# Anatomical Consistency and Adaptive Prior-informed Transformation for Multi-contrast MR Image Synthesis via Diffusion Model

Yejee Shin[1*] Yeeun Lee[1*] Hanbyol Jang[1]
Geonhui Son[1] Hyeongyu Kim[1] Dosik Hwang[1,2†]
[1] Yonsei University [2] Korea Institute of Science and Technology
{yejeeshin, dldpdms0109, hanstar4, higun2, lion4309, dosik.hwang}@yonsei.ac.kr

## Abstract

*Multi-contrast magnetic resonance (MR) images offer critical diagnostic information but are limited by long scan times and high cost. While diffusion models (DMs) excel in medical image synthesis, they often struggle to maintain anatomical consistency and utilize the diverse characteristics of multi-contrast MR images effectively. We propose APT, a unified diffusion model designed to generate accurate and anatomically consistent multi-contrast MR images. APT introduces a mutual information fusion module and an anatomical consistency loss to preserve critical anatomical structures across multiple contrast inputs. To enhance synthesis, APT incorporates a two-stage inference process: in the first stage, a prior codebook provides coarse anatomical structures by selecting appropriate guidance based on precomputed similarity mappings and Bézier curve transformations. The second stage applies iterative unrolling with weighted averaging to refine the initial output, enhancing fine anatomical details and ensuring structural consistency. This approach enables the preservation of both global structures and local details, resulting in realistic and diagnostically valuable synthesized images. Extensive experiments on public multi-contrast MR brain images demonstrate that our approach significantly outperforms state-of-the-art methods. The source codes are available at* https://github.com/yejees/APT.

## 1. Introduction

Multi-contrast magnetic resonance (MR) images serve as a crucial non-invasive modality for the detection and treatment of brain disorders. Multiple modalities offer complementary information that enhances patient characterization and guides clinical decisions. By integrating data from multiple imaging modalities, clinicians can obtain a more comprehensive view of tumors and lesions, thereby improving diagnostic accuracy and optimizing treatment planning. However, acquiring multi-contrast MR images is time-consuming and costly, posing discomfort for patients and increasing the risk of motion artifacts that reduce image quality [9, 24, 36, 40].

Diffusion Models (DMs) [8, 32] have gained increasing attention in recent research due to their robust mode diversity and high-quality image generation capabilities [6, 37], with growing applications in medical image synthesis [16, 26, 28]. DMs iteratively denoise samples to capture complex structural details, making them valuable for tasks requiring fine-grained anatomical accuracy. However, the stochastic nature of the denoising process can disrupt the preservation of consistent anatomical structures [25], potentially compromising the clinical relevance of synthesized images. Moreover, translation across multiple modalities presents challenges in maintaining anatomical consistency while effectively integrating the diverse characteristics of multi-contrast MR images [14].

In this study, we introduce **APT**, a unified diffusion model that incorporates **A**natomical Consistency and an **A**daptive **P**rior-informed **T**ransformation to ensure anatomically consistent synthesis of multi-contrast MR images. To maintain anatomical consistency throughout the synthesis process, APT implements a comprehensive framework. First, we utilize a pretrained biomedical vision-language model [41] to extract anatomical features from multi-contrast images, which are then integrated through a mutual information fusion module. To enhance structural alignment, we propose an anatomical consistency loss that compares fused features from randomly selected multi-contrast images of the same patient with the overall feature set. These anatomically-aware features guide the diffusion model via cross-attention, ensuring the preservation of critical anatomical structures during synthesis.

The prior codebook is a memory mechanism designed to provide optimally initialized images for missing contrast images during inference. It achieves this by leveraging the

*Equal contribution
†Corresponding author

similarity between each contrast image and augmented versions of multi-contrast images generated using Bézier curve transformation functions [27]. It comprises three key components for each contrast image: index mappings, transformation functions, and intensity scaling weights, enabling the generation of more accurate and high-fidelity images.

For inference, we propose a two-stage process to generate missing contrast MR images. The first stage initializes the generation from an anatomically-informed prior provided by the pre-computed codebook, rather than starting from a complete noise state. This prior serves as a structured starting point, which is enhanced in the subsequent reverse diffusion process [8]. The second stage refines detailed anatomical structures affected by the stochastic diffusion process through iterative unrolling near the final step, applying weighted averaging to ensure anatomical precision in the generated images. This approach ensures both global anatomical consistency and accurate local structural details. We validate our approach with experiments on two public multi-contrast brain MRI datasets, demonstrating that our method outperforms existing methods in multi-contrast MR image synthesis.

In summary, our contributions are as follows:

- We propose APT, an anatomically-aware diffusion model that utilizes multi-contrast MR images to generate accurate and anatomically consistent MR image synthesis.
- We introduce a mutual information fusion module and an anatomical consistency loss to enhance anatomical alignment in synthesized MR images.
- We designed a prior codebook to provide anatomically-informed initialization for target image synthesis.
- Two-stage inference strategy employed prior-informed transformation and iterative unrolling to refine anatomical details, ensuring consistent and accurate MR image generation.

## 2. Related Work

### 2.1. CLIP

Contrastive Language-Image Pre-training (CLIP) [29] is a powerful pre-training framework that has been trained on a large-scale dataset, demonstrating exceptional generalization capabilities and interpretability. This text supervision allows CLIP vision encoder to excel in various tasks [17, 22, 23, 30, 35]. Recently, CLIP has gained prominence in the medical imaging field, leading to the development of biomedical vision-language foundation models (VLMs) built on contrastive learning [7, 34, 41]. PubMedCLIP [7] is a vision-language encoder pretrained on medical image-caption pairs sourced from PubMed articles. BiomedCLIP [41] adapts the CLIP model for the biomedical domain and optimizes the tokenizer and context size to effectively handle lengthy biomedical literature. These domain-specific adjustments enable BiomedCLIP to achieve enhanced performance across various tasks compared to existing biomedical VLMs.

### 2.2. Multi-Contrast MR Image Synthesis

Multi-contrast MR images utilize multiple contrast technique to image the identical anatomical structures, providing essential complementary insights for the diagnosis and assessment of tumors and lesions [12, 20]. However, acquiring multi-modality MR imaging requires substantial time and expense [11, 15, 38], limiting its practicality in clinical settings. Trans-modal MR image synthesis addresses these limitations by mapping available MRI modalities to the target domain of missing modalities, effectively generating missing contrasts and enhancing MRI's clinical utility [3, 10, 42]. Generative adversarial networks (GANs) are widely utilized in medical image translation by learning mappings between distinct image distributions [5, 13, 18, 19, 39, 44]. MM-GAN [31] consists of a CNN-based generator and discriminator, designed to handle diverse source-target modality mappings within a single network. Despite these advances, GANs face inherent instability during training due to the nature of adversarial loss, leading to problems like mode collapse and vanishing gradients [6, 13].

