 time steps. As delineated in Section 3.3, a binary mask was generated using the Canny Edge detector, where the threshold hyperparameters were configured to a minimum of 100 and a maximum of 200. For the text prompts, guidance was provided via the CLIP model.

4.3. Quantitative Results

Image quality and diversity. The quality of the generated images is evaluated using the frechet inception distance (FID) metric [8]. Table 1 presents quantitative results derived from the synthetic MR dataset. These results show improved performance when employing a combination of guidance techniques, in contrast to the results using alternative approaches such as StyleGAN2 [13] and vanilla LDM. Both models compared were trained from scratch while LDM utilized the same pre-trained CLIP. Figure 3 provides qualitative evidence, illustrating the results generated based on prompt and structural mask instructions in the input. The higher FID scores recorded across variety MR imaging modalities are reflected in the qualitative analysis, demonstrating that the generated images are precisely customized to the specific MR modality and tumor type.

Correspondence to multi-modality slice sequence. Maintaining a correlation among various medical image modalities—such as FLAIR, T1, T1CE, and T2—is crucial

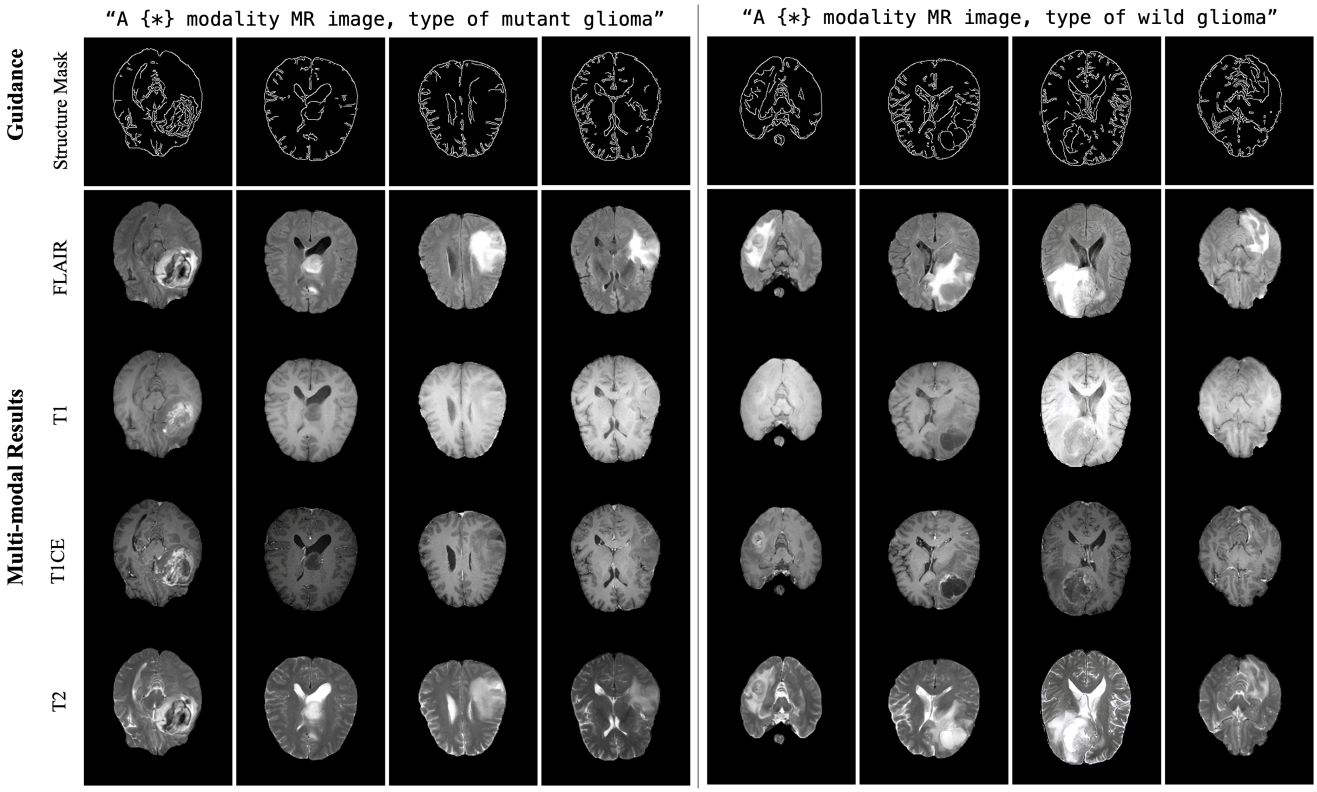

Figure 3. Synthetic results of multi-modal MR images generated based on both input prompt and structural mask guidance. The asterisk {*} in the prompt serves as a placeholder for various MR modality input text, which include FLAIR, T1, T1CE, and T2.

