Diffusion-based Multi-modal MR Fusion for TOF-MRA Image Synthesis

Tianen Yu[1], Xinyu Song[2], Lei Xiang[3], Tao Zhou[1] (✉)

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, China.
taozhou.ai@gmail.com
[2] Shanghai Sixth Peoples's Hospital, China.
[3] Subtle Medical, Menlo Park, USA.

Abstract. Time-of-flight magnetic resonance angiography (TOF-MRA) is widely recognized as the gold standard for non-invasive assessment of cerebrovascular lesions. However, its long scanning times and susceptibility to motion artifacts often result in image blurring and loss of diagnostic information. To address these limitations, the synthesis of TOF-MRA images from multi-modal MR images has emerged as an effective solution. In this paper, we propose a novel Multi-Modal Diffusion Model (MMDM) for TOF-MRA image synthesis, which fully leverages complementary anatomical and pathological information from multi-modal MR images to enhance synthesis performance. Specifically, we introduce modality-specific diffusion modules, each of which independently models the deterministic mapping from a source domain to the target domain, preserving modality-specific prior knowledge. Then, we propose a cross-modal dynamic fusion module to integrate multi-path diffusion features. Additionally, we present a Maximum Intensity Projection (MIP) loss, which constrains the consistency of adjacent slices in the maximum intensity projection space, addressing the issue of vascular discontinuities caused by 2D training. Finally, we propose a Noise-adaptive Weighting Strategy (NAWS) that dynamically balances the multi-objective loss weights based on the data distribution of the diffusion model, ensuring stable convergence during training. Experimental results demonstrate that our method significantly outperforms existing approaches on both the original images and MIP images. Our code is available at https://github.com/taozh2017/MMDM-Syn.

Keywords: Multi-modality, Diffusion model, Max intensity projection

1 Introduction

Multi-modal magnetic resonance imaging, including T1-weighted (T1W), T2-weighted (T2W), and fluid-attenuated inversion recovery (FLAIR) sequences, offers detailed insights into brain anatomy and pathology [5]. Despite their diagnostic synergy [3], Time-of-Flight Magnetic Resonance Angiography (TOF-MRA) is still needed in clinical practice to visualize vascular structures. TOF-MRA is widely used for detecting cerebrovascular abnormalities, stenosis, and hemodynamic changes [2]. However, its long scan times can lead to motion artifacts, causing image blurring and quality degradation [7]. To address these issues, synthesizing TOF-MRA images from multi-modal MR images has emerged as an effective solution.

Existing cross-modal synthesis methods in medical imaging are mainly based on Generative Adversarial Networks (GAN) [17,4,24] and Diffusion Models [11,16]. However, the mode collapse and instability [1] in the GAN training process often result in structural distortions limiting the clinical reliability of the model. Original diffusion models focus on iteratively denoising random noise, gradually aligning the noise distribution with the target data distribution [8,20,19]. Although this mechanism injects diversity into the generated images, it also weakens the determinism of cross-modal mapping due to the uncontrollable interference from random noise. A recently proposed Brownian Bridge Diffusion Model (BBDM [12]) abandons the *noise-to-data* mapping paradigm and instead directly models the diffusion process between the source (*e.g.*, T1W) and the target (*e.g.*, TOF-MRA) domain based on the Brownian bridge diffusion process. By constraining the starting and ending points of the diffusion path, BBDM can explicitly learn the cross-modal feature correspondences.

Despite recent advances in medical image synthesis, which have demonstrated the potential to generate missing modalities [17,4]. Existing approaches primarily focus on single-modality translation [6,22] or simply concatenating multi-modal images during input [4,15,21], often neglecting the rich vascular features embedded in multi-modal MRI data. As a result, synthesized TOF-MRA images often suffer from incomplete vessel continuity or insufficient resolution of microvascular details, limiting their clinical utility.

In this paper, we propose a Multi-Modal Diffusion Model (MMDM) for synthesizing high-fidelity TOF-MRA images from multi-modal MR images. Specifically, the architecture comprises two core components: (1) Three modality-specific diffusion modules, each of which independently learns a deterministic mapping relationship between a source modality and the target TOF-MRA modality, explicitly preserving vascular prior knowledge. (2) Cross-modal dynamic fusion module, which adaptively integrates intermediate features from multiple diffusion paths to mitigate inter-modal conflicts. To address vascular discontinuities in 3D rendering caused by 2D slice training, we present a Maximum Intensity Projection (MIP) loss that enforces distribution alignment of adjacent slices in MIP space, thereby optimizing 3D vascular structural coherence. Furthermore, to tackle the challenge of manual weight tuning in multi-objective optimization, we propose a Noise-adaptive Weighting Strategy (NAWS) based on the data distribution in Brownian bridge diffusion progress. NAWS dynamically adjusts the weight ratio between losses across diffusion timesteps, ensuring stable convergence of multi-objective training. Experimental results show that our model outperforms other state-of-the-art synthesis methods.

