RLBCD: Residual-guided Latent Brownian-bridge Co-Diffusion for Anatomical-to-Metabolic Image Synthesis

Tianxu Lv[1] , **Hongnian Tian**[1] , **Jiansong Fan**[1] , **Yuan Liu**[1] , **Lihua Li**[2] and **Xiang Pan**[1,3*]
[1]School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China
[2]Institute of Biomedical Engineering and Instrumentation, Hangzhou Dianzi University, Hangzhou 310018, China
[3]The PRC Ministry of Education Engineering Research Center of Intelligent Technology for Healthcare, Wuxi, Jiangsu 214122, China
xiangpan@jiangnan.edu.cn

Abstract

While metabolic imaging can facilitate early diagnosis by revealing physiological changes of lesions, it is limited by high cost, high radiation risk, and potential renal impairment. Thus, developing an effective approach for Anatomical-to-Metabolic Image Synthesis (A2MIS) is highly required. However, existing methods are heavily hindered by the gap between distinct domains, and fail to provide a confidence score for the synthesized images, severely restricting their clinical applications. Here, we propose a novel **R**esidual-guided **L**atent **B**rownian-bridge **C**o-**D**iffusion (**RLBCD**) model for A2MIS. Specifically, RLBCD starts with a co-diffusion process that leverages a residual diffusion branch to capture inter-domain differences, which are injected into an enhanced diffusion branch to maximally reconstruct modality-specific details. Furthermore, to explore desired residual guidance, we investigate the encoder and decoder features in diffusion models, and accordingly design a Hybrid-Granularity Fusion to integrate consistent semantics and complementary information for fine-grained reconstruction. Additionally, a latent consistency score is developed to enhance the restoration of modality-specific information, which also serves as an indicator of the inherent confidence of the synthesized images. Extensive experiments conducted on five public and in-house datasets demonstrate that RLBCD not only outperforms state-of-the-art methods for A2MIS, but also is valuable for downstream clinic applications.

1 Introduction

Metabolic imaging (*e.g.* computed tomography angiography (CTA) and positron emission tomography (PET)) provides essential insights into physiological and metabolic status of lesions, facilitating early diagnosis and treatment [Torigian *et al.*, 2007; Lv *et al.*, 2024]. However, compared to anatomical imaging, the acquisition of metabolic imaging is limited by high cost, slow scanning speed, high radiation risk, and potential renal impairment caused by contrast agents [Faucon *et al.*, 2019; Lv *et al.*, 2022]. Therefore, it is necessary to develop computational methods for automatic and effective Anatomical-to-Metabolic Image Synthesis (A2MIS).

Given that the A2MIS problem can be regarded as an image-to-image (I2I) task, various deep learning-based approaches, including Generative Adversarial Network (GAN)-based methods, diffusion model (DM)-based methods, and other techniques, have the potential to address this challenge. GAN-based methods [Goodfellow *et al.*, 2014], such as Pix2Pix [Isola *et al.*, 2017] and CycleGAN [Zhu *et al.*, 2017], have shown promising results by learning the conditional distribution of the metabolic images given the samples from the anatomical domain. However, GAN-based I2I methods are notoriously hard to train and often suffer from mode collapse in the output distribution. Other methods, such as Autoregressive Models [Parmar *et al.*, 2018], Variational Autoencoders (VAEs) [Vahdat and Kautz, 2020], and Normalizing Flows [Kingma and Dhariwal, 2018], have achieved success in specific applications but fail to acquire the same level of sample quality and general applicability as GANs. Recently, DM [Ho *et al.*, 2020; Song *et al.*, 2021] has emerged as a competitive alternative, showing the ability to generate high-quality images compared to GAN-based models. Several conditional diffusion models (CDMs) [Batzolis *et al.*, 2021; Saharia *et al.*, 2022; Rombach *et al.*, 2022] have proposed for the I2I task by integrating the source image into the reverse diffusion process to guide generation toward the target domain. However, CDMs struggle to generate consistent results due to their inherent stochasticity. Even when employing deterministic samplers like DDIM, their reliance on sampling from random noise introduces uncertainty, undermining reproducibility and medical reliability.

