TexDC: Text-Driven Disease-Aware 4D Cardiac Cine MRI Images Generation

Cong Liu[1], Xiaohan Yuan[1], ZhiPeng Yu[1], and Yangang Wang[1⋆]

School of Automation, Southeast University

Abstract. Generating disease-aware cardiac cine magnetic resonance imaging (cine MRI) images has immense potential in medical research, with recent advancements in text-driven image generation technology offering a viable solution. However, establishing clear correlations between textual descriptions and subtle disease regions, especially in capturing their dynamic complexities within cardiac contexts, remains a challenge. To tackle this, our approach emphasizes pre-aligning textual and cardiac cine MRI image features to highlight critical disease areas, establishing interactive relationships between disease text features and spatiotemporal image features during generation. We propose a text-driven framework for synthesizing disease-aware cardiac cine MRI images. Initially, knowledge is transferred from large language models, refining input semantics by updating learnable contexts. By introducing disease-aware pre-alignment, we emphasize and align key disease features across textual and spatiotemporal dimensions, effectively guiding image generation while maintaining spatiotemporal coherence. To our knowledge, this represents the first application of text-driven medical image generation in 4D modalities. We evaluate the superiority of our method on multi-center cardiac cine MRI datasets. Code is publicly available at https://github.com/me-congliu/TexDC.

Keywords: Text-Driven Generation · Disease-Aware · Spatiotemporal Coherent · Cardiac Cine MRI Images.

1 Introduction

Cardiac cine MRI scans, as a 4D (3D+t) modality, offer both spatial and temporal information, enhancing the detailed diagnosis of cardiac structure and function. Generating disease-aware cardiac cine MRI images has significant potential in cardiac imaging research, aiding in understanding disease processes, early diagnosis, targeted therapy, and treatment response monitoring [22].

⋆ Corresponding author: Yangang Wang. E-mail: yangangwang@seu.edu.cn. All the authors from Southeast University are affiliated with the Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing, China. This work was supported in part by the National Natural Science Foundation of China (No. 62076061), the Natural Science Foundation of Jiangsu Province (No. BK20220127).

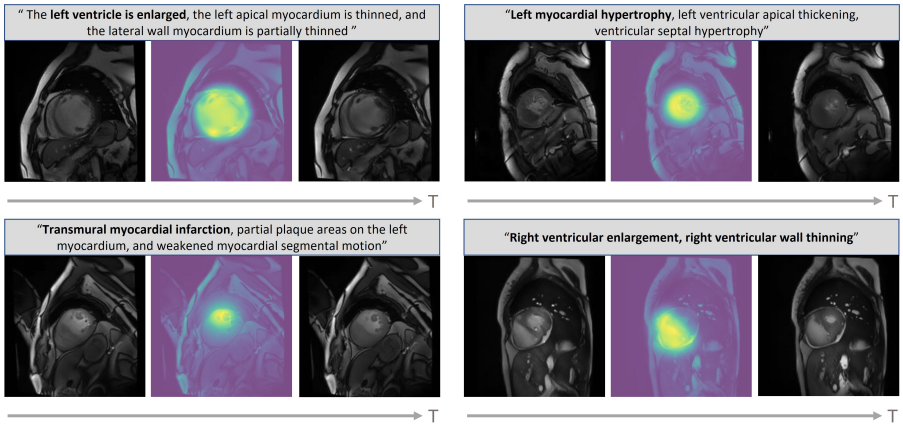

Fig. 1. Our disease-aware and spatiotemporal sequence generation results. Two frames of the sequence are shown, along the temporal dimension, with attention weights of disease-related text keywords (bold) visualized on the images, demonstrating strong visual-textual consistency.

However, training models to generate cine MRI images with specific disease features is hindered by the lack of high-quality data with fine-grained labels. Recent advances in text-driven image generation offer a promising solution, particularly in medical imaging where corresponding medical reports are available [37]. Nevertheless, establishing correlations between text and cardiac cine MRI images, especially at the local level of disease regions, remains challenging. Disease patterns are elusive due to their motion and structural complexity, as well as the small size of affected areas, complicating textual descriptions of cardiac disease dynamics. Despite the growing demand for text-driven medical image generation and the diagnostic advantages of 4D modalities over 2D or 3D, a mature paradigm for handling 4D cardiac cine MRI data has yet to be developed.

To tackle these challenges effectively and capitalize on the strengths of each modality, **our key idea is to focus on pre-aligning textual and cardiac cine MRI image features**. This allows us to concentrate attention on critical disease regions and efficiently establish the interplay between disease-specific textual cues and spatiotemporal image characteristics during generation, as illustrated in Fig. 1. We introduce a text-driven framework for generating disease-aware cardiac cine MRI images. To our knowledge, this marks the first attempt to generate 4D cardiac medical data from textual descriptions, exploring how to maintain authenticity across spatiotemporal details in medical contexts.

Diverging from traditional methods that directly pair medical reports and images for training, our lightweight approach utilizes knowledge transfer from large language model (LLM) tailored to our dataset. We distill disease descriptions from a diagnostic perspective, refining input semantics by updating learnable contexts without the need for extensive model fine-tuning. Once effective text features are acquired, we introduce disease-aware pre-alignment to emphasize

and synchronize critical disease characteristics across textual and spatiotemporal dimensions, particularly focusing on subtle yet crucial disease areas. Leveraging the semantic advantages of text, we construct disease features while harnessing the diffusion model's capability to model probability distributions, capturing spatiotemporal motion patterns and thereby enhancing the generation process. Furthermore, we integrate temporal-aware and slice-aware layers into the pre-trained latent diffusion model (LDM) to ensure temporal and spatial coherence.

In summary, our main contributions are:

• We push text-driven medical image generation into the 4D modality for the first time, achieving cardiac cine MRI text-driven synthesis with disease-awareness.

• We transfer knowledge from LLM and refine input semantics by updating learnable contexts. By introducing disease-aware pre-alignment, we effectively learn associations among disease-related textual keywords, specific image regions, and motion distributions.