### 2.3. Diffusion for MR Image Synthesis

Diffusion models have recently emerged as a leading approach in generative modeling, known for their robust mode coverage and high-quality image generation [6, 8, 37]. Diffusion models are probabilistic frameworks based on a bidirectional Markov chain, where Denoising Diffusion Probabilistic Models (DDPM) [8] optimizes conditional probability by aligning noise in the forward and reverse processes using a specialized loss function. LDM [30] integrates cross-attention modules within a U-Net architecture, facilitating the generation of images under various conditions with greater efficiency. Diffusion models have shown great potential in the field of medical imaging synthesis [16, 26, 28], offering new possibilities for enhancing diagnostic accuracy in clinical settings by generating target domain images using source domain data. SynDiff [28] is the first adversarial diffusion model that leverages a cycle-consistent architecture for unsupervised training of a conditional diffusion model. M2DN [26] employs a multi-input multi-output framework, enabling efficient modality synthesis across various missing scenarios.

## 3. Methodology

We introduce APT, a unified model for multi-contrast MR image synthesis, integrating anatomical consistency and adaptive prior-informed transformation. Fig. 1 illustrates

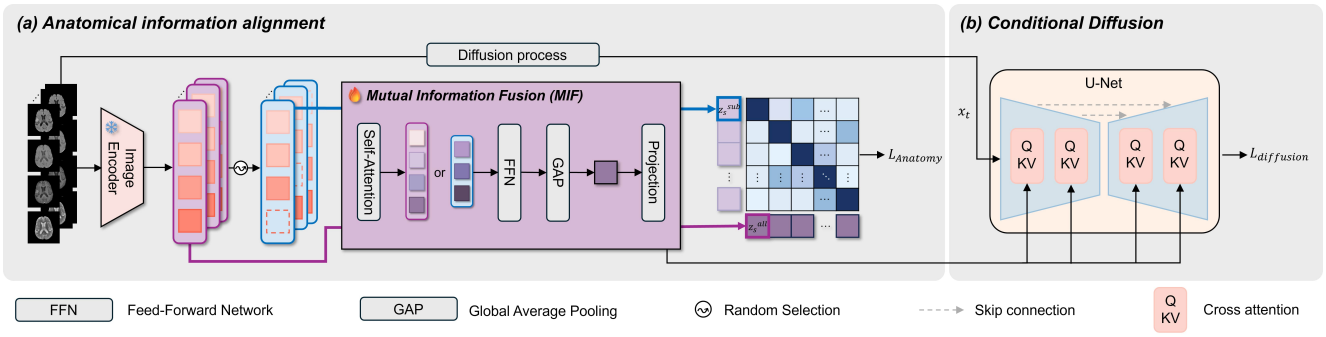


Figure 1. Overview of the proposed training framework for multi-contrast MR image synthesis using a conditional diffusion model. $z_s^{all}$ denotes the vector representing all multi-contrast images of subject $s$, while $z_s^{sub}$ denotes the vector representing a sub-sampled image of subject $u$.

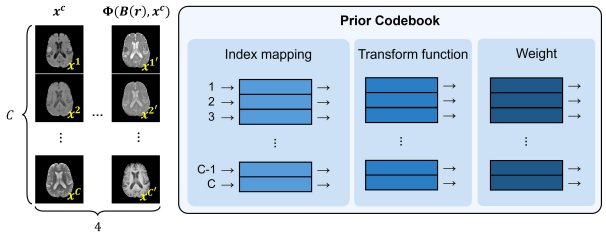


Figure 2. Prior Codebook for Prior-informed Transformation. The left section illustrates augmented versions of target contrast image $x^c$ generated using Bézier curves. $x^{c'}$ denotes transformed images of each contrast $x^c$.

the main training framework, which consists of (a) anatomical information alignment using a mutual information fusion module and (b) a conditional diffusion model. MR image synthesis is guided by a prior codebook (Fig. 2) and employs a two-stage inference strategy (Fig. 3) to refine generated images from coarse to fine details.

### 3.1. Anatomical Information Alignment

**Mutual Information Fusion** Let $x = \{x^1, x^2, \ldots, x^C\}$ represent a set of multi-contrast MR images for the same subject, where each $x^c$ corresponds to a different modality. As demonstrated in [21], text-image joint training can enrich visual features with semantically aligned information, enhancing the model's ability to capture more comprehensive representations. To capture these features from each contrast image, the vision encoder processes each $x^c$ within the pretrained VLM, generating feature maps for each contrast. The mutual information fusion (MIF) module, shown in Fig. 1 (a), utilizes self-attention to learn how each contrast image influences others, capturing dependencies across contrasts. A feed-forward network (FFN) refines the feature maps for enhanced representations. The refined feature maps are aggregated to obtain a compact representation that captures the mutual information across contrasts, resulting in the final feature representation $z_s^{all}$, where $s$ denotes the subject-level fusion of all contrasts.

**Anatomical Consistency Loss** To align the anatomical information between different contrasts effectively, we utilize a contrastive learning framework [29]. Two types of feature vectors are generated: $z_s^{all}$, derived from processing all $C$ contrasts $\{x^1, x^2, \ldots, x^C\}$ through the MIF module, and $z_s^{sub}$, obtained from a randomly selected subset of $C-1$ contrasts. The vectors $z_s^{all}$ and $z_s^{sub}$ represent a positive pair, with other combinations serving as negatives.

The contrastive loss is then defined as:

$$\mathcal{L}_{\text{Anatomy}} = -\log \frac{\exp\left(z_s^{all\top} z_s^{sub}/\tau\right)}{\exp\left(z_s^{all\top} z_s^{sub}/\tau\right) + \sum_{i \neq s} \exp\left(z_s^{all\top} z_i^{s}/\tau\right)} \quad (1)$$

where $\tau$ is a learnable temperature parameter, and $z_i$ represents a negative sample feature vector from a different subject. This loss function promotes similarity between $z_s^{all}$ and $z_s^{sub}$, while discouraging similarity with negative samples $z_i$. This alignment facilitates a unified multi-contrast representation, ensuring anatomical consistency across modalities.