In Particular, the aforementioned methods face several challenges when directly applied to A2MIS. First, anatomical and metabolic images often exhibit similarities in shape (structure) and style (brightness and contrast) at certain locations. This resemblance presents a significant challenge for existing I2I methods, as they struggle to effectively capture the differences between the two domains, leading to synthesized images that do not accurately reflect the intended

*Corresponding Author.

metabolic characteristics. Second, DM-based methods can sample diverse results for a given anatomical image. It is essential to provide a confidence score associated with these synthesized images for physicians and patients, such as AlphaFold [Jumper *et al.*, 2021; Abramson *et al.*, 2024]. This knowledge enables professionals to make informed decisions. Unfortunately, existing methods primarily overlook this crucial point. Thus, a core question for A2MIS is how to optimally decouple and leverage anatomical and metabolic features while incorporating a synthesis confidence score.

To address these challenges, we propose a novel **R**esidual-guided **L**atent **B**rownian-bridge **C**o-**D**iffusion (**RLBCD**) model for precise and reliable A2MIS. Specifically, RLBCD constructs a co-diffusion process that leverages a latent residual diffusion branch to capture inter-domain differences, which are then injected into an latent enhanced diffusion branch to optimally reconstruct modality-specific details. Our method employs a stochastic Brownian bridge process [Li *et al.*, 2023a] that directly learns translation between two domains without any conditioning mechanism. To achieve the desired residual guidance, we investigate the encoder and decoder features in diffusion models and accordingly propose a hybrid-granularity fusion, which integrates consistent semantics and complementary information for fine-grained reconstruction. Moreover, we develop a latent consistency score to improve the restoration of modality-specific information, which also serves as an indicator of the inherent confidence of the synthesized images. We conduct extensive experiments on various public and in-house datasets and tasks, including CT-to-CTA, CT-to-PET, and pre-contrast MRI-to-post-contrast MRI generation, as well as downstream diagnosis and segmentation tasks. Experimental results demonstrate the superiority of our approach not only performance but also valuable medical applications. Specific contributions of this work can be summarized as follows:

- We propose RLBCD, a novel framework for A2MIS, leveraging a designed residual-guided latent Brownian-bridge co-diffusion process for the optimal reconstruction of modality-specific details.
- We propose Hybrid-Granularity Fusion to integrate consistent semantics and complementary information from a residual diffusion for fine-grained generation.
- We propose latent consistency score, serving as an indicator of the inherent confidence of synthesized images.
- Extensive experiments demonstrate the superiority of our approach in terms of translation performance and its value in downstream diagnosis and segmentation tasks.

2 Related Work

2.1 Anatomical-to-Metabolic Image Synthesis

Metabolic images, compared to anatomical images, provide a different perspective on the body's physiological functioning, thereby facilitating early disease detection and monitoring treatment efficacy [Torigian *et al.*, 2007]. However, metabolic imaging presents several obvious limitations, including high cost, significant radiation exposure, and potential risks to renal function. Several approaches have been proposed for A2MIS, including 1) Synthesis of PET from MRI or CT [Hu *et al.*, 2021; Lee *et al.*, 2024; Vega *et al.*, 2024]; 2) Synthesis of CTA from CT [Lyu *et al.*, 2023]; 3) Synthesis of Post-contrast CT from Pre-contrast CT [Kim *et al.*, 2021; Choi *et al.*, 2021]; and 4) Synthesis of Post-contrast MRI from Pre-contrast MRI [Calabrese *et al.*, 2021; Wang *et al.*, 2022]. Current studies mainly utilize or modify existing I2I methods, overlooking the unique characteristics inherent to A2MIS, such as the relations between anatomical and metabolic images, and the requirement for confidence scores of the synthesized images. This confidence assessment is crucial for clinical applications, as it provides a measure of the reliability of the synthesized images. Thus, our objective is to develop an approach for precise and reliable A2MIS.