The contributions of this paper are highlighted as follows: **i)** To the best of our knowledge, it is the first multi-modal diffusion framework for medical image translation, which jointly optimizes cross-modal diffusion paths from multi-

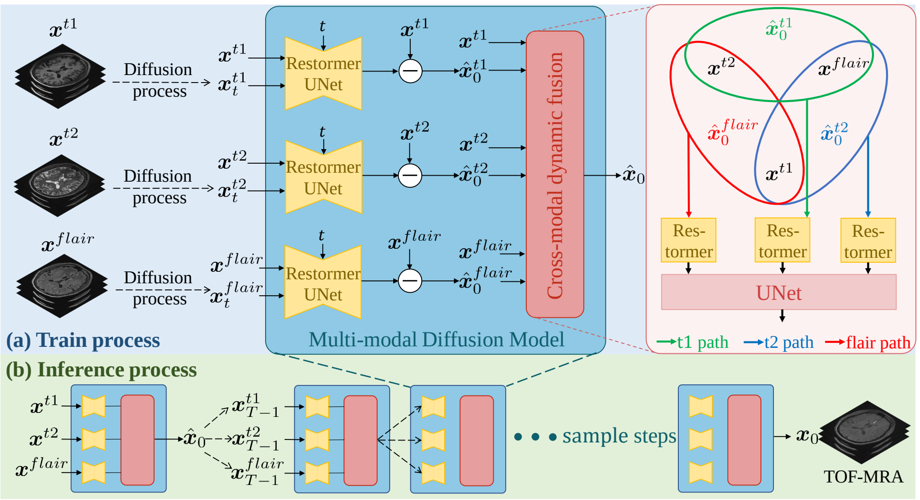

Fig. 1. Overview of our multi-modal diffusion model. ***(a)*** During the training phase, the model generates three modality *independent* predictions $\hat{\boldsymbol{x}}_0^{t1}, \hat{\boldsymbol{x}}_0^{t2}, \hat{\boldsymbol{x}}_0^{flair}$ and a *cooperative* prediction $\hat{\boldsymbol{x}}_0$ all of which are constrained using the MIP loss function $\mathcal{L}_{\mathrm{MIP}}$ and dynamically weighted through Noise-adaptive Weighting Strategy. ***(b)*** In the inference phase, the model utilizes the cooperative prediction $\hat{\boldsymbol{x}}_0$ and applies the reverse process ($\dashrightarrow$) as described in Eq.(2).

modal MR images to TOF-MRA images. **ii)** A vascular continuity-enhanced loss is presented to enforce consistency via maximum intensity projection, effectively mitigating vessel discontinuities caused by 2D slice-wise training. **iii)** A noise-adaptive weighting strategy is presented to dynamically adjust loss weights, ensuring stable optimization under complex training objectives.

2 Proposed Method

Overview: Fig. 1 shows the structure of our model, which generates three independent predictions and dynamically fuses them to obtain the final output.

2.1 Multi-Modal Diffusion Model

BBDM [12] defines the Brownian bridge diffusion process between the source (*e.g.*, T1W) and the target (*e.g.*, TOF-MRA) data distribution. From the two paired data distributions $(\mathcal{X}, \mathcal{Y})$, given a sample $(\boldsymbol{x}, \boldsymbol{y})$, the distribution of the intermediate state at time step t is defined according to diffusion process by

$$q_{BB}(\boldsymbol{x}_t \mid \boldsymbol{x}_0, \boldsymbol{y}) := \mathcal{N}(\boldsymbol{x}_t;\ (1 - m_t)\boldsymbol{x}_0 + m_t\boldsymbol{y},\ \delta_t\boldsymbol{I}), \tag{1}$$

where $\boldsymbol{x}_0 = \boldsymbol{x}$ as the target, $\boldsymbol{y} = \boldsymbol{x}_T$ as the start, $m_t = t/T,\ \delta_t = 2(m_t - m_t^2)$ and T is the total diffusion time steps. Based on the reparameterization trick in

