nnDiffusion: A Standardized 3D Diffusion Framework for Medical Image Synthesis

Haowen Pang[1,*], Pengli Zhu[2,*], Shannan Chen[3], Xiaoming Hong[1], and Chuyang Ye[1(✉)]

[1] School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing, China
chuyang.ye@bit.edu.cn
[2] Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China
[3] College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China

Abstract. Multimodal medical images provide complementary information for clinical diagnosis and treatment planning, but complete acquisitions are often infeasible because of scanning time, patient discomfort, contraindications, and cost. Medical image synthesis can recover missing modalities from available scans, and diffusion models have recently shown promising performance for this task. However, existing methods commonly use heterogeneous preprocessing pipelines, architectures, and training protocols, hindering reproducibility and controlled comparison. We propose nnDiffusion, a standardized 3D diffusion framework for medical image synthesis built upon the nnU-Net pipeline. It implements Denoising Diffusion Probabilistic Models, Flow Matching, and Diffusion Bridges using a common preprocessing pipeline, an automatically configured backbone, and a common training protocol. By inheriting the mature engineering design of nnU-Net, nnDiffusion enables systematic, reproducible, and practical evaluation of diffusion-based synthesis methods for 3D medical images. We validate nnDiffusion on the BraTS dataset with four MRI modalities: T1W, T2W, FLAIR, and T1CE, under both one-to-one and three-to-one synthesis settings. Experimental results demonstrate the feasibility, flexibility, and versatility of the proposed framework across diverse modality synthesis tasks. The code is available at https://github.com/PangHaowen-hub/nnDiffusion.

Keywords: Medical Image Synthesis · Diffusion Model · nnU-Net

1 Introduction

In clinical practice, different medical imaging modalities provide distinct yet complementary views of anatomy and pathology [2,20]. However, acquiring a complete set of modalities for every patient is often impractical due to prolonged scanning time, patient intolerance, motion artifacts, or restrictions on

[*] These authors contributed equally to this work.

contrast-agent administration [4,26]. As a result, cross-modality image synthesis has become an important research direction in medical image analysis [8].

Early studies on medical image synthesis primarily relied on deep regression models [11] or generative adversarial networks (GANs) [21,28]. More recently, diffusion models have emerged as a powerful class of generative models and have shown strong potential for medical image synthesis [19,22,29]. Compared with GAN-based approaches, diffusion models generally offer more stable optimization and a stronger capacity to model complex data distributions, making them particularly attractive for modality synthesis [3,7,10].

Despite the growing popularity of diffusion models, the medical image synthesis community still lacks a standardized 3D implementation framework for diffusion-based medical image synthesis. Existing studies are often implemented as isolated and task-specific codebases, with substantial differences in preprocessing, network backbones, data handling, and training strategies. Recent medical image generation frameworks, such as NV-Generate-CTMR and MAISI [5,27], aim to generate diverse CT or MRI volumes from learned generative priors. In contrast, this work focuses on modality translation, where the goal is to synthesize missing target modalities from available source modalities.

nnU-Net [9] is a widely recognized standard framework for medical image segmentation, owing to its automated preprocessing, adaptive configuration, and robust optimization heuristics. Although originally developed for segmentation, its pipeline provides a strong engineering foundation for reliable 3D medical image synthesis [8,16]. In particular, the standardized design of nnU-Net makes it well suited as a common experimental platform in which different diffusion formulations can be implemented and evaluated under comparable conditions.

In this work, we propose nnDiffusion, an nnU-Net-based framework for medical image synthesis. Our goal is not to introduce a new diffusion model, but to provide the community with a reproducible and extensible implementation platform for representative diffusion paradigms in medical image synthesis. Within this framework, we implement several widely used diffusion formulations, including Denoising Diffusion Probabilistic Models (DDPM), Flow Matching (FM), and Diffusion Bridge (DB). We validate nnDiffusion on the BraTS dataset [1,2,18] using four MRI modalities: T1W, T1CE, T2W, and FLAIR. To assess its flexibility, we consider both one-to-one and three-to-one synthesis settings. Experimental results demonstrate the feasibility, flexibility, and versatility of nnDiffusion across diverse modality synthesis tasks.