2.2 Diffusion Models in Image Translation

Diffusion models (DMs) are initially developed for image generation such as DDPM [Ho *et al.*, 2020] and DDIM [Song *et al.*, 2021]. Subsequent image translation methods [Batzolis *et al.*, 2021; Saharia *et al.*, 2022; Rombach *et al.*, 2022] based on DMs treat the I2I task as conditional image generation, guiding the diffusion towards the target domain by feeding the source image or its encoded feature into the denoising U-Net during the reverse process. Despite achieving some practical success, this condition mechanism lacks a clear theoretical guarantee that the final diffusion result will yield the desired conditional distribution. Recently, a Brownian bridge stochastic process-based diffusion model termed BBDM [Li *et al.*, 2023a], which directly learns the translation between two domains through a bidirectional diffusion process, provides a promising tool for effective I2I. However, it fails to capture discriminative features between source and target domains, especially when they exhibit some similarities in shape and style. Thus, it is important to enhance the model's understanding of the discriminative features between domains while maintaining its ability to capture semantic content of source images. Motivated by this, we present a residual-guided latent Brownian-bridge co-diffusion process to capture inter-domain differences for optimal reconstruction of modality-specific details.

2.3 Encoder-decoder Features in Diffusion Models

Recent studies in exploring the possibility of using DM for representation learning [Preechakul *et al.*, 2022] have demonstrated the encoder-decoder features in denoising networks *e.g.* U-Net, for downstream semantic tasks, such as classification [Sigger *et al.*, 2024], segmentation [Baranchuk *et al.*, 2021; Lv *et al.*, 2023], and fusion[Yang *et al.*, 2025]. More recently, several studies show that the encoder features in denoising U-Net change minimally, whereas the decoder features exhibit substantial variations across different time steps [Li *et al.*, 2023b; Ma *et al.*, 2024]. Inspired by these works, we investigate the encoder-decoder features in DMs and accordingly propose a Hybrid-Granularity Fusion to integrate consistent semantics and complementary information from a residual diffusion for fine-grained generation.