### 3.2. Conditional Diffusion Model

After pretraining the MIF module to capture shared anatomical features across contrasts, we leverage these pretrained features as conditioning inputs for a conditional diffusion model (CDM) [30] to synthesize multi-contrast MR images, as shown in Fig. 1 (b). The feature vector $z_s^{all}$, generated from the pretrained MIF module, is integrated into the U-Net architecture through cross-attention layers. Within these layers, the incorporation of $z_s^{all}$ ensures that anatomical consistency is preserved at each diffusion step, guiding the model to synthesize images that are structurally coherent across contrasts. The objective function is defined as:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t, z_s^{all})\|_2^2\right] \quad (2)$$

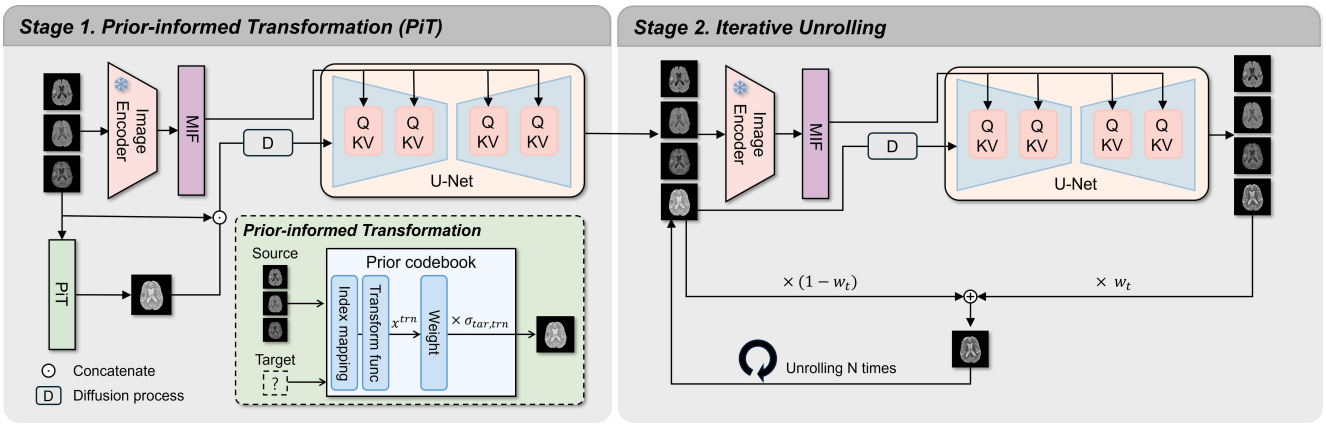


Figure 3. Overview of the proposed inference framework. The process includes (1) structurally consistent initialization using prior-informed guidance to generate coarse anatomical structure, and (2) iterative refinement with weighted averaging to enhance fine details in the synthesized images.

where $x_0$ is the original multi-contrast images that integrate information from all available contrasts, $x_t$ the noisy image at step $t$, $\epsilon$ the added noise, and $\epsilon_\theta(x_t, t, z_s^{all})$ the predicted noise. At each step $t$, the model receives a noisy image $x_t$ and progressively denoises it with $z_s^{all}$, maintaining anatomical consistency.

### 3.3. Prior Codebook

To facilitate effective multi-contrast MR image synthesis, we construct a prior codebook that captures the relationships between different MR contrasts. The codebook stores three essential components: (1) index mappings for high-similarity image pairs, (2) transformation functions, and (3) intensity scaling weights, as shown in Fig. 2.

To generate diverse variations of multi-contrast images, we employ Bézier curves [27] defined with two endpoints ($P_0$ and $P_3$) and two control points ($P_1$ and $P_2$):

$$B(r) = \sum_{i=0}^{n} \binom{n}{i} P_i (1-r)^{n-i} r^i, \quad n = 3,\ r \in [0,1] \quad (3)$$

where $r$ is a parameter along the curve, with domain and range set within $[-1, 1]$. The transformed contrast images are denoted as $\Phi(B(r), x^c)$. Through different Bézier curve parameters, diverse variations are generated for each contrast, resulting in transformed images for each of the C contrasts with 4 different parameter settings.

Using the VLM vision encoder, we compute attention maps between all image pairs in the training dataset to identify similar pairs, excluding self-similar and self-transformed images. These mappings are stored in the codebook for efficient retrieval during inference. For each high-similarity pair identified, we compute contrast-specific scaling factors $\sigma_{\text{tar,trn}}$ by extracting mean values from the images, enabling adjustment of overall intensity levels to match the target contrast characteristics.

### 3.4. Two-Stage Inference Process

The inference process in our framework is designed to synthesize missing MR images through a two-stage approach, as illustrated in Fig. 3.

**Stage1: Prior-informed Transformation** In the first stage, we leverage the prior codebook to generate suitable guidance images for each missing contrast. Given the available contrast images, we first identify corresponding index mappings from the codebook for the missing modality. These mappings are then used to select appropriate transformation functions, which are applied to the source images. The transformed images are further adjusted using pre-computed intensity scaling factors $\sigma_{\text{tar,trn}}$ to ensure global intensity consistency. This prior-informed transformation process generates anatomically consistent guidance images that serve as structural references for the missing contrast synthesis.

The reverse diffusion process starts at intermediate step $M$, with $T > T/2 > M > 0$, where $T$ denotes the total number of steps in the forward diffusion process. This enables the model to begin from a more informed point, preserving structural consistency by leveraging prior knowledge. At each step $t$, the model iteratively refines the predicted image $x_0$ using the following process:

$$\hat{x}_0^{(t)} \leftarrow \hat{x}_0^{(t)} - \alpha \cdot \nabla_{x_t} \| A(\hat{x}_0^{(t)}) - y \|_2^2 \quad (4)$$

where $\hat{x}_0^{(t)}$ represents the predicted image at time step $t$, $\alpha$ is a step-size parameter for gradient descent, $A(.)$ denotes an operator that extracts only the known contrast regions from $\hat{x}_0^{(t)}$, and $y$ represents the known multi-contrast images. This gradient-driven refinement iteratively improves $x_0$, ensuring that the generated image stays consistent with the available contrasts and prior guidance.