• The generation process harnesses the learned interactive relationship between disease text features and spatiotemporal image features, ensuring temporal and spatial coherence. Experimental validation shows the high fidelity of lesion features in generated images compared to detailed descriptions in reports.

2 Related Work

3D medical image generation. Generative Adversarial Networks (GANs) have found application in the field of medical imaging to generate realistic and high-quality 3D images [13, 20, 38]. Previous studies [18, 34] have employed 3D GANs for image generation. Nevertheless, memory limitations during training have confined the generated images to smaller sizes, resulting in a loss of fine details. To address this issue, Sun *et al* [29] introduced a hierarchical structure that concurrently produces a low-resolution image and a randomly selected subvolume of a high-resolution image. Recently, the emergence of Diffusion Probabilistic Models (DPM) has introduced a new solution for high-quality medical image generation, offering better distribution coverage and more stable training [8, 23, 26]. Pinaya *et al* [27] employed LDM with conditions such as age, gender, and ventricular volume to generate high-resolution 3D brain MRI synthetic images. Khader *et al* [16] appended diffusion probabilistic models to the latent space of VQ-GAN for the unconditional generation of the unconditional generation CT and MRI scans. Make-A-Volume [41] introduced a generic paradigm for 3D image synthesis using a 2D backbone, which extends the slice-wise model to be a volume-wise model by insert volumetric layers and fine-tuning the model efficiently. Peng *et al* [25] proposed a memory-efficient 2D conditional DPM to generate high-quality 3D MRI volumes by training an attention network on arbitrary combinations of condition and target slices to learn interdependencies between distant 2D slices. MedGen3D [10] can generate paired 3D medical images and masks by proposing a multi-condition DPM to generate multi-label mask sequences and then utilizing an image sequence generator and semantic diffusion refiner conditioned on the generated mask sequences to produce

realistic 3D medical images aligned with the generated masks. Existing work has begun to attempt to solve sequence generation problems with sequential or temporal dependencies using DPM. Yoon *et al* [36] proposed a sequence-aware diffusion model for generating longitudinal medical images by incorporating a sequence-aware transformer and verified on cardiac and brain datasets. Kim et al. [17] introduced a Diffusion Deformation Model (DDM) composed of diffusion and deformation modules to generate intermediate temporal volumes between source and target volumes, specifically creating 4D cardiac MR images between diastolic and systolic phases for each subject. It only learns motion distribution, whereas we simultaneously model both images and motion of the entire sequence.
Text-driven medical image generation. Text-driven medical image generation enables the creation of personalized medical images by leveraging specific clinical information and medical text descriptions, enhancing the diversity of generated images, compared to models that are not conditioned on text. Recent research have explored the generation of 2D medical images based on medical language text prompts. RoentGen [5] adapts a pre-trained LDM using chest X-ray data and radiology reports, overcoming the large natural medical distributional shift. UniXGen [19], serving as a unified model for bidirectional CXRs and reports generation, has the capability to generate specific view X-rays and leverage multi-view inputs to enhance its generative capacity. Taupetgen [15] can generate 2D tau PET images from textual descriptions and the MR images of subjects based on latent diffusion models. However, extending these models to 3D medical imaging remains challenging due to increased computational complexity. GenerateCT [9] claims to be the first model generating 3D chest CT scans from medical language text prompts. It includes a pre-trained LLM, a transformer-based text-conditioned 3D chest CT generation architecture, and a text-conditioned spatial super-resolution diffusion model. MedSyn [35] can generate high-quality 3D lung CT images conditioned on radiology reports. It starts by synthesizing a low-resolution volume along with its anatomical components and then seamlessly upscales the volume to high resolution.

3 Method

Given a text description about a disease (such as a description in a medical report), we feed it into a learnable prompt using the medical language model BiomedCLIP (Sec. 3.1). Subsequently, we use the text features extracted from BiomedCLIP as the condition and leverage the LDM model to generate an entire 4D sequence (Sec. 3.3). To better enable the image encoder and text encoder to learn features related to heart diseases, we introduce pre-alignment (Sec. 3.2). The overall framework of our method is illustrated in Fig. 2.

3.1 Text-image feature representation

Text features High-quality aligned data for diseased cardiac images and corresponding text is scarce. Directly aligning image-report pairs neglects disease-level semantic correspondences and is constrained by computational resources.

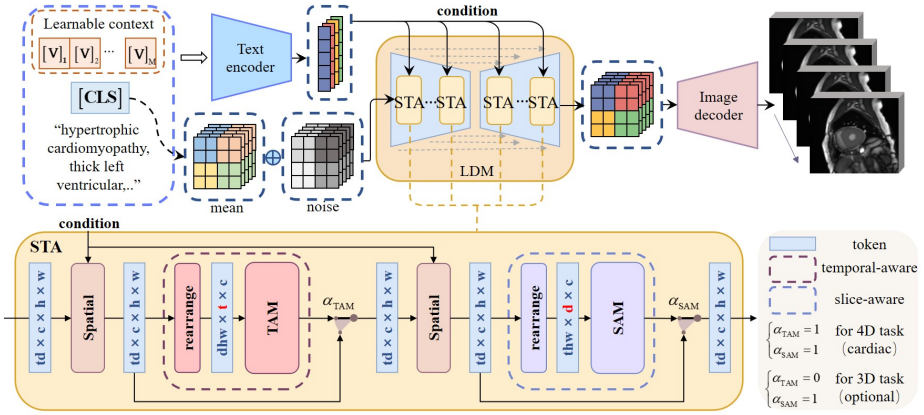

Fig. 2. Pipeline of the proposed cardiac cine MRI generalization. The noise and disease text description are jointly input into the LDM. During the denoising process, the disease text features that have been aligned with the spatiotemporal local features will guide the synthesis of disease images efficiently, while the STA module will model temporal information to conform the disease motion distribution.