DDPM [20], we can obtain the samples $\boldsymbol{x}_t$ of all data at any time step t within time step T by $\boldsymbol{x}_t = (1-m_t)\boldsymbol{x}_0 + m_t\boldsymbol{y} + \sqrt{\delta_t}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is the noise sampled from a standard normal distribution $\mathcal{N}(0,\ 1)$. By deriving the backward process using the Markov chain and Bayes' theorem, we can obtain the conditional distribution of the reverse process:

$$p(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{y}) := \mathcal{N}(\boldsymbol{x}_{t-1};\ \tilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0, \boldsymbol{y}),\ \tilde{\delta}_t\boldsymbol{I}), \tag{2}$$

where $\tilde{\boldsymbol{\mu}}_t = c_1\boldsymbol{x}_t + c_1\boldsymbol{x}_0 + c_3\boldsymbol{y}$ with $\tilde{\delta}_t,\ c_1,\ c_2,\ c_3$ are constants with respect to time t. As target $\boldsymbol{x}_0$ is unknown, the objective of the model $\boldsymbol{\epsilon}_\theta$ is to fit $\tilde{\boldsymbol{\mu}}_t$ and then, through conditional distribution Eq.(2) and $\boldsymbol{x}_{t-1} = \tilde{\boldsymbol{\mu}}_t^\theta + \sqrt{\tilde{\delta}_t}\boldsymbol{\epsilon}$, sample to reverse along the time step t to recover the target $\boldsymbol{x}_0$. The training process aims to make the two distributions $p(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t,\ \boldsymbol{x}_0,\ \boldsymbol{y})$ and $p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t,\ \boldsymbol{y})$ identical, where θ means the model's parameter. The loss function of this process is

$$\mathcal{L}_{\mathrm{BBDM}} = \|m_t(\boldsymbol{y} - \boldsymbol{x}_{\mathbf{0}}) + \sqrt{\delta_t}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t,\ t)\|_2^2, \tag{3}$$

which is equal to $\|\boldsymbol{x}_t - \boldsymbol{x}_{\mathbf{0}} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\|_2^2$.

We define the symbols as follows: $\boldsymbol{x}_0,\ \boldsymbol{x}^{t1},\ \boldsymbol{x}^{t2},\ \boldsymbol{x}^{flair}$ represent TOF-MRA, T1W, T2W and FLAIR slices. $\boldsymbol{x}_t^m$ are latent slices in diffusion process. $\hat{\boldsymbol{x}}_0^m$ are the *independent* prediction of the target TOF-MRA and $\hat{\boldsymbol{x}}_0$ is the *cooperative* prediction, where $m \in \{t1,\ t2,\ flair\}$.

To avoid detail loss caused by naive concatenation or gating mechanisms, based on BBDM, we design three modality-specific diffusion modules ($\boldsymbol{\epsilon}_\theta^{t1}$, $\boldsymbol{\epsilon}_\theta^{t2}$, $\boldsymbol{\epsilon}_\theta^{flair}$), each built on Restormer-UNet [23]. Every module independently models the Brownian bridge diffusion path from a single modality to TOF-MRA and we add the source modality slice $\boldsymbol{x}^m$ into the model's input as the condition. Unlike the original BBDM, the objective of our diffusion module is $\|\boldsymbol{x}^m - \boldsymbol{x}_0 - \boldsymbol{\epsilon}_\theta^m(\boldsymbol{x}_t^m, t, \boldsymbol{x}^m)\|$. Given different optimization objectives $(\boldsymbol{x}^m - \boldsymbol{x}_0)$ across modules, we unify their outputs $\boldsymbol{\epsilon}_\theta^m(\boldsymbol{x}_t^m, t, \boldsymbol{x}^m)$ into the target domain's latent representation $\hat{\boldsymbol{x}}_0^m$ by $\hat{\boldsymbol{x}}_0^m = \boldsymbol{x}^m - \boldsymbol{\epsilon}_\theta^m$. Restormer significantly reduces computational complexity through a channel-wise attention mechanism, achieving efficient high-resolution feature modeling in image restoration tasks, while preserving global dependencies, making it suitable for detail-sensitive medical imaging tasks.