2 Methods

2.1 Diffusion Model

Let $\mathbf{x}_{\text{src}}$ denote the available source modality, and let $\mathbf{x}_{\text{trg}}$ denote the target modality to be synthesized. Medical image synthesis aims to learn the conditional distribution $p(\mathbf{x}_{\text{trg}} \mid \mathbf{x}_{\text{src}})$ so that a plausible target-modality image can be generated from the available scans. In this work, we unify DDPM, FM, and

DB under a common conditional trajectory-learning perspective. All variants can be interpreted as learning a parametric neural network $\mathbf{f}_\theta(\mathbf{x}_t, t, \mathbf{x}_{\text{src}})$ that predicts a formulation-specific target variable along a trajectory connecting a simple prior or source-dependent state to the target data distribution.

DDPM. In DDPM [7], the target image $\mathbf{x}_0 = \mathbf{x}_{\text{trg}}$ is progressively perturbed by Gaussian noise through a predefined forward noising process. The noisy sample $\mathbf{x}_t$ at timestep t is defined as

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \quad t \in \{1, \dots, T\}, \tag{1}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$ denotes the cumulative noise schedule.

Following [24], the network learns to recover the clean target image $\mathbf{x}_0$ from $\mathbf{x}_t$ while conditioning on the source modality $\mathbf{x}_{\text{src}}$. We parameterize the model as $\mathbf{f}_\theta(\mathbf{x}_t, t, \mathbf{x}_{\text{src}})$ and use $\mathbf{x}_0$ as the prediction target. The DDPM objective is

$$\mathcal{L}_{\text{DDPM}} = w_t \left\| \mathbf{f}_\theta(\mathbf{x}_t, t, \mathbf{x}_{\text{src}}) - \mathbf{x}_0 \right\|_2^2, \tag{2}$$

where t is uniformly sampled from $\{1, \dots, T\}$. Inspired by Min-SNR weighting [6], we use a clipped SNR-based weight $w_t = \min(\bar{\alpha}_t/(1-\bar{\alpha}_t), \gamma)$, with $\gamma = 5$ by default.

During inference, generation starts from Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ and proceeds through deterministic DDIM [25] sampling conditioned on $\mathbf{x}_{\text{src}}$. At each reverse step, the model first predicts the clean target image as $\hat{\mathbf{x}}_0 = \mathbf{f}_\theta(\mathbf{x}_t, t, \mathbf{x}_{\text{src}})$, and the sample is updated by

$$\mathbf{x}_{t-\Delta t} = \sqrt{\bar{\alpha}_{t-\Delta t}}\,\hat{\mathbf{x}}_0 + \sqrt{\frac{1-\bar{\alpha}_{t-\Delta t}}{1-\bar{\alpha}_t}}\left(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\,\hat{\mathbf{x}}_0\right), \tag{3}$$

where Δt denotes the stride between two sampling timesteps. This process progressively denoises the initial Gaussian sample into the synthesized target modality.

Flow Matching. For FM, the target-modality image distribution is modeled by learning a continuous deterministic transport between a simple prior distribution and the target data distribution [13,15]. Given a paired training sample $(\mathbf{x}_{\text{src}}, \mathbf{x}_{\text{trg}})$ and a prior sample $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$, we denote the target-modality image as $\mathbf{x}_0 = \mathbf{x}_{\text{trg}}$ and the Gaussian prior sample as $\mathbf{x}_1 = \mathbf{z}$. The linear interpolation path is defined as

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \quad t \in [0, 1], \tag{4}$$

where $\mathbf{x}_0$ corresponds to the target image and $\mathbf{x}_1$ corresponds to the Gaussian prior. By differentiating this path with respect to t, the target velocity along the path is obtained as

$$\mathbf{u}_t = \frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}t} = \mathbf{x}_1 - \mathbf{x}_0 = \mathbf{z} - \mathbf{x}_{\text{trg}}. \tag{5}$$

Under this formulation, the network predicts the conditional velocity field along the interpolation path conditioned on the source modality. We parameterize the model as $\mathbf{f}_\theta(\mathbf{x}_t, t, \mathbf{x}_{\text{src}})$ and train it to predict $\mathbf{u}_t$. The FM objective is

$$\mathcal{L}_{\text{FM}} = \|\mathbf{f}_\theta(\mathbf{x}_t, t, \mathbf{x}_{\text{src}}) - \mathbf{u}_t\|_2^2, \tag{6}$$

where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ and t is uniformly sampled from $[0, 1]$.