Additionally, vague and non-uniform reports bring difficulties for the model in understanding the semantics of images. To address this issue, we incorporate learnable contexts and diagnostic-oriented disease impressions into the prompt architecture, effectively leveraging the knowledge obtained from the pre-trained medical model BiomedCLIP [39] without the need for fine-tuning the large model itself. We employ a few-shot learning approach inspired by CoOp [40] to enable the large model to recognize what thrombus looks like in the context of a cardiac MRI scenario, thereby effectively extracting text features. Specifically, our prompt is defined as follows, primarily consisting of three parts:

$$\boldsymbol{t} = [\mathrm{V}]_1[\ \mathrm{V}]_2 \ldots [\mathrm{V}]_M[\mathrm{CLS}][\mathrm{impression}], \tag{1}$$

where each $[\mathrm{V}]_m (m \in 1, ..., M)$ is a learnable vector with the same dimension as word embeddings and M is a hyperparameter that specifies the number of learnable vectors. [CLS] represents disease classification, where each class has its own embedding. [impression] is the disease-specific critical detail description that is condensed according to the doctors' diagnostic focus. Notably, while [CLS] and [impression] contribute valuable disease-related insights from high-level semantics and critical details respectively, the design of the learnable context is also crucial as it captures nuanced meanings that discrete embedding spaces fall short of conveying. While individual $[\mathrm{V}]_m$ may be task-relevant, their combination can result in nonsensical prompts [40]. Learnable context can automatically integrate complex concepts without manual prompt engineering, effectively transferring and leveraging the knowledge and capabilities of existing large models.

The constructed text prompt $\boldsymbol{t}$ is feeded into the pre-trained text encoder of BiomedCLIP to obtain text tokens $\mathbf{W}_i = \left\{\mathbf{w}_i^1, \mathbf{w}_i^2, \ldots, \mathbf{w}_i^L\right\}$ $(L = 77)$, and

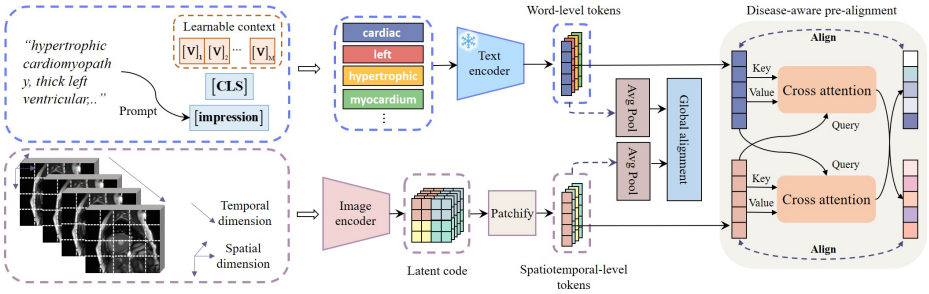

Fig. 3. Pipeline of disease-aware feature representation learning. We employ a large model to extract text features and update the learnable context vectors by aligning global features. We align word-level token embeddings with spatiotemporal tokens to link disease text keywords with local features. Our disease-aware feature representation effectively guides the generation process to learn the specific locations of disease areas and model spatiotemporal information.

further obtain a global feature $\boldsymbol{T}_i$ through an Average Pooling, where i represents the i-th image-report pair, and L represents the number of text tokens.

Image features We assume the input cine MRI is denoted as $x \in \mathbb{R}^{T\times D\times C\times H\times W}$, where the sequence length is T, the number of slices is D, the number of channels is C (where $C = 1$), and the height and width are H and W respectively. We first pre-train an VAE on cine MRI data, extracting its encoder as the image encoder. To facilitate pre-alignment with text, we apply patchify (similar to ViT) after the encoder to obtain spatiotemporal tokens $\mathbf{R}_i = \left\{\mathbf{r}_i^1, \mathbf{r}_i^2, \ldots, \mathbf{r}_i^P\right\}$ ($P = 2940$), and further obtain a global feature $\boldsymbol{I}_i$ through an Average Pooling, where i represents the i-th image-report pair, and P represents the number of spatiotemporal tokens. These tokens are then used for cross attention(pre-alignment) and finally unpatchified back to the original dimensions, resulting in $\mathbf{z}$ for LDM.

3.2 Disease-aware pre-alignment

In medical imaging, disease regions typically constitute only a small portion of the entire image but are crucial due to their diagnostic relevance. Previous studies [14, 32] have also acknowledged the importance of local features, but we are the first to emphasize their role in generation. Here, we introduce a disease-aware pre-alignment method(see Fig. 3) that employs a cross-attention mechanism to establish fine-grained correspondences between spatiotemporal regions and disease descriptions, focusing attention on critical local (disease) areas.

The disease descriptions occupy a significant proportion of tokens in the textual domain, whereas disease regions in the image context account for only a small fraction of spatiotemporal tokens, creating an imbalance. In attention models, tokens are typically treated equally, which results in a more dispersed

distribution of attention when the number of tokens is large. This dispersion can lead to critical disease features being overshadowed by less relevant ones. Therefore, localizing attention is necessary to prioritize the features of disease regions during generation and assign them higher weights accordingly:

$$\mathcal{L}_{\text{expl}} = \lambda_{\text{expl}}(-\frac{1}{|S|}(\sum_{w \in S} \mathcal{R}^{self}_{\text{expl}}(w))), \tag{2}$$

where S is the set of word-level tokens in [impression] and [CLS], and λ_{expl} is the hyperparameter. $\mathcal{R}^{self}_{\text{expl}}$ [6] is an explainable relevance score.