Table 1. Quantitative results of the comparison study on BraTS2021 and ADNI datasets. The best PSNR and SSIM values are in bold.

| | BraTS | | | | | | | | ADNI | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FLAIR | | T1-w | | T1Gd | | T2-w | | T1-w | | T2-w | | PD | |
| | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| MM-GAN [31] | 23.35 | 0.882 | 25.13 | 0.925 | 20.40 | 0.869 | 24.93 | 0.920 | 20.40 | 0.784 | 23.61 | 0.820 | 24.5 | 0.826 |
| HiNET [43] | 24.36 | 0.898 | 23.50 | 0.923 | 23.64 | 0.886 | 26.10 | 0.930 | 20.14 | 0.793 | 23.97 | 0.838 | 24.44 | 0.833 |
| ResViT [4] | 23.72 | 0.871 | 24.13 | 0.914 | 23.71 | 0.877 | 26.30 | 0.928 | 19.62 | 0.768 | 23.95 | 0.833 | 24.65 | 0.827 |
| ADM [6] | 21.58 | 0.833 | 22.79 | 0.887 | 23.83 | 0.868 | 22.84 | 0.855 | 16.53 | 0.659 | 21.37 | 0.767 | 22.31 | 0.795 |
| SynDiff [28] | 21.98 | 0.846 | 23.74 | 0.924 | 24.49 | 0.856 | 25.17 | 0.916 | 20.40 | 0.794 | 22.83 | 0.813 | **24.75** | 0.821 |
| M2DN [26] | 23.83 | 0.870 | 24.01 | 0.905 | 22.65 | 0.899 | 24.66 | 0.901 | 19.59 | 0.773 | 22.99 | 0.777 | 23.99 | 0.801 |
| Ours | **25.52** | **0.920** | **25.88** | **0.942** | **24.54** | **0.903** | **27.44** | **0.951** | **20.82** | **0.804** | **24.56** | **0.854** | 24.48 | **0.849** |

**Stage2: Iterative Unrolling** In the second stage, we enhance the details of the coarsely generated image from Stage 1 by iteratively unrolling the conditional diffusion process with a weighted averaging technique, aiming to generate a refined image. After completing $S$ steps of the reverse diffusion process in Stage 1, Stage 2 performs $N$ unrolling iterations, starting from a step close to zero, to progressively refine the initial output. This iterative process improves both anatomical details and overall image quality by fine-tuning the structure at each iteration.

At each iteration $n$ within the current diffusion step $t$, the predicted image $\hat{x}_0^{(t,n+1)}$ is updated through weighted averaging. The update is a combination of the previous prediction $\hat{x}_0^{(t,n)}$ and the current prediction $\hat{x}_0^{(t,n+1)}$, expressed as:

$$\hat{x}_0^{(t,n+1)} \leftarrow \hat{x}_0^{(t,n)} \cdot (1 - w_n) + \hat{x}_0^{(t,n+1)} \cdot w_n \qquad (5)$$

where $w_n$ is the weighting factor at iteration $n$, controlling the balance between the previous and current predictions. This update mechanism aligns with the reverse diffusion process, allowing for iterative refinement of both the global and finer details of the generated image.

# 4. Experiments and Analysis

## 4.1. Setup

**Datasets** We evaluate the model on BraTS 2021 [2] and Alzheimer's Disease Neuroimaging Initiative (ADNI) [1] datasets. The BraTS 2021 dataset comprises a total of 2,040 cases of multi-parametric brain tumor MRI scans, including T1-weighted (T1-w), post-contrast T1-weighted (T1Gd (Gadolinium)), T2-weighted (T2-w), and T2 Fluid-Attenuated Inversion Recovery (FLAIR) volumes. These scans, collected from multiple institutions, reflect variations in imaging protocols and equipments. All images were resized to 256 × 256 pixels and processed as 2D slices. 80% of the data was allocated for training, and the test set includes 200 slices from 50 randomly selected subjects. The ADNI dataset includes 737 MRI scans from both cognitively unimpaired and Alzheimer's Disease (AD) patients. The dataset includes T1-w, T2-w, and Proton Density (PD) sequences, which were resized to 256 × 256 pixels. For training, 80% of the data was used, while the test set consists of 150 slices from 30 randomly selected subjects.

**Implementation Details** The proposed method was implemented with the MIF module trained while maintaining a frozen BiomedCLIP [41] vision encoder. The parameter $C$ was configured to 4 for the BraTS 2021 dataset and 3 for the ADNI dataset, with a subset of $C-1$ randomly selected during each training iteration. The training utilized a batch size of 512 and was conducted for 500 epochs. For MR image synthesis, a UNet-based conditional diffusion model (CDM) [30] was employed. The training was performed for 250k iterations on the BraTS 2021 dataset and 350k iterations on the ADNI dataset, with an Adam optimizer set at a learning rate of $1 \times 10^{-4}$ and a batch size of 16. The diffusion model employed a cosine noise schedule over 1000 timesteps $T$. For Bézier curve transformation, we employed four parameter settings for $(P_0, P_1, P_2, P_3)$: two increasing transformations $(-1, -1, 1, 1)$, $(-1, -0.5, 0.5, 1)$ and two decreasing transformations $(1, 1, -1, -1)$, $(1, 0.5, -0.5, -1)$. The parameters $M$, $S$, $\alpha$, $w_n$ were set to 400, 20, 20, and 0.5 for the BraTS 2021 dataset, and 200, 20, 50, and 0.5 for the ADNI dataset, respectively. In the second stage, the parameter $N$ was set to 50. All experiments were conducted on an Ubuntu 20.04 operating system, utilizing two NVIDIA RTX A6000 Tensor Core GPUs.

## 4.2. Comparison with State-of-the-art Methods

**Quantitative Results** We compared our method with six recent state-of-the-art (SOTA) methods for multi-contrast MR image synthesis, including three advanced generative adversarial networks [4, 31, 43] and three diffusion-based approaches [6, 26, 28]. In Table 1, quantitative comparisons for the BraTS 2021 and ADNI datasets are presented using two widely recognized evaluation metrics: Peak Signal-to-Noise Ratio (PSNR) [33] and Structural Similarity Index (SSIM). The results illustrate the performance in synthesizing the missing target contrast image using $C-1$ provided source contrast images. We evaluate all possible combina-