During inference, generation is performed by solving the ordinary differential equation defined by the learned conditional velocity field in the reverse direction, from $t = 1$ to $t = 0$. Starting from an initial prior sample $\mathbf{x}_1 \sim \mathcal{N}(0, \mathbf{I})$, we use Euler integration:

$$\mathbf{x}_{t-\Delta t} = \mathbf{x}_t - \mathbf{f}_\theta(\mathbf{x}_t, t, \mathbf{x}_{\text{src}})\Delta t. \tag{7}$$

This reverse integration maps a Gaussian prior sample to the conditional target-modality image distribution guided by the source image $\mathbf{x}_{\text{src}}$.

Diffusion Bridge. For the DB, the source and target modalities are directly connected through a stochastic bridge process [12,14]. For notational simplicity, we first describe the one-to-one synthesis setting, where the source and target modalities have the same spatial and channel dimensionality. Let $\mathbf{x}_0 = \mathbf{x}_{\text{trg}}$ denote the clean target image and let $\mathbf{x}_1 = \mathbf{x}_{\text{src}}$ denote the available source modality used as the conditional endpoint at $t = 1$. In the forward bridge process, the target image is gradually transformed toward the source image over continuous time $t \in [0, 1]$. Following image-to-image Brownian bridge and Schrödinger bridge formulations [12,14], we adopt a simplified continuous-time bridge process with linear mean interpolation and a symmetric variance schedule:

$$\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\mathbf{x}_1 + \sigma_{\text{B}}\sqrt{t(1 - t)}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{8}$$

where σ_{B} denotes the bridge noise-scale coefficient and is set to 0.1 in all experiments.

Under this formulation, the network predicts the clean target image from the noisy bridge state conditioned on the source modality. We parameterize the model as $\mathbf{f}_\theta(\mathbf{x}_t, t, \mathbf{x}_{\text{src}})$ and use $\mathbf{x}_0$ as the prediction target. The DB objective is defined as

$$\mathcal{L}_{\text{DB}} = \|\mathbf{f}_\theta(\mathbf{x}_t, t, \mathbf{x}_{\text{src}}) - \mathbf{x}_0\|_2^2, \tag{9}$$

where t is uniformly sampled from $[0, 1]$.

During inference, generation starts from the source condition $\mathbf{x}_1$ and progressively samples the reverse bridge toward the target image. At each reverse step from t to s with $0 \le s < t \le 1$, the model first estimates the clean target image as $\hat{\mathbf{x}}_0 = \mathbf{f}_\theta(\mathbf{x}_t, t, \mathbf{x}_{\text{src}})$. The reverse update is sampled as

$$\mathbf{x}_s = \frac{s}{t}\mathbf{x}_t + \left(1 - \frac{s}{t}\right)\hat{\mathbf{x}}_0 + \eta\sigma_{\text{B}}\sqrt{\frac{s(t - s)}{t}}\,\mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}). \tag{10}$$

Here, η controls the stochasticity of the reverse process. We use $\eta = 1.0$ by default, corresponding to full posterior sampling.

For many-to-one synthesis, a separate bridge is constructed from each source modality to the shared target. The target is replicated across source channels, and the intermediate bridge states are concatenated with all source modalities and processed by the same time-conditioned network. During inference, the source modalities initialize their corresponding reverse bridges, and the final channel-wise target predictions are averaged to produce one synthesized image.

2.2 Backbone Network Architecture

nnDiffusion uses the nnU-Net planner to automatically instantiate the backbone architecture from the training data. The planner determines the patch size, batch size, number of resolution stages, convolution kernel sizes, downsampling strides, and feature-map widths according to the image geometry and the configured memory target.