	Modality				
Method	FLAIR	T1	T1CE	T2	Average
StyeGAN2	56.20	75.85	**34.63**	33.03	49.18
LDM	56.86	77.15	61.57	37.27	58.21
Ours	**43.23**	**46.58**	45.47	**30.15**	**41.35**

Table 1. Quantitative comparison of different models using FID. The best results are shown in bold.

	SSIM	NMI	NCC
GT	0.757 ± 0.08	0.487 ± 0.06	0.910 ± 0.02
Prediction	0.743 ± 0.17	0.457 ± 0.06	0.909 ± 0.03
Random	0.659 ± 0.04	0.271 ± 0.04	0.735 ± 0.09

Table 2. Quantitatively comparing the correlations of multimodal sequences.

for the fidelity of the synthesized outcomes. However, tumor features often display different characteristics across these modalities. To address this, multiple metrics have been selected for medical image registration [4], including the structural similarity index measure (SSIM) [33], normalized mutual information (NMI) [16], and normalized cross correlation (NCC). To define an acceptable range for these quantitative metrics, upper and lower bounds were established based on true MR multi-modality sequences. The upper bound was determined using actual sequences to measure the correlation between images across different modalities, while the lower bound was set using randomly sampled images from each MR modality. This approach ensures that the upper bound not only measures inter-modality correlation but also accommodates the unique tumor morphology that may vary between modalities. The computational results for a total of 500 sequences—comprising 250 IDH mutant types and 250 wild types—are presented in Table 2. The findings confirm that all sequences fall within the calculated category and closely approach the upper limit. This suggests that a multi-modal set, conditioned on both modality and tumor type, has been successfully generated.

Correspondence to prompt guidance. The CLIP score [20], defined as the correlation between CLIP text and image embeddings, serves as a metric to evaluate the similarity of synthetic results under given prompt guidance. Specifically, to account for minimum object similarity, the

	FLAIR		T1		T1CE		T2		
Text	Mutant	Wild	Mutant	Wild	Mutant	Wild	Mutant	Wild	Average
Full-Prompt	30.8779	31.0675	31.3073	30.3258	31.6367	31.2587	31.2019	30.9289	31.0755
Modality Type Subset	24.4757	24.4810	28.7178	28.8115	26.1714	26.1978	29.1843	29.3515	27.1738
IDH Type Subset	30.0574	29.9842	30.0087	29.2547	30.0681	30.1175	29.8715	29.8513	29.9016

Table 3. Average image-text similarities measured by CLIP score between the text prompt and the generated images.