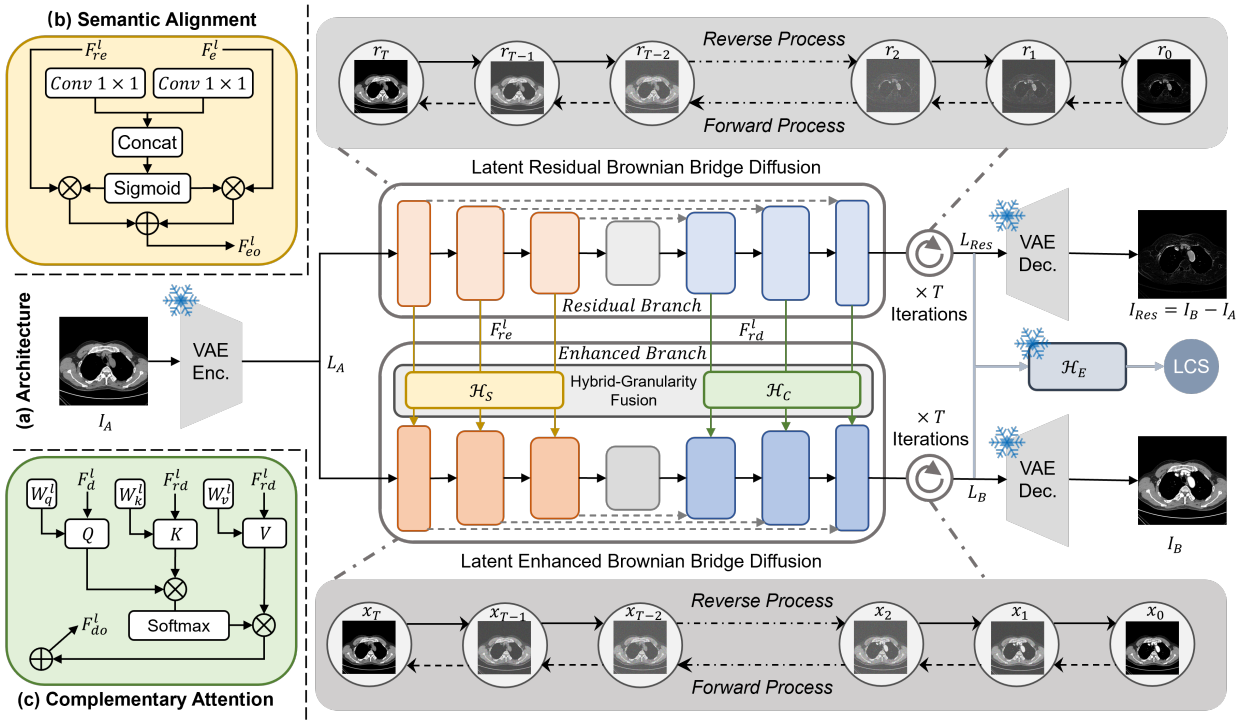

Figure 1: **a**) Architecture of RLBCD, composed of a latent co-diffusion process (*i.e.* a residual Brownian-bridge diffusion to capture inter-domain differences and an enhanced Brownian-bridge diffusion for precise A2MIS), a Hybrid-Granularity Fusion (HGF) including a Semantic Alignment network $\mathcal{H}_S$ (**b**) and a Complementary Attention network $\mathcal{H}_C$ (**c**) to integrate consistent semantics and complementary information from the residual diffusion, as well as a latent evaluation network $\mathcal{H}_E$ to evaluate the confidence of synthesized images.

3 Methodology

3.1 Problem Formulation

Given an anatomical image I_A and a metabolic image I_B from domain A and B, the inter-domain difference can be calculated as $I_{Res} = I_B - I_A$, and A2MIS aims to learn a mapping from domain A to domain B. The key to our method is the infusion of residual information by transforming I_A to I_{Res}. To improve the learning efficiency and model generalization, the diffusion process is conducted in a latent space. Given an image I_A sampled from domain A, we first extract its latent feature L_A using a VAE Encoder [Esser *et al.*, 2021] that maps between raw-voxel space and low-dimensional latent space, followed by a co-diffusion process to obtain the corresponding latent representation L_B and L_{Res} in domain B and B-A. The final translated images I_B and I_{Res} can be generated by a pre-trained VAE decoder [Esser *et al.*, 2021].

3.2 Overview

Fig. 1(a) shows the architecture of RLBCD, which consists of a latent residual-guided co-diffusion process designed to capture inter-domain differences and maximally reconstruct modality-specific details, a Hybrid-Granularity Fusion composed of a Semantic Alignment network $\mathcal{H}_S$ (b) and a Complementary Attention network $\mathcal{H}_C$ (c) to integrate consistent semantics and complementary information from the residual diffusion, as well as a latent evaluation network $\mathcal{H}_E$ to assess the confidence of synthesized images.

3.3 Residual-guided Co-Diffusion

We take similar notations as DDPM, and let $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{r})$ represent the paired data from domains A, B, and Res. To speed up the training and inference process, we conduct diffusion process in the latent space of VQGAN [Esser *et al.*, 2021]. For simplicity, we use $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{r}$ to denote the corresponding latent features ($\boldsymbol{x} := L_A(\boldsymbol{x}), \boldsymbol{y} := L_B(\boldsymbol{y}), \boldsymbol{r} := L_{Res}(\boldsymbol{r})$).

Residual Branch. To effectively capture inter-domain differences, we employ a Brownian bridge diffusion that directly learns the translation from the source domain to the residual domain. Unlike DDPM that conclude at Gaussian noise $\boldsymbol{x}_T \sim \mathcal{N}(0, \mathbf{I})$, Brownian bridge diffusion assumes that both endpoints of the diffusion process as fixed data points from an arbitrary joint distribution, *i.e.* $(\boldsymbol{x}_T, \boldsymbol{x}_0) \sim q_{data}(\boldsymbol{x}, \boldsymbol{r})$. The forward process of the Brownian bridge forms a bridge between two fixed endpoints at $t = 0$ and T:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0, \boldsymbol{r}) = \mathcal{N}(\boldsymbol{x}_t; (1 - m_t)\boldsymbol{x}_0 + m_t\boldsymbol{r}, \delta_t\mathbf{I}) \quad (1)$$