To resolve modality-specific prediction discrepancies, we introduce the cross-modal dynamic fusion module ($\boldsymbol{\epsilon}_\theta^0$) that adaptively aggregates complementary information from other modalities, generating spatially consistent TOF-MRA predictions. Finally, target TOF-MRA's *cooperative* prediction $\hat{\boldsymbol{x}}_0$ can be represented as:

$$\hat{\boldsymbol{x}}_0 = \boldsymbol{\epsilon}_\theta^0(\hat{\boldsymbol{x}}_0^{t1},\ \hat{\boldsymbol{x}}_0^{t2},\ \hat{\boldsymbol{x}}_0^{flair}, \boldsymbol{x}^{t1}, \boldsymbol{x}^{t2}, \boldsymbol{x}^{flair},\ t) \tag{4}$$

2.2 Maximum Intensity Projection Loss

To eliminate stripe artifacts along non-imaging planes in 3D views, we adopt a 2.5D input approach. However, this strategy may introduce vascular discontinuities in 3D rendering due to the limitations of 2D slice training. Inspired by

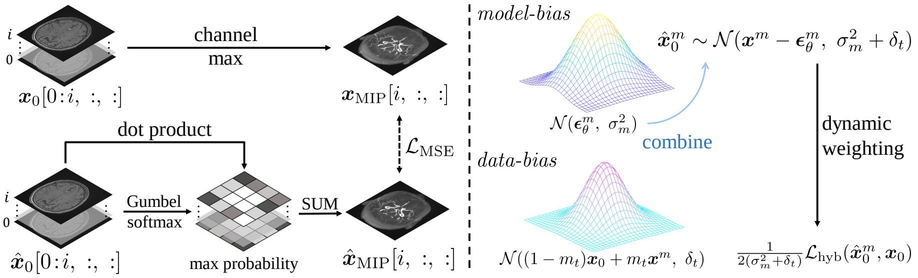
i
0
channel
max
$\boldsymbol{x}_0[0\!:\!i,\ :,\ :]$
$\boldsymbol{x}_{\text{MIP}}[i,\ :,\ :]$
dot product
$\mathcal{L}_{\text{MSE}}$
Gumbel
softmax
SUM
max probability
$\hat{\boldsymbol{x}}_0[0\!:\!i,\ :,\ :]$
$\hat{\boldsymbol{x}}_{\text{MIP}}[i,\ :,\ :]$
model-bias
$\hat{\boldsymbol{x}}_0^m \sim \mathcal{N}(\boldsymbol{x}^m - \boldsymbol{\epsilon}_\theta^m,\ \sigma_m^2 + \delta_t)$
$\mathcal{N}(\boldsymbol{\epsilon}_\theta^m,\ \sigma_m^2)$
combine
dynamic
weighting
data-bias
$\mathcal{N}((1-m_t)\boldsymbol{x}_0 + m_t\boldsymbol{x}^m,\ \delta_t)$
$\frac{1}{2(\sigma_m^2+\delta_t)}\mathcal{L}_{\text{hyb}}(\hat{\boldsymbol{x}}_0^m, \boldsymbol{x}_0)$

Fig. 2. Left: The calculation process of the i-th layer of the GT MIP $\boldsymbol{x}_{\text{MIP}}$ and the predicted MIP $\hat{\boldsymbol{x}}_{\text{MIP}}$. **Right**: We combine the two uncertainties of the diffusion model's output, *model-bias* and *data-bias* to dynamically adjust the weight of $\mathcal{L}_{\text{hyb}}$.

the Maximum Intensity Projection (MIP) can model spatial correlations [18], we propose a MIP loss function to constrain the distribution of maximum values of the generated images. By enforcing the distribution alignment of adjacent slices in the maximum intensity projection space, this loss encourages cross-slice continuity of vascular structures.

In TOF-MRA images, vascular structures are characterized by high-intensity regions. The Gumbel sampling mechanism approximates the distribution of high-intensity signals, enforcing constraints on vascular regions and improving the quality and accuracy of the generated vascular features. In this case, we leverage the Gumbel sampling mechanism to define the MIP loss. Specifically, given a GT image $\boldsymbol{x}_0 \in \mathbb{R}^{N\times W\times H}$, we compute its channel-wise MIP image by

$$\boldsymbol{x}_{\text{MIP}}[i,\ :,\ :] = max(\boldsymbol{x}_0[0\!:\!i,\ :,\ :]) \in \mathbb{R}^{1\times W\times H},\ i \in \{0, ..., N-1\}, \tag{5}$$

where N is the number of 2D slices, and W and H denote the width and height of the image. For the prediction map $\hat{\boldsymbol{x}}_0 \in \mathbb{R}^{N\times W\times H}$, we calculate its MIP image $\hat{\boldsymbol{x}}_{\text{MIP}} \in \mathbb{R}^{N\times W\times H}$ through Gumbel random sampling, which models the maximum values and enable gradient backpropagation by