To adapt the nnU-Net backbone from segmentation to diffusion-based medical image synthesis, temporal conditioning is incorporated through a wrapper around the planned nnU-Net backbone. A sinusoidal timestep encoder maps each training or sampling timestep to a 64-dimensional embedding. For each encoder and decoder stage, a stage-specific multilayer perceptron projects the timestep embedding into feature-wise scale and shift parameters, which are used to modulate the corresponding feature tensors via FiLM conditioning [23]. The network input is formed by channel-wise concatenation of the current state and the available source modalities [24]. Accordingly, the network predicts continuous image-valued quantities rather than discrete segmentation logits. The deep-supervision structure is preserved, where auxiliary prediction heads at intermediate decoder resolutions are optimized with downsampled continuous targets.

2.3 Implementation Details

Data Preprocessing. We follow the original nnU-Net preprocessing pipeline for geometry normalization, including resampling and cropping. Specifically, all images are resampled according to the target spacing determined by the nnU-Net planner. The cropping bounding box is computed exclusively from the nonzero region of the source modalities, and the same source-derived cropping procedure is used during both training and inference. After geometry preprocessing, source and target channels are processed independently. Each channel is clipped at its 0.05th and 99.95th intensity percentiles and then linearly normalized to $[-1, 1]$.

Training Protocol. All models are trained using AdamW [17] with an initial learning rate of 10^{-4}, a weight decay of 10^{-4}, and $(\beta_1, \beta_2) = (0.9, 0.99)$. The learning rate is linearly warmed up for 20 epochs and then decayed to 10^{-6} following a cosine schedule. We further employ gradient clipping with a maximum norm of 1.0, mixed-precision training on CUDA devices, and exponential moving average (EMA) model weights with a decay rate of 0.999. Following the

training-length convention of nnU-Net, each model is trained for 1000 epochs. For diffusion-based optimization, each epoch contains 500 training iterations.

To adapt the nnU-Net training pipeline from discrete segmentation to continuous image synthesis, we redesign both the validation and augmentation strategies. First, the original Dice-based model selection criterion is replaced with validation reconstruction loss. Specifically, we store the negative validation loss in the original nnU-Net checkpoint-selection interface, so that maximizing the moving-average validation score is equivalent to minimizing a smoothed reconstruction loss. The best checkpoint is therefore selected according to the EMA of the validation loss. Second, we revise the nnU-Net augmentation policy to maintain the voxel-wise correspondence between paired source and target images. Intensity perturbations originally designed for segmentation, including Gaussian noise, Gaussian blurring, multiplicative brightness adjustment, contrast adjustment, gamma augmentation, and simulated low-resolution degradation, are removed because they may alter the source-domain appearance without preserving the paired synthesis target. Instead, nnDiffusion retains only synchronized geometric transformations, including rotation, scaling, and mirroring, which are applied consistently to all source and target channels.

Inference and Export. During inference, to handle full-volume MRI scans under GPU memory constraints, sliding-window prediction is applied at every sampling step. The current generative state is concatenated with the source condition and divided into overlapping 3D patches with a step size of 0.5 relative to the planned patch size. Local predictions are fused into a whole-volume estimate using Gaussian-weighted accumulation. We use 100 sampling steps for all three synthesis variants.

After synthesis, the continuous prediction is resampled to the original cropped resolution, inserted into the original volume extent, inverse-transposed to restore the input orientation, and exported as a floating-point medical image with the original geometry metadata. Since target-specific intensity statistics are unavailable at inference time, predictions are exported in the normalized intensity space.

3 Results

We conduct experiments on the BraTS dataset, which contains 1,251 multimodal brain MRI scans for training and validation and 219 scans for testing, with four routinely used modalities: T1W, T2W, FLAIR, and T1CE. We construct both one-to-one and three-to-one synthesis tasks across all four modalities. The fidelity of synthesized images is measured by peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). For the three-to-one setting, we further evaluate tumor-structure preservation through downstream segmentation consistency. An nnU-Net segmentation model is first trained using real multimodal MRI inputs. At test time, we compare the segmentation produced from all real modalities with the segmentation produced after replacing one real modality with its synthesized counterpart while keeping the remaining three modalities

Table 1. Quantitative comparison of one-to-one cross-modality MRI synthesis.