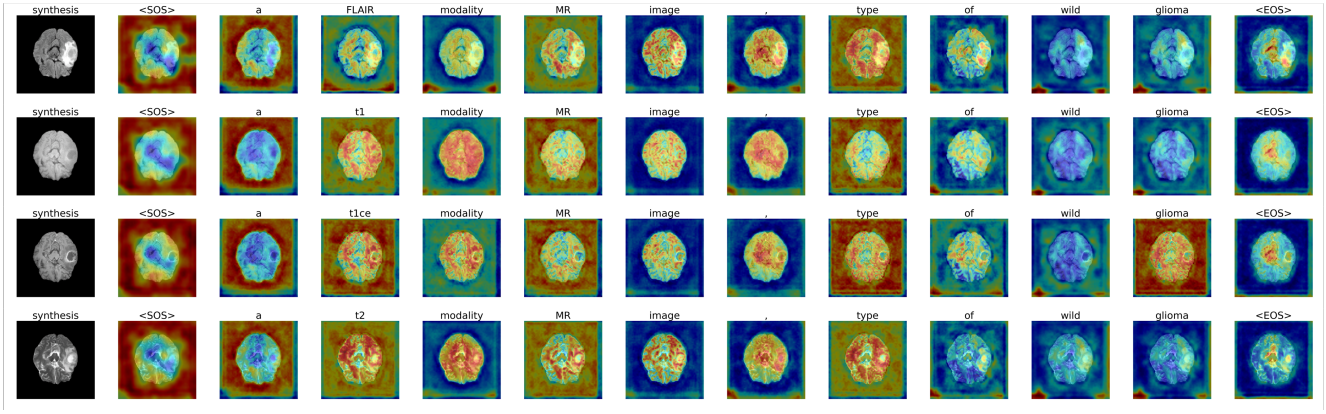

Figure 4. Visualization of the attention map displays text elements corresponding to synthetic outcomes. Along the y-axis, the map displays results for various MR modalities, including FLAIR, T1, T1CE, and T2. Meanwhile, the x-axis arranges the words derived from the input prompt in order.

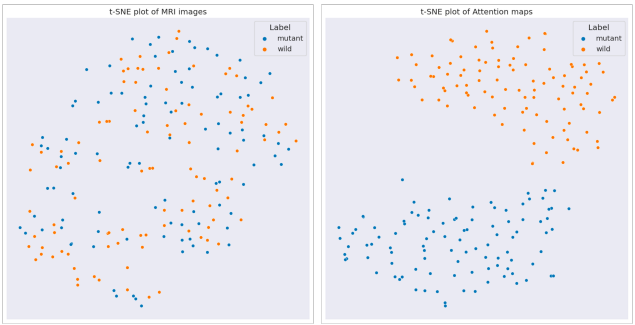

Figure 5. Comparison between the t-SNE dimensional reduction results for MR images (on the left) and the average activation maps categorized by tumor type (on the right).

entire prompt was segmented into two subsets: modality and IDH type, which are key elements of the sentences. The "Full-Prompt" category measures the image-text similarity based on complete sentences, while the subsets for modality type and IDH type employ the phrases "a specified modality type modality MR image" and "type of specified IDH type glioma," respectively. The specific CLIP model utilized for this evaluation is ViT-B/16. Upon reviewing the CLIP scores summarized in Table 3, it was found that the degree of agreement was highest across all prompts. Additionally, the similarity was marginally higher in the IDH type subset compared to the modality type subset.

4.4. Cross-Attention Map Visualization

By incorporating both prompt and structural conditions, the cross-attention maps aim to offer comprehensive control over the image generation process. The aggregated results of these maps are displayed in Figure 4, which highlights text elements corresponding to synthetic outcomes across various imaging modalities. A recurring pattern observed in the attention maps is the global activation of image areas when words like "FLAIR", "modality", "MR", and "image" are used. In contrast, localized areas are emphasized when terms such as "wild", and "glioma" are present. Notably, these maps highlight regions where a tumor's presence can be inferred, even without the use of specific tumor location information like a tumor mask. Such insights open up possibilities for future research, particularly in exploring the correlation between gene displacement and the specific locations of different IDH tumors.

Based on critical observations and visualization results, the cross-attention layer emerges as a key factor in determining the attribution between text and image for each word. This finding suggests the potential for feature separation, particularly in the context of complex and difficult-to-interpret medical images like tumors. Representative tumor

activation maps for each sequence were obtained by averaging the activation maps associated with the words "glioma" or "wild" across the four imaging modalities. Subsequently, dimension reduction was performed using t-SNE [31], focusing on the average attention map associated with the representative IDH type. Initially, both sets of images, with dimensions of 256×256 pixels, were downsized to a feature size of 1000 using ResNet-18, further reduced to a two-dimensional space using t-SNE.

Compared to MR images—which often require expert evaluation, intricate medical information, and frequently exhibit overlapping features—the selectively activated map demonstrates significant separation in the 2D reduction results as in Figure 5. This occurs even when the map dimensions are identical. The distinct results in the reduced dimension not only validate that the synthesis process yielded condition-appropriate outcomes, but also indicate that this is an effective approach for extracting and utilizing desired feature information.