where $\boldsymbol{x}_0 = \boldsymbol{x}$, $\boldsymbol{x}_t = \boldsymbol{r}$, $m_t = t/T$ and the variance term $\delta_t = 2(m_t - m_t^2)$. The reverse process of residual diffusion aims to predict $\boldsymbol{x}_{t-1}$ based on $\boldsymbol{x}_t$:

$$p_{\theta_r}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{r}) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{\theta_r}(\boldsymbol{x}_t, t)\widetilde{\delta}_t\mathbf{I}) \quad (2)$$

where $\widetilde{\delta}_t$ is the variance of Gaussian noise at step t and $\boldsymbol{\mu}_{\theta_r}(\boldsymbol{x}_t, t)$ is the predicted mean value of the noise to be learned. The training objective of Residual Branch is optimizing the Evidence Lower Bound (ELBO):

$$\mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{r}, \boldsymbol{\epsilon}}[c_{\epsilon t} \parallel m_t(\boldsymbol{r} - \boldsymbol{x}_0) + \sqrt{\delta}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\theta_r}(\boldsymbol{x}_t, t) \parallel^2] \quad (3)$$

where $c_{\epsilon t}$ denotes the coefficient term of the estimated noise $\boldsymbol{\epsilon}_{\theta_r}$ in mean value term $\widetilde{\boldsymbol{\mu}}_t$. As thus, we build a map from the source domain A to the residual domain Res, which can provide a guidance for the following enhanced branch.

Enhanced Branch is designed to build a mapping from the source domain A to the target domain B by integrating complementary information from the residual branch. To maintain consistency with the residual branch, the enhanced branch uses a Brownian bridge diffusion process that learns to establish the bridge from a fixed initial point $\boldsymbol{x}_T = \boldsymbol{x}$ to the target $\boldsymbol{x}_0 = \boldsymbol{y}$. Formally, the forward enhanced diffusion process is defined as follows:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}_t; (1 - m_t)\boldsymbol{x}_0 + m_t\boldsymbol{y}, \delta_t\mathbf{I}) \quad (4)$$

where $\boldsymbol{x}_0 = \boldsymbol{x}$, $\boldsymbol{x}_t = \boldsymbol{y}$, $m_t = t/T$ and the variance term $\delta_t = 2(m_t - m_t^2)$. Unlike conventional diffusion processes, the reverse process of the enhanced diffusion aims to predict $\boldsymbol{x}_{t-1}$ based on $\boldsymbol{x}_t$ and the estimated residual information $\boldsymbol{r}_t$ from the residual branch:

$$p_{\theta_e}(\boldsymbol{x}_{t-1}|(\boldsymbol{x}_t, \boldsymbol{r}_t), \boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{\theta_e}(\boldsymbol{x}_t, \boldsymbol{r}_t, t)\widetilde{\delta}_t\mathbf{I}) \quad (5)$$

where $\boldsymbol{\mu}_{\theta_e}(\boldsymbol{x}_t, \boldsymbol{r}_t, t)$ is the learned mean value of the noise. In particular, we employ an enhanced noise predictor network ($\boldsymbol{\epsilon}_{\theta_e}(\boldsymbol{x_t}, \boldsymbol{r_t}, t)$) to predict the noise component at the step t:

$$\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{y},\boldsymbol{\epsilon}}[c_{\epsilon t} \parallel m_t(\boldsymbol{y} - \boldsymbol{x}_0) + \sqrt{\delta}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\theta_e}(\boldsymbol{x}_t, \boldsymbol{r}_t, t) \parallel^2] \quad (6)$$

3.4 Hybrid-Granularity Fusion

As aforementioned, we establish a bridge between the co-diffusion branches, and the information from the residual branch is injected into the enhanced branch to guide the A2MIS process. However, despite the abundant semantic information contained in the encoder-decoder features of denoising models, its application in optimizing I2I has not been fully explored, especially for the Brownian bridge diffusion. Thus, we investigate the encoder-decoder features within denoising models in the Brownian bridge diffusion process. For clarity, we implement diffusion in the voxel space ranter than latent space. As shown in Fig. 2, for the same source image, the encoder features are similar whether diffusing to the residual image or the target image, mainly capturing the semantic information of the source image. However, we observe significant differences in the decoder features, which are primarily related to the translated images. Building on the observations, we propose the Hybrid-Granularity Fusion, including a Semantic Alignment module to strengthen semantic representations and a Complementary Attention mechanism to integrate complementary features.