Disease-aware representation is achieved by two symmetrical cross-attention modules [7, 31] to facilitate soft matching between text and images. The input consists of the word-level tokens $\mathbf{W}_i = \left\{\mathbf{w}_i^1, \mathbf{w}_i^2, \ldots, \mathbf{w}_i^L\right\}$ and the spatiotemporal-level tokens $\mathbf{R}_i = \left\{\mathbf{r}_i^1, \mathbf{r}_i^2, \ldots, \mathbf{r}_i^P\right\}$ obtained from Sec. 3.1 for the i-th image-report pair. We alternately treat these two modalities as queries, keys, and values to learn their correct alignment. Formally, for the j-th spatiotemporal-level token embedding $\mathbf{r}_i^j$ in the i-th image-report pair, we consider it as a query, using $\mathbf{w}_i^j$ as both key and value, allowing $\mathbf{r}_i^j$ to attend to all word-level token embeddings in $\mathbf{W}_i$, and then calculate its corresponding cross-modal word-level embedding $\mathbf{o}_i^j$,

$$\mathbf{o}_i^j = \sum_{k=1}^{N} \mathbf{O}(\alpha_i^{j2k}(\mathbf{V}\mathbf{w}_i^k)), \alpha_i^{j2k} = \text{softmax}(\frac{(\mathbf{Q}\mathbf{r}_i^j)^T(\mathbf{K}\mathbf{w}_i^k)}{\sqrt{d}}), \tag{3}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{d \times d}$ are learnable cross-attention parameters. Subsequently, we employ the disease-aware loss to pull $\mathbf{r}_i^j$ toward its cross-modal word-level embedding $\mathbf{o}_i^j$ while pushing it away from other cross-modal word-level embeddings. This maximizes the lower bound of the local cross-modal mutual information within each image-report pair. Therefore, the disease-aware loss for spatiotemporal-level is derived from two symmetrical InfoNCE losses [24]:

$$\mathcal{L}^1_{\text{DAS}} = \log \frac{\exp\left(\mathcal{R}^{cross}_{\text{expl}}\left(\mathbf{r}_i^j, \mathbf{o}_i^j\right)/\tau\right)}{\sum_{k=1}^{P} \exp\left(\mathcal{R}^{cross}_{\text{expl}}\left(\mathbf{r}_i^j, \mathbf{o}_i^k\right)/\tau\right)}, \tag{4}$$

$$\mathcal{L}^2_{\text{DAS}} = \log \frac{\exp\left(\mathcal{R}^{cross}_{\text{expl}}\left(\mathbf{o}_i^j, \mathbf{r}_i^j\right)/\tau\right)}{\sum_{k=1}^{P} \exp\left(\mathcal{R}^{cross}_{\text{expl}}\left(\mathbf{o}_i^j, \mathbf{r}_i^k\right)/\tau\right)}, \tag{5}$$

where $\mathcal{R}^{cross}_{\text{expl}}(\mathbf{a}, \mathbf{b})$ [6] represents the relevance score between token $\mathbf{a}$ and token $\mathbf{b}$, and τ is the token-level temperature hyperparameter.

It is worth noting that, we use relevance scores $\mathcal{R}^{self}_{\text{expl}}$ and $\mathcal{R}^{cross}_{\text{expl}}$ to compute interaction scores instead of the traditional attention values, which only consider the product of queries and keys when handling interactions between different tokens. This approach disregards the impact of intermediate Transformer model components on attention scores. Such oversight can result in unrelated parts

receiving disproportionately high attention scores, thus introducing biases into the interaction process (See Sup.Mat for details).

The overall $\mathcal{L}_{\text{DAS}}$ is defined as:

$$\mathcal{L}_{\text{DAS}} = -\frac{1}{2NP}\sum_{i=1}^{N}\sum_{j=1}^{P}\lambda_i^j\left(\mathcal{L}_{\text{DAS}}^1 + \mathcal{L}_{\text{DAS}}^2\right), \tag{6}$$

where N is the number of image-report pairs, P is the number of spatiotemporal-level tokens, and λ_i^j represents the learnable weight assigned to the j-th spatiotemporal-level token. Since spatiotemporal-level tokens containing pathological information are evidently more critical than those with unrelated information, they are assigned different weights.

Similarly, for the j-th word-level token embedding $\mathbf{w}_i^j$ in the i-th image-report pair, we treat it as a query, using $\mathbf{r}_i^j$ as both key and value, and perform cross-attention operations, obtaining the disease-aware loss at the word-level, denoted as L_{DAW}.

Finally, the overall disease-aware loss is the combination of L_{DAS} and L_{DAW}:

$$\mathcal{L}_{\text{DA}} = -\frac{1}{2}\left(\mathcal{L}_{\text{DAS}} + \mathcal{L}_{\text{DAW}}\right). \tag{7}$$

Note that this loss is used to simultaneously update the cross-attention parameters and the learnable context from Sec. 3.1. This ensures that the word-level tokens focus on specific disease region features rather than just global features. The obtained high-quality, fine-grained feature representations will be utilized to guide the downstream generative task.

Beyond local losses, we employ $\mathcal{L}_{\text{G}}$ to align global semantics:

$$\mathcal{L}_{\text{G}} = -\log\frac{\exp\left(\cos\left(\boldsymbol{I}_i, \boldsymbol{T}_i\right)/\tau\right)}{\sum_{j=1}^{K}\exp\left(\cos\left(\boldsymbol{I}_j, \boldsymbol{T}_j\right)/\tau\right)}, \tag{8}$$

where τ is a learnable temperature parameter and $cos(\cdot,\cdot)$ denotes cosine similarity, K is the number of disease classes. The overall loss $\mathcal{L}$ is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{G}} + \mathcal{L}_{\text{DA}} + \mathcal{L}_{\text{expl}}, \tag{9}$$

3.3 Spatiotemporal Disease Region Generation

We observe distinct motion distributions for different diseases, thus emphasizing the necessity of preserving inter-frame motion information during disease image generation. Inspired by [3] and leveraging the obtained pre-trained LDM, we integrate the spatiotemporal attention mechanism (STA) into the generation process. This integration aims to further learn temporal dynamics and inter-layer correlations, specifically the probabilistic distribution of cardiac deformation. It ensures temporal continuity across frames and spatial coherence among slices in the generated cine MRI (see Fig. 2).