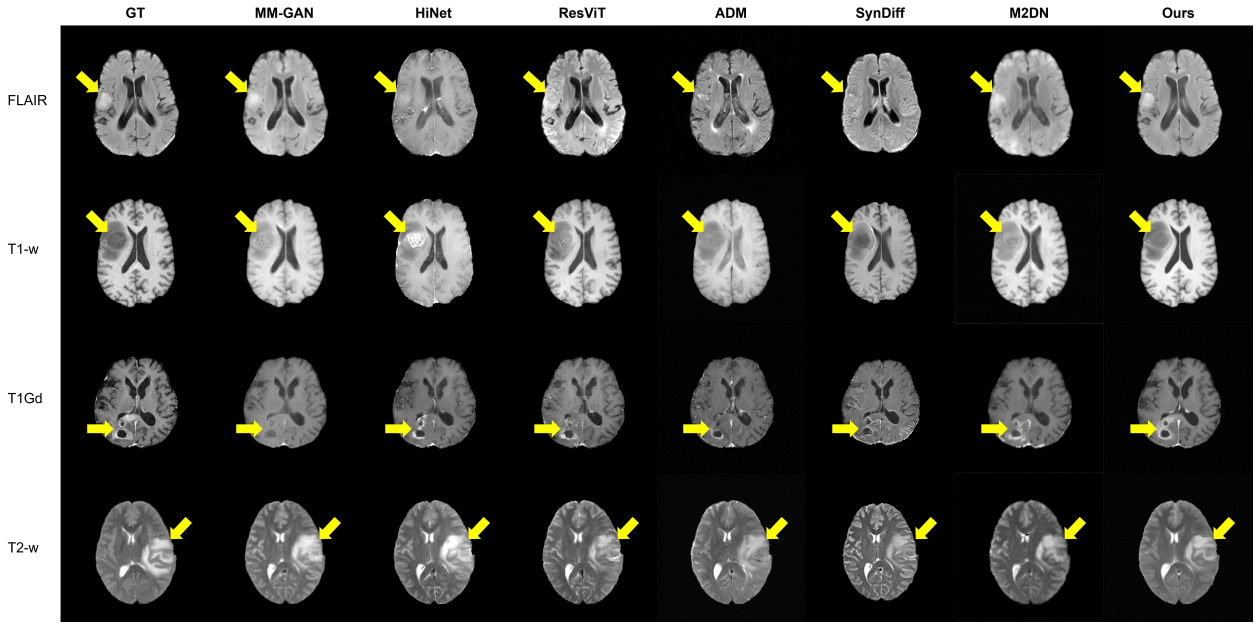


Figure 4. Qualitative results of comparison study on BraTS 2021 dataset. Yellow arrows indicate regions where the methods differ.

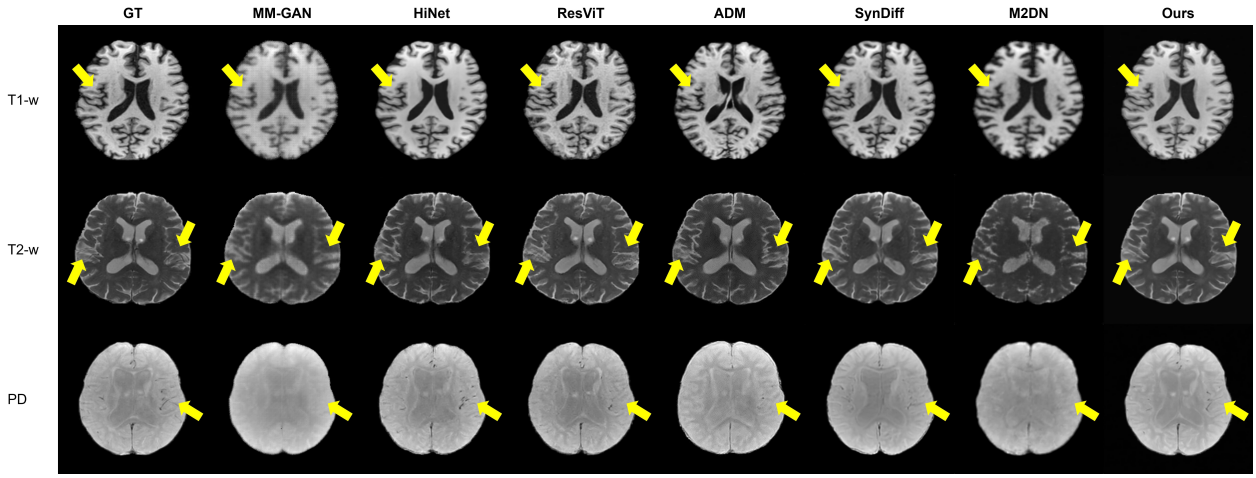


Figure 5. Qualitative results of comparison study on ADNI dataset. Yellow arrows indicate regions where the methods differ.

Table 2. Quantitative results of the ablation study on BraTS2021 and ADNI datasets. CDM denotes conditional diffusion models. U denotes unified synthesis model; A denotes synthesis with adaptive prior information; I denotes synthesis unrolled iteratively in stage 2. For start step $M$, 400 and 200 correspond to BraTS and ADNI datasets, respectively. The best PSNR and SSIM values are in bold.

| | | | BraTS | | | | | | | | ADNI | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | FLAIR | | T1-w | | T1Gd | | T2-w | | T1-w | | T2-w | | PD | |
| Method | Start step M | Sampling steps S | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| CDM (U) | 1000 | 50 | 25.40 | 0.919 | 25.18 | 0.935 | 22.02 | 0.889 | 26.56 | 0.946 | 20.11 | 0.784 | 23.77 | 0.830 | 24.65 | 0.840 |
| CDM (U) | 1000 | 20 | 25.02 | 0.912 | 24.80 | 0.934 | 21.93 | 0.891 | 26.32 | 0.94 | 19.59 | 0.765 | 23.38 | 0.819 | 24.30 | 0.836 |
| CDM (U) | 400 / 200 | 20 | 24.42 | 0.917 | 24.12 | 0.934 | 21.06 | 0.885 | 24.08 | 0.936 | 19.42 | 0.767 | 23.07 | 0.819 | 24.09 | 0.835 |
| CDM (U+A) | 400 / 200 | 20 | 25.09 | 0.915 | 25.58 | 0.938 | **24.81** | 0.902 | 27.30 | 0.947 | 20.45 | 0.796 | 23.71 | 0.831 | 24.46 | 0.839 |
| Ours (U+A+I) | 400 / 200 | 20 | **25.52** | **0.920** | **25.88** | **0.942** | 24.54 | **0.903** | **27.44** | **0.951** | **20.82** | **0.804** | **24.56** | **0.854** | **24.48** | **0.849** |

tions, showing that our method generally achieves higher quantitative results compared to other models. The improvement in evaluation metrics highlights our method's capability to effectively capture intricate details and structural

Table 3. Quantitative results of the ablation study of adaptive selected prior information on BraTS2021 and ADNI datasets. The best PSNR and SSIM values are in bold.