Semantic Alignment. To effectively enhance the semantic representation, we introduce the Semantic Alignment (SAM) that is embedded in different scales ($l \in \{1, 2, 3\}$) of the down-sampling blocks in $\boldsymbol{\epsilon}_{\theta_e}$, where we refer F_e^l and F_{eo}^l to its input and output. Specifically, the SAM is designed to refine the learned representations by aligning and integrating semantic information from the residual branch F_{re}^l:

$$AliCof = \sigma(P(\mathcal{C}_3^l(\mathcal{C}_1^l(F_{re}^l) \oplus \mathcal{C}_2^l(F_e^l)))) \quad (7)$$

$$F_{eo}^l = F_e^l AliCof + F_{re}^l(1 - AliCof) \quad (8)$$

where $\mathcal{C}_1^l$, $\mathcal{C}_2^l$, $\mathcal{C}_3^l$ are 1×1 convolutions for the l-th scale down-sampling block and P is the adaptive average pooling.

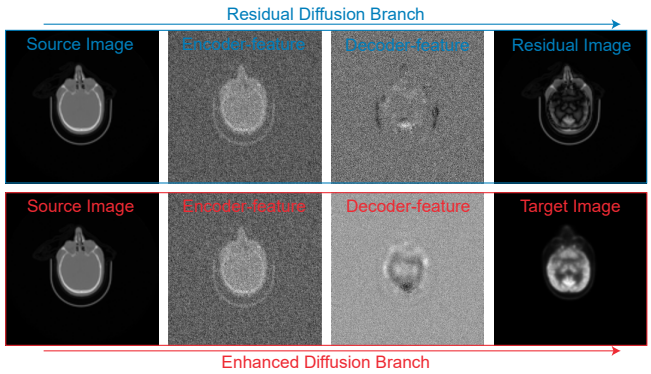

Figure 2: Visualization of encoder-decoder features of the denoising model during the Brownian bridge diffusion process, with the transformation from the same source image to the residual image (top) and the target image (bottom).

Complementary Attention. To adaptively incorporate the residual details from the decoder features in the residual branch, we introduce the Complementary Attention (CMA) that is embedded in different scales ($l \in \{1, 2, 3\}$) of up-sampling blocks in $\boldsymbol{\epsilon}_{\theta_e}$, where we refer F_d^l and F_{do}^l to its input and output. Given the multi-scale residual feature maps F_{re}^l from ϵ_{θ_r}, the CMA module aims to compensate for the missing necessary details. Specifically, we propose to inject multi-scale residual details by biasing the queries of cross-attention n the up-sampling blocks as shown in Fig. 1(c), *i.e.*,

$$Q = W_q^l F_d^l,\ K = W_k^l F_{rd}^l,\ V = W_v^l F_{rd}^l \quad (9)$$

$$A = \text{softmax}(\frac{QK^T}{\sqrt{d}}),\ F_{do} = AV \quad (10)$$

where W_q^l, W_k^l, W_v^l are specific projection layers for the l-th scale up-sampling block of dimension d.

3.5 Latent Consistency Score

To enhance the credibility of synthesized images, we propose a self-constraint latent consistency score (LCS). It does not require target images, allowing it to serve as an indicator for sampled images. Specifically, we consider that the latent features L_{Res} and L_B can be effectively fused to generate L_A. Thus, we pre-train a sample convolution-based fusion model that takes L_{Res} and L_B as inputs and outputs the corresponding L_A. We first utilize a paired datasets to pre-train this fusion network, enabling that it accurately fuses L_B and L_{Res} to generate L_A. After pre-training, we incorporate it into our proposed RLBCD as a latent evaluation network $\mathcal{H}_E$. It receives estimated representations $\hat{L}_B$ and $\hat{L}_{Res}$ to predict the fused representation $\hat{L}_A$. During the training stage of RLBCD, we impose a constraint on the similarity between the predicted $\hat{L}_A$ and the true latent representation L_A as an auxiliary loss::

$$\mathbb{E}_{L_A,\hat{L}_A}[\parallel L_A - \hat{L}_A \parallel^2] \quad (11)$$

Besides, during the generation phase, we can evaluate the reliability of the synthesized images by assessing the similarity to select the best-performing one.