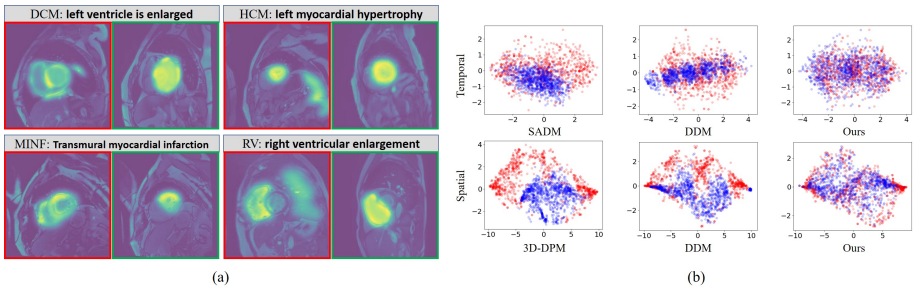

Fig. 4. (a) **Visualization of attention weights for different diseases and ablation study on L_{expl}.** Red boxes (w/o L_{expl}), green boxes (w/ L_{expl}). (b) **Comparing data distributions across temporal and spatial dimensions from different methods.** t-SNE embedding of generated samples (blue) from each model and real samples (red) for temporal features (1st line) and spatial features (2nd line).

Expanding on the pre-trained spatial layer in LDM and utilizing word-level tokens from 3.2 as condition, this representation achieves improved semantic clarity through disease-level alignment. The core concept of the newly integrated STA module lies in envisioning the 4D sequence as a coordinate system with temporal and spatial axes, allowing for attention focus modification through rearrangement operations. Specifically, we introduce an additional temporal-aware attention module (TAM), denoted as l_{TAM}, after the spatial layer to incorporate temporal information into the model. The input latent is $\mathbf{z}_\tau \in \mathbb{R}^{td\times c\times h\times w}$, $\tau = T \dots 0$. Then, we separate the time dimension t from the batch dimension, establishing temporal relationships for these t tokens. Here, c serves as the feature dimension for each token, and we rearrange the data dimension to $\mathbf{z}'_\tau \in \mathbb{R}^{dhw\times t\times c}$. Next, the denoising process is guided by TAM to ensure temporal consistency, with supervised image encoding conducted along the temporal dimension. In all our implementations, $t = 5, d = 3, c = 768, h = 14, w = 14$.

The activation and deactivation of the switch TAM can be achieved through $\alpha_{\text{TAM}} \in [0, 1]$, where when $\alpha_{\text{TAM}} = 0$, it corresponds to activation, and the obtained results serve as inputs for the next stage.

$$\mathbf{z}_{\tau+1} = \alpha_{\text{TAM}}\mathbf{z}_\tau + (1 - \alpha_{\text{TAM}})\, l_{\text{TAM}}(\mathbf{z}'_\tau). \tag{10}$$

In addition to considering temporal relationships in motion, further attention is needed for spatial connections. Since medical images typically consist of stacked 2D slices, we introduce the slice-aware attention module (SAM), denoted as l_{SAM}. After obtaining $\mathbf{z}_{\tau+1}$ and passing through the spatial layer, we separate the layer dimension d from the batch dimension and rearrange the data dimension to $\mathbf{z}'_{\tau+1} \in \mathbb{R}^{thw\times d\times c}$ to establish spatial correlations among these d tokens. Next, the switch $\alpha_{\text{SAM}} \in [0, 1]$ is applied.

$$\mathbf{z}_{\tau+2} = \alpha_{\text{SAM}}\mathbf{z}_{\tau+1} + (1 - \alpha_{\text{SAM}})\, l_{\text{SAM}}(\mathbf{z}'_{\tau+1}), \tag{11}$$

The STA module will be iteratively utilized in the network. It is noteworthy that the utilization of TAM and SAM can be flexibly controlled through switches.

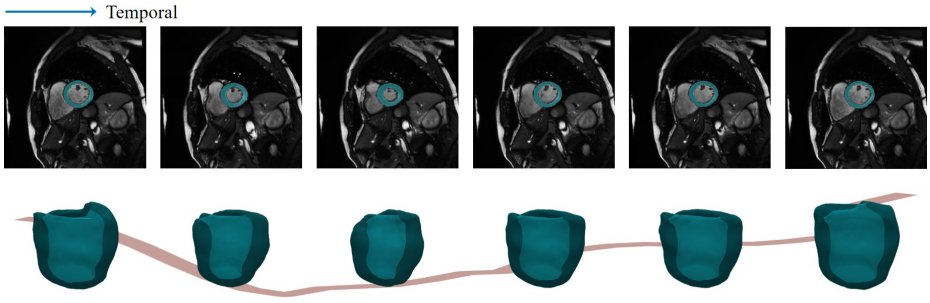

Fig. 5. Visualization of spatiotemporal consistency. 1st line: Middle slices along the temporal dimension, myocardium segmentation highlighted. 2nd line: Mesh sequences generated from stacked slice segmentation, curve representing volume changes.

Activating both switches enables 4D generation tasks (such as cardiac imaging). Alternatively, activating only the SAM allows for general 3D generation tasks (additional applications can be found in Sup.Mat).

During the inference stage, the input condition is the word-level token encoded from the report, and the input is noise sampled with the mean of spatiotemporal tokens under the corresponding disease category.

3.4 Training

Here, we present the training process for our entire framework. Initially, we perform pretraining on an auto-encoder for cine MRI to obtain the image encoder and image decoder. The text encoder utilizes the pretrained encoder from BiomedCLIP. Subsequently, we employ $\mathcal{L}_{\mathrm{G}}$ for global feature constraints and $\mathcal{L}_{\mathrm{DA}}$ for disease-aware representation learning. During this process, we simultaneously update the parameters of the disease-aware alignment cross-attention module and fine-tune both the text encoder and image encoder. Specifically, fine-tuning the text encoder involves optimizing the learnable context in the prompt. This results in a more finely aligned feature space for disease perception, utilized in downstream generation tasks.