Method	T1W-to-T2W		T1W-to-FLAIR		T1W-to-T1CE		T2W-to-T1W		T2W-to-FLAIR		T2W-to-T1CE	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DDPM	23.46	0.882	21.92	0.834	23.44	0.857	21.38	0.842	22.24	0.847	22.28	0.855
FM	26.15	0.902	25.02	0.863	25.61	0.871	25.28	0.893	25.49	0.873	25.81	0.873
DB	24.56	0.889	24.20	0.847	25.66	0.870	24.91	0.893	25.39	0.865	25.27	0.866
Method	FLAIR-to-T1W		FLAIR-to-T2W		FLAIR-to-T1CE		T1CE-to-T1W		T1CE-to-T2W		T1CE-to-FLAIR	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DDPM	22.87	0.868	23.57	0.882	22.84	0.852	23.40	0.874	23.87	0.885	22.07	0.839
FM	25.87	0.897	26.45	0.903	25.56	0.867	26.55	0.903	26.43	0.901	25.28	0.867
DB	24.87	0.885	25.29	0.889	24.81	0.856	26.77	0.908	25.52	0.891	24.31	0.851

Note: Cell colors indicate Best, Second-best, and Lowest results per column.

Table 2. Quantitative comparison of three-to-one missing-modality MRI synthesis. DSC_{WT}, DSC_{TC}, and DSC_{ET} denote the DSC of the whole tumor, tumor core, and enhancing tumor regions, respectively.

Method	Three-to-T1W					Three-to-T2W				
	PSNR	SSIM	DSC_{WT}	DSC_{TC}	DSC_{ET}	PSNR	SSIM	DSC_{WT}	DSC_{TC}	DSC_{ET}
DDPM	23.43	0.874	0.981	0.942	0.911	23.00	0.888	0.967	0.942	0.909
FM	27.47	0.915	0.986	0.952	0.932	27.42	0.915	0.977	0.962	0.946
DB	26.91	0.910	0.986	0.954	0.917	26.17	0.901	0.972	0.957	0.945
Method	Three-to-FLAIR					Three-to-T1CE				
	PSNR	SSIM	DSC_{WT}	DSC_{TC}	DSC_{ET}	PSNR	SSIM	DSC_{WT}	DSC_{TC}	DSC_{ET}
DDPM	22.29	0.847	0.819	0.915	0.900	23.60	0.863	0.938	0.427	0.203
FM	26.78	0.884	0.914	0.942	0.926	26.50	0.881	0.962	0.625	0.380
DB	25.62	0.865	0.890	0.941	0.921	26.20	0.874	0.953	0.482	0.214

Note: Cell colors indicate Best, Second-best, and Lowest results per column.

real. Dice similarity coefficient (DSC) is then computed between these two predicted masks for whole tumor (WT), tumor core (TC), and enhancing tumor (ET) regions.

Tables 1 and 2 summarize the quantitative results under the one-to-one and three-to-one synthesis settings, respectively. In the one-to-one setting, FM achieves the best performance in most modality pairs, demonstrating its strong ability to reconstruct target-modality intensities and preserve structural details. DB also shows competitive performance and obtains the best results on several pairs, such as T1CE-to-T1W, suggesting that the bridge-based formulation is also effective for cross-modality synthesis. In the three-to-one setting, incorporating multiple source modalities generally improves synthesis performance, particularly for FM and DB, although the degree of improvement varies across methods and target modalities. This indicates that multi-source inputs can provide complementary anatomical and contrast information, but their benefits depend on the target modality and the synthesis strategy. Across different targets,