4 Experiments

4.1 Datasets

To comprehensively validate our proposed method, we conduct extensive experiments on five datasets involving various anatomical-metabolic modality translations. Initially, we conduct experiments on three datasets, including two public datasets, *i.e.* Head-Neck-PET-CT [Vallières *et al.*, 2017] and Duke-Breast-Cancer-MRI [Saha *et al.*, 2021], as well as an in-house dataset, *i.e.* Chest-CTA. Head-Neck-PET-CT contains a total of 93 patients with the FDG-PET/CT scan. Each PET/CT volume consists of 119 slices and the size of each slice is 128×128. Duke-Breast-Cancer-MRI contains a total of 922 patients with the contrast-enhanced MRI scan. Each MR volume consists of 60 slices and the size of each slice is 256×256. Chest-CTA contains a total of 114 patients with the CT/CTA scan. Each CT/CTA volume consists of 560 slices and the size of each slice is 256×256. The above three datasets are split in a 7:1:2 ratio based on case-level for training, validation and testing. Subsequently, we utilize two independent datasets to assess the models' generalization ability and downstream value, including a public dataset, *i.e.* Breast-MRI-NACT-PiloT [Newitt and Hylton, 2016], and an in-house dataset, *i.e.* HUASHANCT-PET. Breast-MRI-NACT-PiloT contains a total of 64 patients with the contrast-enhanced MRI protocol. Each MR volume consists of 60 slices and the size of each slice is 256×256. Ground truth segmentation of the data are provided in the dataset for tumor annotation. HUASHANCT-PET contains a total of 43 patients with the FDG-PET/CT scan and corresponding cardiac amyloidosis diagnostic labels. Each volume consists of 148 slices and the size of each slice is 128×128.

4.2 Baselines and Evaluation Metic

We compare our proposed method with six state-of-the-art A2MIS approaches, including Pix2Pix [Isola *et al.*, 2017], CycleGAN [Zhu *et al.*, 2017], VQI2I [Chen *et al.*, 2022], QS-Attn [Hu *et al.*, 2022], BBDM [Li *et al.*, 2023a], and UNSB [Kim *et al.*, 2024]. All baselines are trained using paired anatomical-metabolic images for a fair comparison. We comprehensively evaluated the experimental results from both the synthesis and downstream segmentation and diagnosis. In the synthesis evaluation, we evaluate the quality of translated images using Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM). As for the segmentation evaluation, we assess the cancer segmentation value of the translated images through a unified segmentation model, *i.e.* U-Net, and quantify the segmentation results using Dice Similarity Coefficient (DSC) and Jaccard Index (JI). As for the diagnosis evaluation, we assess the cardiac amyloidosis prediction value of the translated images through a unified classification model, *i.e.* ResNet, and quantify the classification results using accuracy (ACC) and the area under the curve (AUC).

4.3 Implementation Details

The developed model is implemented using Pytorch and the experiments in this study are executed on a platform comprising four NVIDIA RTX A6000 GPUs to accelerate the training process. We first pre-train VQGAN [Esser *et al.*, 2021] with downsampling factor of 8 using the collected datasets. The number of time steps of Brownian bridge diffusion is set to be 1000 during the training stage, and then we employ 200 sampling steps during the sample stage with the considerations of both sample quality and efficiency following [Li *et al.*, 2023a]. All models are first trained and evaluated on the Head-Neck-PET-CT, Duke-Breast-Cancer-MRI, and Chest-CTA datasets. Then we directly assess the performance of the models trained on the Duke-Breast-Cancer-MRI dataset using the Breast-MRI-NACT-PiloT dataset, and we evaluate the performance of the models trained on the Head-Neck-PET-CT dataset using the HUASHANCT-PET dataset to verify their generalization ability. For the downstream diagnosis and segmentation tasks, we use a five-fold cross-validation strategy and the mean scores of results are presented.