The generative network LDM is trained using noise prediction to facilitate denoising. The image input is encoded into the latent space by the image encoder, where noise is added and removed in the latent space. The tokens extracted by BiomedCLIP serve as conditional guidance, with cross-attention aggregating input and conditional information. We adopt a classifier-free training approach [12] and perform sampling using the following linear combination of the conditional and unconditional score estimates,

$$\tilde{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}\left(\mathbf{z}_{\tau}, \mathbf{c}, t\right) = (1-w)\boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\mathbf{z}_{\tau}, \mathbf{c}, t\right) + w\boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\mathbf{z}_{\tau}, t\right), \tag{12}$$

where $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ is the noise generation function, $\mathbf{z}_{\tau}$ is the latent code, $\mathbf{c}$ is the condition, and t denotes the time step. The term w controls the weighting between noise conditioned on both image and text and noise conditioned solely on time.

4 Experiments

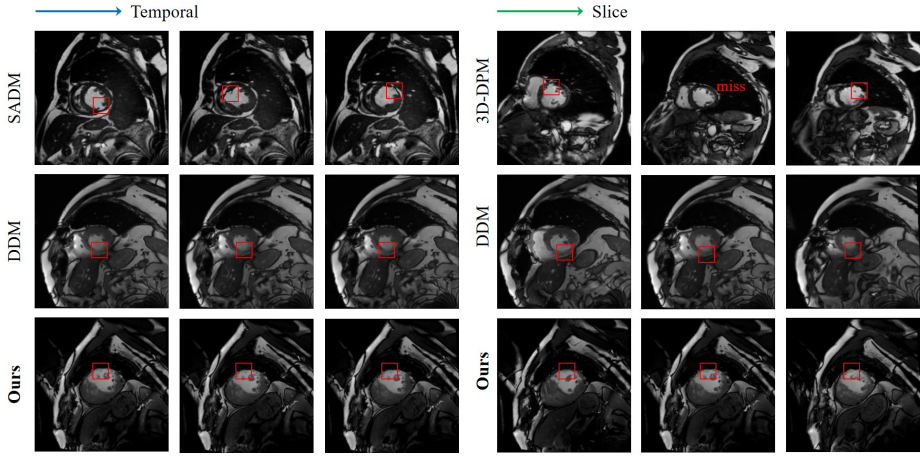

Fig. 6. Comparison of disease region generation across temporal and spatial dimensions. Disease regions highlighted in red boxes, showing three consecutive frames and slices. Our method demonstrates the most stable localization.

4.1 Preparation

Data. Our training dataset comprises data collected from Jiangsu Province Hospital and Subei People's Hospital, along with three publicly available datasets containing disease subjects, including ACDC [2], M&Ms [4], and Data Science Bowl [1], totaling over 1,700 sequences (details are shown in Sup.Mat). The dataset encompasses the following main categories: normal cases (NOR), patients with previous myocardial infarction (MINF), dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), and abnormal right ventricle (RV). We sought professional advice from doctors and obtained specialized descriptions for each category. Due to collection from different centers, each patient includes multiple slices (8-10 slices) with an image resolution of $1.25 \times 1.25mm$, and sequences vary in length. We standardized all data formats, with volume size set to $224 \times 224 \times 9$ and a sequence length of 25. The training was conducted on a piece of RTX 3090 GPU, completing 100,000 iterations. Each batch size is a 4D sequence (25×9 images).

Model architectures. Our LDM is based on [28]. They use convolutional encoders and decoders, and its architecture is built on the U-Net. Our change occurs in Unet, the temporal and spatial dimensions of the temporal attention layer are interleaved in the middle of the original layer, so that the denoising process models the temporal relationship we need. It is noting that the position where the text condition of the original layer added is remains unchanged.

Component	Variants	FID↓	F4D↓	CLIPSIM↑
learnable context vectors	$M = 4$	56.57	764.71	0.2751
	$M = 8$	56.12	760.99	0.2783
	$M = 12$	55.38	754.61	0.2876
	$M = 16$	**54.82**	**743.41**	**0.2917**
	$M = 24$	55.78	762.95	0.2762
pre-alignment (L_{DA}, L_{expl})	w/o L_{DA}	50.81	785.43	0.2149
	w/o L_{expl}	49.87	772.81	0.2382
	w/o pre-alignment	55.37	834.79	0.1955
	w/ pre-alignment	**54.82**	**743.41**	**0.2917**
spatiotemporal alignment	w/o TAM	54.73	863.28	0.2631
	w/o SAM	54.78	855.37	0.2593
	w/ STA module	**54.82**	**743.41**	**0.2917**

Table 1. Ablation study of components in our framework. Ablation experiments prove the effectiveness of each component, including, learnable context vectors, L_{DA} and STA module (SAM and TAM).

We have verified through experiments that the combination of text condition and the timing module we added enables the generation process to capture the correspondence between disease-related text keywords, specific image areas, and motion distributions. For text feature extraction, we use the text encoder of BiomedCLIP model as our text encoder. It is noting that the text encoder of BiomedCLIP remains fixed, while the parameters of learnable context vector and cross attention layers in the disease-aware moudle are updated.

Evaluation metrics. To assess the quality of generation, we use Frechet Inception Distance (FID) [11] as well as Frechet Video Distance (FVD) [30], which measure both visual quality and temporal consistency. Specifically, FVD-T assesses the motion of each slice, while FVD-S evaluates the coherence of different slices within each frame.

Furthermore, we propose a variant for 4D evaluation, named Frechet 4D Distance (F4D), which measures both temporal and spatial consistency. Specifically, we use the pre-trained network as a 4D feature extractor to compute the Fréchet distance based on the feature vector space between two image distributions, similar to FID and FVD. For our text-driven method, we also evaluate CLIP similarity (CLIPSIM) [33] and visualize attention weights (as described in Eqn. 3) for judging the disease-related semantic quality. To further validate the effectiveness of the proposed work in terms of temporal consistency and semantic quality, we additionally perform doctors' evaluation and t-SNE [21] embedding visualization. Additionally, a survey is conducted among medical professionals to evaluate the clinical usability of our method (details are shown in Sup.Mat).