$$\hat{\boldsymbol{x}}_{\text{MIP}}[i,\ :,\ :] = \sum_{j=0}^{i}(Softmax\left(\frac{\hat{\boldsymbol{x}}_0[0\!:\!j,\ :,\ :] + R_{Gumbel}}{\tau}\right) \cdot \hat{\boldsymbol{x}}_0[0\!:\!j,\ :,\ :]), \tag{6}$$

where $R_{Gumbel} = -ln(-ln(u + 1^{-20}) + 1^{-20})$, τ is a temperature coefficient and $u \in \mathbb{R}^{N\times W\times H}$ is a random matrix following the $\mathcal{U}(0,1)$ distribution. Finally, the MIP loss is calculated as: $\mathcal{L}_{\text{MIP}}(\hat{\boldsymbol{x}}_0,\ \boldsymbol{x}_0) = \mathcal{L}_{\text{MSE}}(\hat{\boldsymbol{x}}_{\text{MIP}},\ \boldsymbol{x}_{\text{MIP}}) + \mathcal{L}_{\text{MSE}}(\hat{\boldsymbol{x}}^r_{\text{MIP}},\ \boldsymbol{x}^r_{\text{MIP}})$, where $\mathcal{L}_{\text{MSE}}$ is the mean square error. $\hat{\boldsymbol{x}}^r_{\text{MIP}}$ and $\boldsymbol{x}^r_{\text{MIP}}$ are MIP images computed by reversing $\hat{\boldsymbol{x}}_0$ and $\boldsymbol{x}_0$ along the channel dimension.

Consequently, the hybrid loss includes two parts: one applied to the original image and the other to the MIP image. The formulation is formulated by

$$\mathcal{L}_{\text{hyd}}(\hat{\boldsymbol{x}}_0,\ \boldsymbol{x}_0) = \mathcal{L}_{\text{MIP}}(\hat{\boldsymbol{x}}_0,\ \boldsymbol{x}_0) + \mathcal{L}_{\text{MSE}}(\hat{\boldsymbol{x}}_0,\ \boldsymbol{x}_0). \tag{7}$$

2.3 Noise-adaptive Weighting Strategy

In this study, two key factors may affect the diffusion model's output, *model-bias* and *data-bias*. The former originates from the model itself, while the latter is introduced by the noise component in the input data of the diffusion model. To address these biases, we present a Noise-adaptive Weighting Strategy. For a regression task, according to the multi-task learning optimization theory [10,13], given the model ϵ_θ and the input x, we have a likelihood function for the probability of GT y given the output $\epsilon_\theta(x)$ as $p(y|\epsilon_\theta) := \mathcal{N}(\epsilon_\theta(x),\ \sigma^2)$, it follows a normal distribution with $\epsilon_\theta(x)$ as the mean and σ as the variance. We refer to this as the *model-bias* effect. Due to length constraints, the model input and GT $\boldsymbol{x}_0$ is omitted in Fig. 2 and Eqs.(8,9). For our modality-specific modules $\boldsymbol{\epsilon}_\theta^{t1},\ \boldsymbol{\epsilon}_\theta^{t2},\ \boldsymbol{\epsilon}_\theta^{flair}$, their input $\boldsymbol{x}_t^m \sim \mathcal{N}(\boldsymbol{x}_t^m;\ (1-m_t)\boldsymbol{x}_0 + m_t\boldsymbol{x}^m,\ \delta_t\boldsymbol{I})$ contains noise with intensity $\sqrt{\delta_t}$, we consider this as *data-bias* for the model prediction:

$$p(\hat{\boldsymbol{x}}_0^m) := \mathcal{N}(\boldsymbol{x}^m - \boldsymbol{\epsilon}_\theta^m,\ \sigma_m^2 + \delta_t). \tag{8}$$

Since the cross-modal dynamic fusion module $\boldsymbol{\epsilon}_\theta^0$ dose not account for data uncertainty introduced by $\boldsymbol{x}_t^m$, for four predictions $\{\hat{\boldsymbol{x}}_0,\ \hat{\boldsymbol{x}}_0^{t1},\ \hat{\boldsymbol{x}}_0^{t2},\ \hat{\boldsymbol{x}}_0^{flair}\}$, we have:

$$p(\hat{\boldsymbol{x}}_0,\ \hat{\boldsymbol{x}}_0^{t1},\ \hat{\boldsymbol{x}}_0^{t2},\ \hat{\boldsymbol{x}}_0^{flair}) := \mathcal{N}(\boldsymbol{\epsilon}_\theta^0,\ \sigma_0^2)\prod_m \mathcal{N}(\boldsymbol{x}^m - \boldsymbol{\epsilon}_\theta^m,\ \sigma_m^2 + \delta_t), \tag{9}$$