| | BraTS | | | | | | | | ADNI | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FLAIR | | T1-w | | T1Gd | | T2-w | | T1-w | | T2-w | | PD | |
| Prior Information | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| Zero | 24.20 | **0.916** | 24.09 | 0.934 | 20.74 | 0.883 | 24.34 | 0.937 | 19.49 | 0.770 | 23.01 | 0.819 | 24.12 | 0.835 |
| Gaussian Noise | 24.42 | 0.917 | 24.12 | 0.934 | 21.06 | 0.885 | 24.08 | 0.936 | 19.42 | 0.767 | 23.07 | 0.819 | 24.09 | 0.835 |
| Random Prior Selection | 24.84 | 0.915 | 25.09 | 0.935 | 20.89 | 0.884 | 23.50 | 0.929 | 19.69 | 0.774 | 23.58 | 0.828 | **24.50** | 0.837 |
| Adaptive Prior Selection | **25.09** | 0.915 | **25.58** | **0.938** | **24.81** | **0.902** | **27.30** | **0.947** | **20.45** | **0.796** | **23.71** | **0.831** | 24.46 | **0.839** |

Table 4. Quantitative results of the ablation study of stage 2 on BraTS 2021 and ADNI datasets. For start step $M$, 40 and 20 correspond to BraTS and ADNI datasets, respectively.

| | | | BraTS | | | | | | | | ADNI | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | FLAIR | | T1-w | | T1Gd | | T2-w | | T1-w | | T2-w | | PD | |
| Stage | Start step M | Total iterations N | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| 2 | 40 / 20 | 10 | 25.31 | 0.917 | 25.93 | 0.942 | 24.83 | 0.906 | 27.44 | 0.949 | 20.63 | 0.800 | 24.26 | 0.849 | 23.96 | 0.847 |
| 2 | 40 / 20 | 25 | 25.35 | 0.919 | 25.99 | 0.943 | 24.71 | 0.905 | 27.52 | 0.952 | 20.74 | 0.803 | 24.48 | 0.851 | 24.27 | 0.849 |
| 2 | 40 / 20 | 50 | 25.52 | 0.920 | 25.88 | 0.942 | 24.54 | 0.903 | 27.44 | 0.951 | 20.82 | 0.804 | 24.56 | 0.854 | 24.48 | 0.849 |
| 2 | 40 / 20 | 100 | 25.61 | 0.920 | 25.85 | 0.942 | 23.75 | 0.900 | 27.60 | 0.952 | 20.73 | 0.804 | 24.59 | 0.855 | 24.38 | 0.849 |
| 2 | 40 / 20 | 200 | 25.53 | 0.920 | 25.77 | 0.940 | 23.60 | 0.898 | 27.53 | 0.952 | 20.63 | 0.803 | 24.62 | 0.857 | 24.25 | 0.849 |

information in MR images.

**Qualitative Results** Figure 4 and Figure 5 show qualitative comparisons of the existing methods, with Figure 4 for the BraTS 2021 dataset and Figure 5 for the ADNI dataset. As shown in Figure 4, our model preserves anatomical details more effectively, particularly in generating tumor areas. In Figure 5, our method also generates finer details, capturing the cortical distortions characteristic of AD and improving the depiction of detailed structures.

### 4.3. Ablation Study

**Effect of Start Step $M$ and Sampling Steps $S$ in CDM** Table 2 illustrates the results of ablation for the start step $M$ and the number of sampling steps $S$ in the conditional diffusion model (CDM). Larger values of $M$ and $S$ are associated with higher PSNR and SSIM values, indicating improvements in image fidelity and anatomical detail; however, this comes with increased computational cost. In contrast, reducing the sampling steps $S$ leads to faster inference times, though it causes a reduction in PSNR and SSIM, reflecting a potential decrease in image quality.

**Ablation of Model Components** Table 2 presents the ablation study on the unified synthesis model (U), adaptive prior information (A), and iterative unrolling (I) to evaluate their contributions to synthesis performance. Adding adaptive prior information (U+A) improves PSNR and SSIM across both datasets, suggesting that it enhances anatomical accuracy and realism. Further improvement is observed with iterative unrolling in Stage 2 (U+A+I), which helps capture finer anatomical details and reduce artifacts. These results highlight the complementary benefits of adaptive prior information and iterative unrolling in producing high-quality synthesized images.

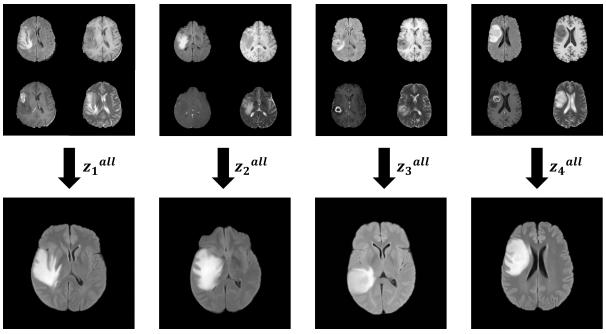


Figure 6. Anatomical alignment analysis using diverse $z^{\text{all}}$. The left column shows the four paired contrast images (FLAIR, T1-w, T1Gd, T2-w) used to generate $z^{\text{all}}$, followed by synthesized FLAIR from this anatomical feature representation.

**Impact of Adaptive Prior Information** Table 3 shows the impact of adaptive prior information on synthesis performance. Different prior selection strategies, including zero prior, Gaussian noise, random prior selection, and adaptive prior selection, are evaluated. The adaptive prior selection generally produces higher PSNR and SSIM values, demonstrating its effectiveness in guiding the synthesis process to generate anatomically realistic images. Notably, for the BraTS dataset, the T1Gd contrast images show a substantial improvement in PSNR, with an increase of approximately 4 dB when using adaptive prior selection rather than random prior selection. This result indicates that incorporating adaptive prior information enhances the model's ability to capture relevant anatomical features across contrasts.