4.2 Ablation study

Ablation study for the learnable context vectors. Table 1 (learnable context vectors) demonstrates consistent performance improvements with the inclusion of learnable context vectors. Minor wording adjustments can significantly

impact performance, and learnable context vectors streamline prompt engineering. However, using more vectors requires additional training data. Our experimental validation suggests that 16 vectors achieve optimal performance when our medical imaging data is limited.

Ablation study for pre-alignment. L_{DA} and L_{expl} ensure that our attention model identifies crucial image regions related to diseases from both local perspectives. L_{expl} directs the model to focus on local regions, while L_{DA} ensures alignment between images and text at the local level. Visualization in Fig. 4 (a) illustrates their roles. Results in the red boxes indicate that relying solely on L_{DA} biases attention, whereas L_{expl} ensures it focuses on the correct disease semantic regions. Results in the green boxes demonstrate that the combination of both losses leads to a more consistent disease-aware representation between text and images. Quantitative results in Tab. 1 (L_{DA}, L_{expl}) show that both losses simultaneously improve F4D and CLIPSIM, indicating that precise alignment of text and images in the representation space enhances the visual-textual consistency during the generation process. The spatiotemporal consistency in disease areas in Fig. 6 further confirms this point.

Ablation study for spatiotemporal alignment. We conduct an ablation study on temporal and spatial consistency using the α_{TAM} and α_{SAM} switches, respectively, as shown in Tab. 1 (STA). The inclusion of TAM and SAM significantly reduces F4D while enhancing both temporal and spatial consistency. To better illustrate spatiotemporal coherence during generation, as depicted in Fig. 5, the myocardial part of all slices is segmented in the generated image sequence, displaying middle slices along the temporal dimension. Inter-layer interpolation is performed on the segmented slices stacked for each frame to generate the corresponding myocardial mesh. We visualize a sequence of meshes and plot the volume change curve over time. It can be observed that the meshes are smooth spatially, while the volume curve accurately reflects the rhythmic cycle of cardiac diastole-systole-diastole over time.

Furthermore, when we deactivate α_{TAM} and activate α_{SAM}, it facilitates the execution of 3D generation tasks with spatial consistency, applicable to most organs. We present the generated images of brain diseases in Sup.Mat, showcasing the scalability of our approach.

4.3 Comparisons

We are the first to achieve cardiac cine MRI text-driven synthesis with disease-awareness and spatiotemporal coherence. Since we are tackling a novel 4D generation task, there are currently no directly comparable methods. Therefore, we consider comparisons with representative works in the 3D modality: SADM [36], DDM [17], and 3D-DPM [8].

Specifically, we decouple time and space to separately evaluate temporal consistency and spatial consistency. The quantitative evaluation in Tab. 2, with lower FID scores and FVD scores, reflects that generated images from our method are more consistent in time and space, closer to the real distribution of diseases. Though sequences from two hospital account for only 5.2% of the

Method	Disease-aware	Spatial-consistent	Temporal-consistent	Text-driven (controllable)	FID↓	FVD-T↓	FVD-S↓
SADM [36]			√		75.28	943.61	–
DDM [17]		√	√		67.34	892.49	879.33
3D-DPM [8]		√			89.71	–	952.67
ours (w/ public)	√	√	√	√	56.13	756.92	769.44
ours	√	√	√	√	**54.82**	**752.33**	**764.21**

Table 2. Capacities and qualitative results of different methods.

entire dataset, we additionally supplemented the experiments conducted only on public datasets, as shown in Tab. 2 (w/ public), which reflects a minimal impact on the metrics and shows that the effect of our method can be verified on the public dataset. Additionally, we utilize t-SNE [21] to reduce the dimensions of temporal and spatial features of generated and real data, visualized in Fig. 4 (b). The higher the overlap, the more similar the data distributions, indicating stronger consistency and superiority of our method.

By aligning disease-related text keywords, specific image areas, and motion distributions, the disease-aware mechanism endows the attended disease features with temporal and spatial attributes, significantly improving the generation of disease images. The results shown in Fig. 6 demonstrate our spatiotemporal consistency, especially in critical disease areas. Specifically, we select three consecutive frames along the temporal axis and three consecutive layers along the slice axis to show, with red boxes highlighting the disease areas of transmural myocardial infarction. It can be observed that the disease regions generated by SADM and 3D-DPM exhibit significant fluctuations or even disappearances in both time and space. Notably, although the results of DDM demonstrate good spatiotemporal consistency similar to ours. This is because DDM interpolates intermediate frames by learning the distribution of motion fields. However, the movement of the heart within the cardiac cycle is not uniformly changing, leading to differences between its output and the actual disease distribution (as shown in Fig. 4 (b)). In summary, our generation results consider both global and local spatiotemporal consistency in disease areas.

5 Conclusion

We introduce a framework that integrates cross-attention between text and cardiac cine MRI, incorporating both disease features and spatiotemporal motion. Our method generates disease-aware, spatiotemporally coherent cardiac cine MRI images, representing a novel approach for 4D cardiac medical data generation from text. Limitations and future work can be found in Sup.Mat.