and the negative log probability can be expressed by

$$-\log\ p \propto \frac{1}{2\sigma_0^2}\|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0\|^2 + \sum\nolimits_m \frac{1}{2\sigma_m^2 + 2\delta_t}\|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0^m\|^2 + A, \tag{10}$$

where $A = \log\sigma_0 + \frac{1}{2}\sum_m \log(\sigma_m^2 + \delta_t)$. Considering that a very small value of A may lead to an excessively small loss, we introduce a regularization term $C = \ \log\ (1 + \sigma_0^2) + \sum_m \log\ (1 + \sigma_m^2 + \delta_t)$.
Overall Loss Function: the total loss of the model is formulated as follows:

$$\mathcal{L}_{total} = \frac{1}{2\sigma_0^2}\mathcal{L}_{\text{hyd}}(\hat{\boldsymbol{x}}_0,\ \boldsymbol{x}_0) + \sum\nolimits_m \frac{1}{2(\sigma_m^2 + \delta_t)}\mathcal{L}_{\text{hyd}}(\hat{\boldsymbol{x}}_0^m,\ \boldsymbol{x}_0) + C, \tag{11}$$

where σ_0 and σ_m are learnable parameters to adaptive adjustment of the loss functions' weights and $m \in \{t1,\ t2,\ flair\}$.

3 Experiments

Dataset. We implement experiments on an in-house dataset, comprising 560 multi-modal brain MRI scans (*i.e.*, T1W, T2W, FLAIR, and TOF-MRA) from 140 patients (age: 68.15±7.95, 62.14% female) at 2022. All scans were anonymized and registered to TOF-MRA's shape (image size: 288 × 320, number of slices: 213.66 ± 11.94, voxel size: 0.65×0.625×0.625). These paired data are randomly divided into 98 cases for training, 28 for testing, and 14 for validation.

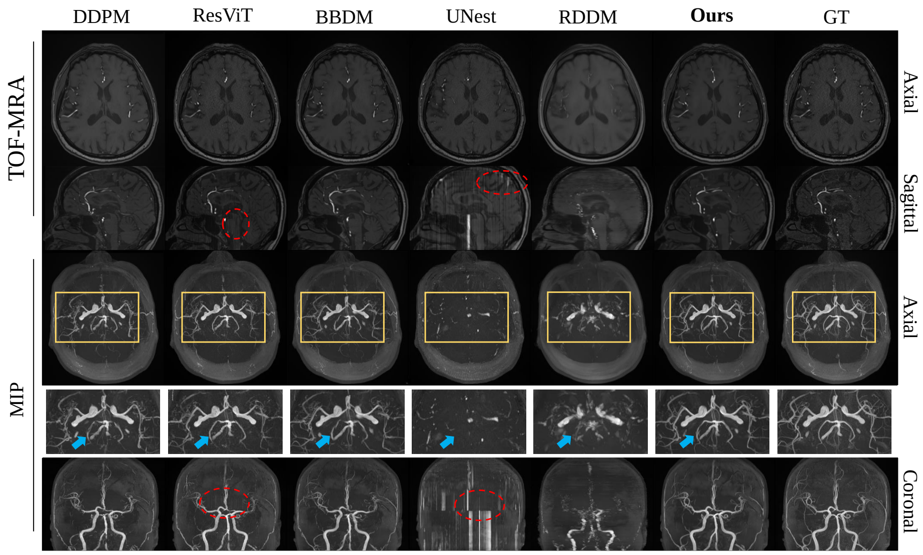

Fig. 3. Qualitative results of different methods. From top to bottom: Axial view of the original TOF-MRA; Sagittal view; Axial MIP; Zoom-in view of axial MIP; and Coronal MIP. All images have been processed by min-max normalization.

Table 1. Quantitative results with comparison methods.

	TOF-MRA			Axial MIP		
Methods	PSNR ↑	SSIM ↑	RMSE(%) ↓	PSNR ↑	SSIM ↑	RMSE(%) ↓
Pix2pix	32.5116	0.8609	2.395	26.1776	0.7721	5.003
DDPM	32.3867	0.8307	2.455	26.0530	0.7508	5.101
ResViT	32.9006	0.8351	2.311	26.3866	0.7346	4.628
BBDM	33.1392	0.8767	2.250	26.3624	0.7750	4.919
UNest	29.6771	0.7898	3.300	21.5071	0.6247	8.476
RDDM	30.1936	0.7897	3.116	22.8504	0.6757	7.280
Ours	**33.6237**	**0.8832**	**2.128**	**27.1845**	**0.792**	**4.463**

Implementation Details. All experiments are conducted in a PyTorch 2.3.1 environment with CUDA 12.0 and two NVIDIA 3090 GPUs. All meth used 2.5D axial slices and all diffusion models are trained with 1000 diffusion steps. Our method employs 200 DDIM sampling steps. The Adam optimizer is employed with a learning rate of 0.0001. Batch size and epochs are set to 4 and 200. The comparison methods are trained with their publicly available codes.