**Effect of Iterative Unrolling in Stage 2** Table 4 presents the results of the ablation study on iterative unrolling in Stage 2. Stage 1 is consistently applied across all exper-

Table 5. Quantitative results evaluating the effect of different starting steps $M$ on synthesis performance. The experiments are conducted without Stage 2.

| | BraTS | | | | | | | | ADNI | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FLAIR | | T1-w | | T1Gd | | T2-w | | T1-w | | T2-w | | PD | |
| Start step M | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| 1000 | 25.02 | 0.912 | 24.80 | 0.934 | 21.93 | 0.891 | 26.32 | 0.94 | 19.59 | 0.765 | 23.38 | 0.819 | 24.30 | 0.836 |
| 500 | 25.47 | 0.918 | 25.32 | 0.937 | 24.46 | 0.901 | 27.16 | 0.947 | 19.84 | 0.777 | 23.35 | 0.823 | 24.15 | 0.837 |
| 400 | 25.09 | 0.915 | 25.58 | 0.938 | 24.81 | 0.902 | 27.30 | 0.947 | 19.91 | 0.780 | 23.46 | 0.825 | 24.35 | 0.838 |
| 200 | 22.90 | 0.879 | 25.23 | 0.939 | 25.12 | 0.903 | 27.06 | 0.947 | 20.45 | 0.796 | 23.71 | 0.831 | 24.46 | 0.839 |
| 100 | 19.75 | 0.829 | 24.55 | 0.935 | 24.61 | 0.899 | 25.02 | 0.933 | 20.40 | 0.797 | 23.93 | 0.839 | 24.23 | 0.838 |
| 40 | 17.97 | 0.782 | 23.26 | 0.921 | 21.34 | 0.878 | 22.83 | 0.905 | 18.70 | 0.769 | 22.67 | 0.825 | 22.07 | 0.808 |
| 20 | 17.44 | 0.761 | 22.02 | 0.91 | 18.70 | 0.858 | 21.99 | 0.889 | 17.45 | 0.748 | 21.25 | 0.799 | 20.76 | 0.750 |

Table 6. Quantitative results for scenarios of random missing modalities on BraTS 2021 dataset.

| Scenarios | | | | FLAIR | | T1-w | | T1Gd | | T2-w | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| FLAIR | T1-w | T1Gd | T2-w | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| ✓ | | | | - | - | 23.54 | 0.914 | 21.20 | 0.869 | 21.67 | 0.904 |
| | ✓ | | | 23.54 | 0.893 | - | - | 22.22 | 0.883 | 25.62 | 0.931 |
| | | ✓ | | 22.25 | 0.869 | 24.27 | 0.912 | - | - | 23.24 | 0.898 |
| | | | ✓ | 23.40 | 0.893 | 23.93 | 0.925 | 23.17 | 0.889 | - | - |
| ✓ | ✓ | | | - | - | - | - | 24.26 | 0.899 | 27.06 | 0.948 |
| ✓ | | ✓ | | - | - | 24.74 | 0.930 | - | - | 23.03 | 0.924 |
| ✓ | | | ✓ | - | - | 24.50 | 0.938 | 21.14 | 0.886 | - | - |
| | ✓ | ✓ | | 24.08 | 0.898 | - | - | - | - | 26.52 | 0.939 |
| | ✓ | | ✓ | 26.40 | 0.923 | - | - | 23.30 | 0.896 | - | - |
| | | ✓ | ✓ | 25.26 | 0.914 | 25.11 | 0.935 | - | - | - | - |

Table 7. Quantitative results for scenarios of random missing modalities on ADNI dataset.

| Scenarios | | | T1-w | | T2-w | | PD | |
|---|---|---|---|---|---|---|---|---|
| T1-w | T2-W | PD | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| ✓ | | | - | - | 22.56 | 0.802 | 22.43 | 0.799 |
| | ✓ | | 20.37 | 0.798 | - | - | 23.98 | 0.851 |
| | | ✓ | 17.04 | 0.663 | 22.05 | 0.764 | - | - |

iments, while Stage 2 explores the impact of varying the number of unrolling iterations ($N$). The results demonstrate that increasing the unrolling iterations in Stage 2 generally improves PSNR and SSIM, indicating that iterative refinement in Stage 2 enhances image quality. A slight performance decrease is observed when $N = 200$, indicating that an excessive number of iterations may lead to overfitting or diminishing returns in terms of image quality. This demonstrates that there is an optimal range for the number of iterations to achieve high-quality synthesized images while avoiding potential performance degradation.

**Anatomical Alignment Analysis** To analyze the effectiveness of anatomical feature alignment, we generate FLAIR image using only $z^{all}$ conditioning from noise, without any additional guidance or refinement steps as shown in Fig. 6. Each row presents synthesized FLAIR image generated using the corresponding $z^{all}$. The distinct results across different $z^{all}$ features, along with the preserved anatomical characteristics from the input contrast combinations, demonstrate that our mutual information fusion module effectively captures and aligns the anatomical features.

**Analysis of Starting Step** $M$ Table 5 demonstrates the impact of different starting steps M on synthesis performance. The results show that initializing from intermediate steps yields better performance compared to starting from random noise. The improvement is attributed to the adaptive prior information at intermediate steps. These steps already contain established structural and contextual details, allowing the model to leverage this information effectively. Consequently, images synthesized from optimal starting steps exhibit higher PSNR and SSIM values, resulting in clearer and more consistent image details.

**Performance with Random Missing Modalities** Table 6 and Table 7 present the quantitative results for scenarios with randomly missing modalities across BraTS and ADNI datasets. For the BraTS dataset, as shown in Table 6, the model maintains consistent performance with PSNR above 23 dB and SSIM above 0.89 in most scenarios, even when multiple input modalities are missing. Similarly, for the ADNI dataset, as shown in Table 7, although performance metrics show some variation depending on which modalities are missing, the model maintains reasonable results. However, when only PD is available as input, the performance metrics show relatively lower values, which can be attributed to the inherently low contrast between gray and white matter in PD-weighted images.

## 5. Conclusion

This paper presents APT, a unified framework for multi-contrast MRI synthesis using conditional diffusion models. Our approach effectively captures anatomical structures using a mutual information fusion module for anatomical consistency. For inference, we propose a two-stage synthesis process that leverages adaptive prior-informed transformation and iterative unrolling with weighted averaging, demonstrating consistent and robust performance across various missing modality scenarios. The results highlight the potential of our method to enhance medical imaging by producing reliable and detailed MR images, contributing to more informed diagnostic and treatment decisions.

**Acknowledgements.** This research was supported by Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Science and ICT (2022R1A2C2008983), Artificial Intelligence Graduate School Program at Yonsei University (RS-2020-II201361), the KIST Institutional Program (2E33801, 2E33800), and Yonsei Signature Research Cluster Program of 2024 (2024-22-0161).