References

1. Second annual data science bowl. https://kaggle.com/competitions/second-annual-data-science-bowl.
2. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging **37**(11), 2514–2525 (2018)
3. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
4. Campello, V.M., Gkontra, P., Izquierdo, C., Martin-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., et al.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. IEEE Transactions on Medical Imaging **40**(12), 3543–3554 (2021)
5. Chambon, P., Bluethgen, C., Delbrouck, J.B., Van der Sluijs, R., Połacin, M., Chaves, J.M.Z., Abraham, T.M., Purohit, S., Langlotz, C.P., Chaudhari, A.: Roentgen: vision-language foundation model for chest x-ray generation. arXiv preprint arXiv:2211.12737 (2022)
6. Chefer, H., Gur, S., Wolf, L.: Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 397–406 (2021)
7. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Learning universal image-text representations (2019)
8. Dorjsembe, Z., Odonchimed, S., Xiao, F.: Three-dimensional medical image synthesis with denoising diffusion probabilistic models. In: Medical Imaging with Deep Learning (2022)
9. Hamamci, I.E., Er, S., Simsar, E., Tezcan, A., Simsek, A.G., Almas, F., Esirgun, S.N., Reynaud, H., Pati, S., Bluethgen, C., et al.: Generatect: Text-guided 3d chest ct generation. arXiv preprint arXiv:2305.16037 (2023)
10. Han, K., Xiong, Y., You, C., Khosravi, P., Sun, S., Yan, X., Duncan, J.S., Xie, X.: Medgen3d: A deep generative framework for paired 3d image and mask generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 759–769. Springer (2023)
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
12. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
13. Hong, S., Marinescu, R., Dalca, A.V., Bonkhoff, A.K., Bretzner, M., Rost, N.S., Golland, P.: 3d-stylegan: A style-based generative adversarial network for generative modeling of three-dimensional medical images. In: Deep Generative Models, and Data Augmentation, Labelling, and Imperfections: First Workshop, DGM4MICCAI 2021, and First Workshop, DALI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 1. pp. 24–34. Springer (2021)
14. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: 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. pp. 3942–3951 (2021)

15. Jang, S.I., Lois, C., Thibault, E., Becker, J.A., Dong, Y., Normandin, M.D., Price, J.C., Johnson, K.A., Fakhri, G.E., Gong, K.: Taupetgen: Text-conditional tau pet image synthesis based on latent diffusion models. arXiv preprint arXiv:2306.11984 (2023)
16. Khader, F., Müller-Franzes, G., Tayebi Arasteh, S., Han, T., Haarburger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baeßler, B., Foersch, S., et al.: Denoising diffusion probabilistic models for 3d medical image generation. Scientific Reports **13**(1), 7303 (2023)
17. Kim, B., Ye, J.C.: Diffusion deformable model for 4d temporal medical image generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 539–548. Springer (2022)
18. Kwon, G., Han, C., Kim, D.s.: Generation of 3d brain mri using auto-encoding generative adversarial networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 118–126. Springer (2019)
19. Lee, H., Kim, W., Kim, J.H., Kim, T., Kim, J., Sunwoo, L., Choi, E.: Unified chest x-ray and radiology report generation model with multi-view chest x-rays. arXiv preprint arXiv:2302.12172 (2023)
20. Liu, Y., Dwivedi, G., Boussaid, F., Sanfilippo, F., Yamada, M., Bennamoun, M.: Inflating 2d convolution weights for efficient generation of 3d medical images. Computer Methods and Programs in Biomedicine p. 107685 (2023)
21. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11), 2579–2605 (2008)
22. Menchón-Lara, R.M., Simmross-Wattenberg, F., Casaseca-de-la Higuera, P., Martín-Fernández, M., Alberola-López, C.: Reconstruction techniques for cardiac cine mri. Insights into imaging **10**, 1–16 (2019)
23. Müller-Franzes, G., Niehues, J.M., Khader, F., Arasteh, S.T., Haarburger, C., Kuhl, C., Wang, T., Han, T., Nebelung, S., Kather, J.N., et al.: Diffusion probabilistic models beat gans on medical images. arXiv preprint arXiv:2212.07501 (2022)
24. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
25. Peng, W., Adeli, E., Bosschieter, T., Park, S.H., Zhao, Q., Pohl, K.M.: Generating realistic brain mris via a conditional diffusion probabilistic model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 14–24. Springer (2023)
26. Peng, W., Adeli, E., Zhao, Q., Pohl, K.M.: Generating realistic 3d brain mris using a conditional diffusion probabilistic model. arXiv preprint arXiv:2212.08034 (2022)
27. Pinaya, W.H., Tudosiu, P.D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J.: Brain imaging generation with latent diffusion models. In: MICCAI Workshop on Deep Generative Models. pp. 117–126. Springer (2022)
28. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
29. Sun, L., Chen, J., Xu, Y., Gong, M., Yu, K., Batmanghelich, K.: Hierarchical amortized gan for 3d high resolution medical image synthesis. IEEE journal of biomedical and health informatics **26**(8), 3966–3975 (2022)
30. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)

31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
32. Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L.: Multi-granularity cross-modal alignment for generalized medical visual representation learning. Advances in Neural Information Processing Systems **35**, 33536–33549 (2022)
33. Wu, C., Huang, L., Zhang, Q., Li, B., Ji, L., Yang, F., Sapiro, G., Duan, N.: Godiva: Generating open-domain videos from natural descriptions. arXiv preprint arXiv:2104.14806 (2021)
34. Xing, S., Sinha, H., Hwang, S.J.: Cycle consistent embedding of 3d brains with auto-encoding generative adversarial networks. In: Medical Imaging with Deep Learning (2021)
35. Xu, Y., Sun, L., Peng, W., Visweswaran, S., Batmanghelich, K.: Medsyn: Text-guided anatomy-aware synthesis of high-fidelity 3d ct images. arXiv preprint arXiv:2310.03559 (2023)
36. Yoon, J.S., Zhang, C., Suk, H.I., Guo, J., Li, X.: Sadm: Sequence-aware diffusion model for longitudinal medical image generation. In: International Conference on Information Processing in Medical Imaging. pp. 388–400. Springer (2023)
37. Żelaszczyk, M., Mańdziuk, J.: Text-to-image cross-modal generation: A systematic review. arXiv preprint arXiv:2401.11631 (2024)
38. Zhang, K., Hu, H., Philbrick, K., Conte, G.M., Sobek, J.D., Rouzrokh, P., Erickson, B.J.: Soup-gan: Super-resolution mri using generative adversarial networks. Tomography **8**(2), 905–919 (2022)
39. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al.: Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915 (2023)
40. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
41. Zhu, L., Xue, Z., Jin, Z., Liu, X., He, J., Liu, Z., Yu, L.: Make-a-volume: Leveraging latent diffusion models for cross-modality 3d brain mri synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 592–601. Springer (2023)