3.1 Comparison with State-of-the-art Methods

Comparison Methods. In this study, we compare our method with several existing GAN and diffusion models, including Pix2Pix [9], DDPM [8], ResViT [4],

Table 2. Quantitative results of ablative studies.

	TOF-MRA			Axial MIP		
Methods	PSNR ↑	SSIM ↑	RMSE(%) ↓	PSNR ↑	SSIM ↑	RMSE(%) ↓
MMDM	32.6016	0.8741	2.379	24.8344	0.7674	5.872
$+\mathcal{L}_{\text{MIP}}$	33.1696	0.8817	2.226	27.0000	0.7903	4.508
+NWAS	33.1019	0.8808	2.241	26.9284	0.7897	4.632
+NWAS, $\mathcal{L}_{\text{MIP}}$	**33.6237**	**0.8832**	**2.128**	**27.1845**	**0.7920**	**4.463**

Table 3. Quantitative results of different modality combinations.

Used modality			TOF-MRA			Axial MIP		
T1W	T2W	FLAIR	PSNR↑	SSIM↑	RMSE(%)↓	PSNR↑	SSIM↑	RMSE(%)↓
✓	✓		32.8268	0.8756	2.310	26.1274	0.7756	5.049
✓		✓	32.6179	0.8553	2.373	25.5160	0.7563	5.426
	✓	✓	32.0945	0.8640	2.508	25.7678	0.7654	5.237
✓	✓	✓	**33.6237**	**0.8832**	**2.128**	**27.1845**	**0.7920**	**4.463**

BBDM [12], UNest [17], and RDDM [14]. ResViT incorporates multi-modal inputs, whereas we employ a gating strategy to integrate multi-modal data as input for other comparison methods.

Results. Table 1 and Fig. 3 shows quantitative and qualitative results for comparison and our methods. Our method demonstrates the highest similarity with both TOF-MRA and axial MIP images. While ResViT achieves good metrics on TOF-MRA, its performance on axial MIP images is subpar. In qualitative result, the vertebral artery of the comparison method was not shown well (blue arrow in Fig. 3) and the GAN-based methods exhibit issues with mode collapse. For instance, in UNest, different cases generate identical vascular structures, and in ResViT, the majority of images exhibit the same noise spots in the center (highlighted by the red dashed circle in Fig. 3). Due to the lack of MIP loss constraint on the maximum values, even though the original TOF-MRA images perform decently, all comparison methods fail to capture the fine details of vascular structures in the axial MIP images.

3.2 Ablation Study

We utilize MMDM as the baseline for our model, which directly aggregates the mean squared errors of the four predictions to construct the loss function. In the ablation study concerning modalities, we eliminate one modality-specific diffusion module to replicate the scenario of synthesizing TOF-MRA from two modalities. As shown in Table 2 and Table 3, we have the following observations: **1)** The proposed MIP loss substantially enhances vascular continuity, improving PSNR, SSIM, and RMSE scores on Axial MIP images by 2.1656 dB, 0.0229, 1.319% compared to baseline; **2)** The NAWS facilitates stable model training through multi-task loss balancing, achieving 2.094 dB higher PSNR; and **3)** The multi-modal ablation study demonstrates that our model with the fusion of three modalities achieves superior image quality compared to all combinations of involving two modalities.

4 Conclusion

We have proposed a multi-modal diffusion model for synthesizing TOF-MRA from multi-modal MR images. Our model establishes deterministic mappings from multi-source domains to the target, overcoming single-modality translation limitations. The MIP loss enhances 3D vascular continuity by constraining maximum intensity distributions, while the noise-adaptive weight strategy balances multi-objectives. Experimental results demonstrate that our model significantly outperforms existing state-of-the-art methods across all metrics. More importantly, the proposed framework can be seamlessly extended to other multi-modal medical image generation tasks.