## References

[1] https://adni.loni.usc.edu/. Accessed: 2024-11-04. 5

[2] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. 5

[3] Nicolas Cordier, Hervé Delingette, Minh Lê, and Nicholas Ayache. Extended modality propagation: image synthesis of pathological cases. *IEEE transactions on medical imaging*, 35(12):2598–2608, 2016. 2

[4] Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. Resvit: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10): 2598–2614, 2022. 5

[5] Shayan U. Dar, Mehmet Yurt, Levent Karacan, Aykut Erdem, Emre Erdem, and Tolga Cukur. Image synthesis in multi-contrast mri with conditional generative adversarial networks. *IEEE transactions on medical imaging*, 38(10): 2375–2388, 2019. 2

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 2, 5

[7] Sherzod Eslami, Christoph Meinel, and Gerard De Melo. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, 2023. 2

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2

[9] Kieren Grant Hollingsworth. Reducing acquisition time in clinical mri by data undersampling and compressed sensing reconstruction. *Physics in Medicine & Biology*, 60(21): R297, 2015. 1

[10] Yong Huang, Ling Shao, and Alejandro F Frangi. Cross-modality image synthesis via weakly coupled and geometry co-regularized joint dictionary learning. *IEEE transactions on medical imaging*, 37(3):815–827, 2017. 2

[11] Thao Huynh, Yang Gao, Jian Kang, Litao Wang, Peng Zhang, Jun Lian, and Dinggang Shen. Estimating ct image from mri data using structured random forest and auto-context model. *IEEE Transactions on Medical Imaging*, 35 (1):174–183, 2015. 2

[12] Juan Eugenio Iglesias, Ender Konukoglu, Darko Zikic, Ben Glocker, Koen Van Leemput, and Bruce Fischl. Is synthesizing mri contrast useful for inter-modality analysis? In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part I*, pages 631–638. Springer Berlin Heidelberg, 2013. 2

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[14] L. Jiang, Y. Mao, X. Wang, X. Chen, and C. Li. Cola-diff: Conditional latent diffusion model for multi-modal mri synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 398–408, Cham, 2023. Springer Nature Switzerland. 1

[15] Amod Jog, Aaron Carass, Snehashis Roy, Dzung L Pham, and Jerry L Prince. Random forest regression for magnetic resonance image synthesis. *Medical Image Analysis*, 35: 475–488, 2017. 2

[16] Amir Kazerouni, Ehsan Kheradmand Aghdam, Mohammad Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023. 1, 2

[17] Kyungwon Kim, Yongmoon Lee, Doohyun Park, Taejoon Eo, Daemyung Youn, Hyesang Lee, and Dosik Hwang. Llm-guided multi-modal multiple instance learning for 5-year overall survival prediction of lung cancer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 239–249. Springer, 2024. 2

[18] Sewon Kim, Hanbyol Jang, Seokjun Hong, Yeong Sang Hong, Won C Bae, Sungjun Kim, and Dosik Hwang. Fat-saturated image generation from multi-contrast mris using generative adversarial networks with bloch equation-based autoencoder regularization. *Medical image analysis*, 73: 102198, 2021. 2

[19] Deokjong Lee, Junho Kim, Woo-Jin Moon, and Jong Chul Ye. Collagan: Collaborative gan for missing image data imputation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2487–2496, 2019. 2

[20] Junghoon Lee et al. Multi-atlas-based ct synthesis from conventional mri with patch-based refinement for mri-based radiotherapy planning. In *Medical Imaging 2017: Image Processing*. SPIE, 2017. 2

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3

[22] Jiarui Lin and Shaogang Gong. Gridclip: One-stage object detection by grid-level clip representation learning. *arXiv preprint arXiv:2303.09252*, 2023. 2

[23] Bowen Liu, Donghuan Lu, Dong Wei, Xiang Wu, Yifan Wang, Yue Zhang, and Yefeng Zheng. Improving medi-

cal vision-language contrastive pretraining with semantics-aware triage. *IEEE Transactions on Medical Imaging*, 2023. 2

[24] Michael Lustig, David Donoho, and John M Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007. 1

[25] Q. Lyu and G. Wang. Conversion between ct and mri images using diffusion and score-matching models. *arXiv preprint arXiv:2209.12104*, 2022. 1

[26] Xiangxi Meng, Kaicong Sun, Jun Xu, Xuming He, and Dinggang Shen. Multi-modal modality-masked diffusion network for brain mri synthesis with random modality missing. *IEEE Transactions on Medical Imaging*, 2024. 1, 2, 5

[27] Michael E Mortenson. *Mathematics for computer graphics applications*. Industrial Press Inc., 1999. 2, 4

[28] Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023. 1, 2, 5

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3

[30] 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. 2, 3, 5

[31] Anmol Sharma and Ghassan Hamarneh. Missing mri pulse sequence synthesis using multi-modal generative adversarial network. *IEEE transactions on medical imaging*, 39(4): 1170–1183, 2019. 2, 5

[32] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1

[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. 5

[34] Zeyu Wang, Ziyan Wu, Deepak Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. 2

[35] Ziyue Wang, Jing Liang, Ran He, Ning Xu, Zhilei Wang, and Tieniu Tan. Improving zero-shot generalization for clip with synthesized prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3032–3042, 2023. 2

[36] M. L. Wood and R. M. Henkelman. Mr image artifacts from periodic motion. *Medical Physics*, 12(2):143–151, 1985. 1

[37] Lantian Yang, Zhiqiang Zhang, Yang Song, Sungjun Hong, Rui Xu, Yuting Zhao, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023. 1, 2

[38] Dong H. Ye, Darko Zikic, Ben Glocker, Antonio Criminisi, and Ender Konukoglu. Modality propagation: coherent synthesis of subject-specific scans with data-driven regularization. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part I*, pages 606–613. Springer Berlin Heidelberg, 2013. 2

[39] Mehmet Yurt, Salman Ul Dar, Aykut Erdem, Emre Erdem, Kader Karli Oguz, and Tolga Çukur. mustgan: multi-stream generative adversarial networks for mr image synthesis. *Medical Image Analysis*, 70:101944, 2021. 2

[40] M. Zaitsev, J. Maclaren, and M. Herbst. Motion artifacts in mri: A complex problem with many partial solutions. *Journal of Magnetic Resonance Imaging*, 42(4):887–901, 2015. 1

[41] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 1, 2, 5

[42] Can Zhao, Aaron Carass, Jisung Lee, Yi He, and Jerry L Prince. Whole brain segmentation and labeling from ct using synthetic mr images. In *Machine Learning in Medical Imaging: 8th International Workshop, MLMI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 10, 2017, Proceedings 8*, pages 291–298. Springer International Publishing, 2017. 2

[43] Tao Zhou, Huazhu Fu, Geng Chen, Jianbing Shen, and Ling Shao. Hi-net: hybrid-fusion network for multi-modal mr image synthesis. *IEEE transactions on medical imaging*, 39(9): 2772–2781, 2020. 5

[44] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 2