4.5. Qualitative Evaluation

The Turing Test was employed to display synthetic medical imagery for clinical expert evaluation. Two distinct evaluations were conducted: (1) a comparative review where clinicians chose synthetic MR images over a randomly paired real and synthetic image set from the same modality, and (2) a qualitative review in which experts rated a sequence of generated MR images on a scale ranging from 0 to 5. Additionally, the latter evaluation involved determining whether the sequence included a tumor with an IDH mutant-like or wild-like phenotype. Figure 3 in the supplementary material provides an example of the test interface. Two expert clinicians from Yonsei University Severance hospital and Asan Medical Center were participated, evaluating 20 questions within 30 minutes time constraint for each test. The average accuracy rates for the test types—identification of Synthetic in comparison to Real and differentiation between Mutant and Wild-type tumors—were 47.5% and 67.5%, respectively. The image scoring averaged 4.375 ($\pm$ 0.7403) out of a maximum of 5. The qualitative evaluation suggests that the synthesized images are clinically comparable to actual MR images of brain tumors. The quantitative findings confirm that the generated images conformed to the specified IDH type in the input prompt, with the majority consistent with the clinical judgments. The results were affirmed by clinical reviewers to be of high quality, reflecting the advanced level of synthetic output achieved.

5. Conclusion

The framework proposed in this study has exhibited exceptional performance in generating synthetic multi-modal MR sequences, guided by specific input conditions. By leveraging a diffusion-based denoising network, enhanced with an additional guidance layer, the model adeptly integrates both structural masks and text prompts as conditional inputs. This approach to multimodal image generation not only eliminates the need for parallelizing large parametric models for each mode, but also promotes greater synergy by integrating diverse medical information. Furthermore, the interpretation of cross-attention maps, based on text-conditional statements, facilitates the disentanglement of intricate features in the generated images. Future research should address the incorporation of more intricate sentence structures and the utilization of diverse information forms. Consequently, this research provides a solid foundation for future work in the field of text-conditional medical image generation and offers potential to promote advances in the field of medical imaging.

Acknowledgements This work was supported in part by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HI21C1161); in part by an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Ministry of Science and ICT (MSIT), Republic of Korea (No.2020-0-01336, Artificial Intelligence Graduate School Program (UNIST)); and in part by a grant from the National Research Foundation of Korea (NRF) funded by the MSIT, Republic of Korea (NRF-2021R1F1A1057818).

References

[1] Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784, 2022. 3

[2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 3

[3] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018. 1

[4] Steffen Czolbe, Paraskevas Pegios, Oswin Krause, and Aasa Feragen. Semantic similarity metrics for image registration. *Medical Image Analysis*, 87:102830, 2023. 6

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[6] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2

[7] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt im-

age editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1, 3, 5

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

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

[10] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022. 3

[11] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021. 3

[12] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 3

[13] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 5

[14] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models for medical image analysis: A comprehensive survey. *arXiv preprint arXiv:2211.07804*, 2022. 1

[15] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[16] George Klir and Mark Wierman. *Uncertainty-based information: elements of generalized information theory*, volume 15. Springer Science & Business Media, 1999. 6

[17] Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. *arXiv preprint arXiv:2106.00132*, 2021. 3

[18] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner. *arXiv preprint arXiv:2307.15189*, 2023. 2

[19] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6

[21] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 1

[22] 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*, pages 10684–10695, 2022. 1, 2

[23] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[24] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv preprint arXiv:2306.14435*, 2023. 1, 3

[25] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, pages 1–9, 2023. 2

[26] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[27] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[28] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34:1415–1428, 2021. 2

[29] Aiham Taleb, Matthias Kirchler, Remo Monti, and Christoph Lippert. Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20908–20921, 2022. 3

[30] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022. 3

[31] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[32] Tonghe Wang, Yang Lei, Yabo Fu, Jacob F Wynne, Walter J Curran, Tian Liu, and Xiaofeng Yang. A review on medical imaging synthesis using deep learning and its clinical applications. *Journal of applied clinical medical physics*, 22(1):11–36, 2021. 3

[33] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[34] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. 3

[35] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019. 1

[36] Biting Yu, Yan Wang, Lei Wang, Dinggang Shen, and Luping Zhou. Medical image synthesis via deep learning. *Deep Learning in Medical Image Analysis: Challenges and Applications*, pages 23–44, 2020. 3

[37] Angela Zhang, Lei Xing, James Zou, and Joseph C Wu. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering*, 6(12):1330–1345, 2022. 1

[38] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 3

[39] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022. 3

[40] Hong-Yu Zhou, Xiaoyu Chen, Yinghao Zhang, Ruibang Luo, Liansheng Wang, and Yizhou Yu. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nature Machine Intelligence*, 4(1):32–40, 2022. 3