Acknowledgement. This work was in part supported by the National Natural Science Foundation of China (No. 62172228).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bau, D., Zhu, J.Y., Wulff, J., Peebles, W., Strobelt, H., Zhou, B., Torralba, A.: Seeing what a gan cannot generate. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4502–4511 (2019)
2. Chung, M.S., Jung, S.C., Kim, S.O., Kim, H.S., Choi, C.G., Kim, S.J., Kwon, S.U., Kang, D.W., Kim, J.S.: Intracranial artery steno-occlusion: diagnosis by using two-dimensional spatially selective radiofrequency excitation pulse MR imaging. Radiology 284(3), 834–843 (2017)
3. Cui, C., Yang, H., Wang, Y., Zhao, S., Asad, Z., Coburn, L.A., Wilson, K.T., Landman, B.A., Huo, Y.: Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. Progress in Biomedical Engineering 5(2), 022001 (2023)
4. Dalmaz, O., Yurt, M., Çukur, T.: Resvit: residual vision transformers for multimodal medical image synthesis. IEEE Transactions on Medical Imaging 41(10), 2598–2614 (2022)
5. Doherty, D., Millen, K.J., Barkovich, A.J.: Midbrain and hindbrain malformations: advances in clinical diagnosis, imaging, and genetics. The Lancet Neurology 12(4), 381–393 (2013)
6. Fujita, S., Hagiwara, A., Otsuka, Y., Hori, M., Takei, N., Hwang, K.P., Irie, R., Andica, C., Kamagata, K., Akashi, T., et al.: Deep learning approach for generating MRA images from 3D quantitative synthetic MRI without additional scans. Investigative Radiology 55(4), 249–256 (2020)

7. Fushimi, Y., Fujimoto, K., Okada, T., Yamamoto, A., Tanaka, T., Kikuchi, T., Miyamoto, S., Togashi, K.: Compressed sensing 3-dimensional time-of-flight magnetic resonance angiography for cerebral aneurysms: optimization and evaluation. Investigative Radiology 51(4), 228–235 (2016)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems 33, 6840–6851 (2020)
9. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1125–1134 (2017)
10. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: CVPR. pp. 7482–7491 (2018)
11. Kim, J., Park, H.: Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7604–7613 (2024)
12. Li, B., Xue, K., Liu, B., Lai, Y.K.: Bbdm: Image-to-image translation with brownian bridge diffusion models. In: CVPR. pp. 1952–1961 (2023)
13. Liebel, L., Körner, M.: Auxiliary tasks in multi-task learning. arXiv preprint arXiv:1805.06334 (2018)
14. Liu, J., Wang, Q., Fan, H., Wang, Y., Tang, Y., Qu, L.: Residual denoising diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2773–2783 (2024)
15. Olut, S., Sahin, Y.H., Demir, U., Unal, G.: Generative adversarial training for mra image synthesis using multi-contrast mri. In: PRedictive Intelligence in MEdicine: First International Workshop. pp. 147–154. Springer (2018)
16. Özbey, M., Dalmaz, O., Dar, S.U., Bedel, H.A., Özturk, Ş., Güngör, A., Çukur, T.: Unsupervised medical image translation with adversarial diffusion models. IEEE Transactions on Medical Imaging 42(12), 3524–3539 (2023)
17. Phan, V.M.H., Xie, Y., Zhang, B., Qi, Y., Liao, Z., Perperidis, A., Phung, S.L., Verjans, J.W., To, M.S.: Structural attention: Rethinking transformer for unpaired medical image synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 690–700. Springer (2024)
18. Radhakrishna, C., Chintalapati, K.V., Kumar, S.C.H.R., Sutrave, R., Mattern, H., Speck, O., Nürnberger, A., Chatterjee, S.: Spockmip: Segmentation of vessels in mras with enhanced continuity using maximum intensity projection as loss. arXiv preprint arXiv:2407.08655 (2024)
19. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)
20. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
21. Xia, Y., Ravikumar, N., Lassila, T., Frangi, A.F.: Virtual high-resolution mr angiography from non-angiographic multi-contrast MRIs: synthetic vascular model populations for in-silico trials. Medical Image Analysis 87, 102814 (2023)
22. You, S.H., Cho, Y., Kim, B., Yang, K.S., Kim, B.K., Park, S.E.: Synthetic time of flight magnetic resonance angiography generation model based on cycle-consistent generative adversarial network using PETRA-MRA in the patients with treated intracranial aneurysm. Journal of Magnetic Resonance Imaging 56(5), 1513–1528 (2022)

23. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: CVPR. pp. 5728–5739 (2022)
24. Zhou, T., Fu, H., Chen, G., Shen, J., Shao, L.: Hi-net: hybrid-fusion network for multi-modal MR image synthesis. IEEE Transactions on Medical Imaging 39(9), 2772–2781 (2020)