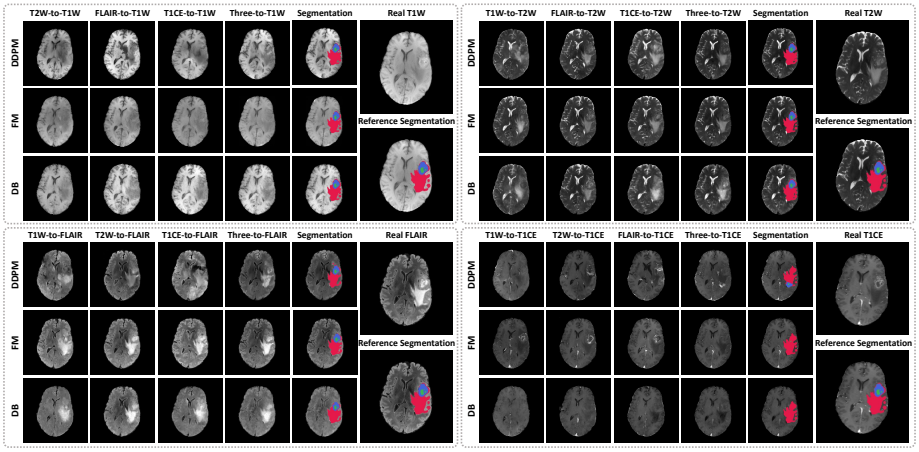

Fig. 1. Qualitative comparison of one-to-one and three-to-one cross-modality MRI synthesis with corresponding tumor segmentation consistency. In the three-to-one setting, segmentation consistency is evaluated by replacing one real modality with its synthesized counterpart while keeping the other three modalities real.

T1W and T2W synthesis generally achieves higher PSNR and SSIM, whereas T1CE synthesis remains more challenging, especially in terms of tumor-structure consistency for the TC and ET regions. This is reasonable because contrast enhancement is spatially localized and highly dependent on subtle tumor-related intensity patterns. Overall, FM achieves the strongest performance across most image-fidelity and tumor-structure metrics, while DB consistently ranks second in most cases.

Figure 1 presents representative qualitative results for cross-modality MRI synthesis under both one-to-one and three-to-one settings. For T1W, T2W, and FLAIR, the synthesized images show clear anatomical structures and tumor-related contrast, and the corresponding segmentation results show good spatial consistency with the reference segmentation masks obtained from real MRI inputs. In contrast, T1CE synthesis remains more challenging. The synthesized T1CE images exhibit relatively weaker enhancement patterns and less accurate tumor boundary preservation, which further leads to less satisfactory segmentation consistency. These observations suggest that most target modalities can be reliably synthesized within the proposed framework, whereas contrast-enhanced T1CE synthesis still requires further improvement.

4 Conclusion

In this work, we propose nnDiffusion, a standardized 3D diffusion framework for medical image modality synthesis. By reformulating the nnU-Net pipeline as a conditional generative framework, nnDiffusion extends a mature and well-validated medical imaging system to multimodal MRI synthesis within modern

diffusion paradigms. Experiments on the BraTS dataset demonstrate its flexibility across both one-to-one and three-to-one synthesis settings. By leveraging the robust preprocessing, architecture configuration, and optimization heuristics of nnU-Net, nnDiffusion enhances reproducibility and facilitates practical deployment. In future work, nnDiffusion can be further extended to broader medical imaging modalities, more diverse clinical scenarios, and advanced generative paradigms, providing a flexible foundation for reliable medical image synthesis.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (62576034) and the Fundamental Research Funds for the Central Universities (2025CX01009).

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

References

1. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
2. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data **4**, 170117 (2017)
3. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. In: Advances in Neural Information Processing Systems. vol. 34, pp. 8780–8794 (2021)
4. Dorent, R., Joutard, S., Modat, M., Ourselin, S., Vercauteren, T.: Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 74–82. Springer (2019)
5. Guo, P., Zhao, C., Yang, D., Xu, Z., Nath, V., Tang, Y., Simon, B., Belue, M., Harmon, S., Turkbey, B., et al.: MAISI: Medical AI for synthetic imaging. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4430–4441. IEEE (2025)
6. Hang, T., Gu, S., Li, C., Bao, J., Chen, D., Hu, H., Geng, X., Guo, B.: Efficient diffusion training via min-SNR weighting strategy. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7441–7451 (2023)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020)
8. Huijben, E.M., Terpstra, M.L., Pai, S., Thummerer, A., Koopmans, P., Afonso, M., Van Eijnatten, M., Gurney-Champion, O., Chen, Z., Zhang, Y., et al.: Generating synthetic computed tomography for radiotherapy: SynthRAD2023 challenge report. Medical Image Analysis **97**, 103276 (2024)
9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)

10. Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D.: Diffusion models in medical imaging: A comprehensive survey. Medical Image Analysis **88**, 102846 (2023)
11. Kläser, K., Markiewicz, P., Ranzini, M., Li, W., Modat, M., Hutton, B.F., Atkinson, D., Thielemans, K., Cardoso, M.J., Ourselin, S.: Deep boosted regression for MR to CT synthesis. In: International Workshop on Simulation and Synthesis in Medical Imaging. pp. 61–70. Springer (2018)
12. Li, B., Xue, K., Liu, B., Lai, Y.K.: BBDM: Image-to-image translation with brownian bridge diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition. pp. 1952–1961 (2023)
13. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: International Conference on Learning Representations (2023)
14. Liu, G.H., Vahdat, A., Huang, D.A., Theodorou, E.A., Nie, W., Anandkumar, A.: I^2SB: Image-to-image schrödinger bridge. In: Proceedings of the 40th International Conference on Machine Learning. pp. 22042–22062 (2023)
15. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: International Conference on Learning Representations (2023)
16. Longuefosse, A., Bot, E.L., De Senneville, B.D., Giraud, R., Mansencal, B., Coupé, P., Desbarats, P., Baldacci, F.: Adapted nnU-Net: a robust baseline for cross-modality synthesis and medical image inpainting. In: International Workshop on Simulation and Synthesis in Medical Imaging. pp. 24–33. Springer (2024)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
18. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2015)
19. Pang, H., Hong, X., Zhang, P., Ye, C.: Cascaded diffusion model and segment anything model for medical image synthesis via uncertainty-guided prompt generation. In: International Conference on Information Processing in Medical Imaging. pp. 203–217. Springer (2025)
20. Pang, H., Qi, S., Wu, Y., Wang, M., Li, C., Sun, Y., Qian, W., Tang, G., Xu, J., Liang, Z., et al.: NCCT-CECT image synthesizers and their application to pulmonary vessel segmentation. Computer Methods and Programs in Biomedicine **231**, 107389 (2023)
21. Pang, H., Xu, S., Zhang, X., An, F., Zhang, X., Wang, Y., Liu, F., Yan, T., Marais, P., Qian, Y., et al.: DRCAM: Dose reduction via contrast amplification modeling for brain contrast-enhanced magnetic resonance images. Biomedical Signal Processing and Control **113**, 109074 (2026)
22. Pang, H., Zhang, P., Hong, X., Chen, S., Ye, C.: D^3M: Deformation-driven diffusion model for synthesis of contrast-enhanced MRI with brain tumors. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 151–160. Springer (2025)
23. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: FiLM: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
24. 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)

25. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)
26. Zhang, T., Pang, H., Wu, Y., Xu, J., Liu, L., Li, S., Xia, S., Chen, R., Liang, Z., Qi, S.: BreathVisionNet: A pulmonary-function-guided CNN-transformer hybrid model for expiratory CT image synthesis. Computer Methods and Programs in Biomedicine **259**, 108516 (2025)
27. Zhao, C., Guo, P., Yang, D., Tang, Y., He, Y., Simon, B., Belue, M., Harmon, S., Turkbey, B., Xu, D.: MAISI-v2: Accelerated 3D high-resolution medical image synthesis with rectified flow and region-specific contrastive loss. Proceedings of the 40th AAAI Conference on Artificial Intelligence (2026)
28. Zhu, P., Fu, Y., Chen, N., Qiu, A.: Q-space guided multi-modal translation network for diffusion-weighted image synthesis. IEEE Transactions on Medical Imaging **45**(3), 1167–1178 (2025)
29. Zhu, P., Liu, C., Fu, Y., Chen, N., Qiu, A.: Cycle-conditional diffusion model for noise correction of diffusion-weighted images using unpaired data. Medical Image Analysis **103**, 103579 (2025)