4.4 Comparison with Baselines

Table 1 and Fig. 3 report the performances of our proposed method and state-of-the-art approaches on the Head-Nect-PET-CT, Duke-Breast-Cancer-MRI, and Chest-CTA datasets for various anatomical-metabolic modality translations. The best score in each column is in **bold** and the second best is underlined. Experimental results demonstrate that the proposed RLBCD comprehensively outperforms other methods on all three datasets across all evaluation metrics. This result verifies the effectiveness of our framework for diverse A2MIS, including translations from CT to PET, pre-contrast MRI post-contrast MRI, and CT to CTA. In three different datasets, our methods achieves SSIM of 5.20%, 5.90%, and 4.52% gain over the best baseline model. This signifies that our RLBCD is capable of integrating critical complementary information in residual branch that other models may overlook. Among all baselines, CycleGAN and BBDM achieves better performance on different datasets, indicating the effectiveness of GAN-based and Diffusion-based approaches. In comparison, the proposed method utilizes a co-diffusion process with a Hybrid-Granularity Fusion to capture consistent semantics and inter-domain differences for the maximal reconstruction of modality-specific details.

4.5 Ablation Study

To explore the impact of each designed component in the proposed method, we conduct an extensive ablation analysis by evaluating different RLBCD variants as follows:

- **RLBCD without the Semantic Alignment (w/o SAM)**: This variant replace the SAM with a direct combination of two encoder features in denoising models.
- **RLBCD without the Complementary Attention (w/o CMA)**: This variant replace the CMA with a direct combination of two decoder features in denoising models.
- **RLBCD without the Hybrid-Granularity Fusion (w/o HGF)**: This variant replace the HGF with a direct combination of encoder-decoder features.
- **RLBCD without the Latent Consistency Score in training (w/o LCS)**: This variant removes the auxiliary latent consistency loss during the training phase.

	Head-Neck-PET-CT CT → PET			Duke-Breast-Cancer-MRI Pre-contrast MRI → Post-contrast MRI			Chest-CTA CT → CTA		
	MAE (Voxel)↓	PSNR (dB)↑	SSIM (%)↑	MAE (Voxel)↓	PSNR (dB)↑	SSIM (%)↑	MAE (Voxel)↓	PSNR (dB)↑	SSIM (%)↑
Pix2Pix	11.42±2.97	19.42±1.87	68.00±5.55	6.73±1.35	23.86±1.53	70.88±4.47	9.74±2.13	21.22±1.99	76.60±4.71
CycleGAN	8.42±4.66	22.06±4.86	65.82±8.64	6.91±1.20	24.42±1.46	65.39±4.84	8.57±2.11	22.56±2.29	81.34±4.82
VQI2I	9.10±4.70	21.90±3.21	58.81±5.88	7.32±1.39	24.50±1.27	68.51±4.93	18.95±2.29	16.06±0.85	58.63±3.71
QS-Attn	10.22±2.70	19.48±1.10	48.49±7.07	9.29±1.43	24.11±1.65	42.69±4.65	10.66±3.33	20.99±2.36	80.14±5.30
BBDM	3.85±2.19	23.91±3.68	86.06±6.98	5.73±1.11	25.65±1.41	75.55±4.65	12.70±1.95	19.31±1.10	72.80±4.61
UNSB	6.09±2.70	23.20±1.10	84.89±7.07	7.68±2.12	23.54±2.16	71.98±6.05	10.00±3.90	21.33±3.26	78.78±6.71
RLBCD	**2.97±1.23**	**25.97±3.77**	**91.26±4.67**	**3.59±1.12**	**27.14±2.35**	**81.45±5.41**	**6.51±0.98**	**24.92±0.85**	**85.86±3.68**

Table 1: Comparison of RLBCD and baselines for A2MIS prediction on two public and one in-house datasets. (Mean ± Std)

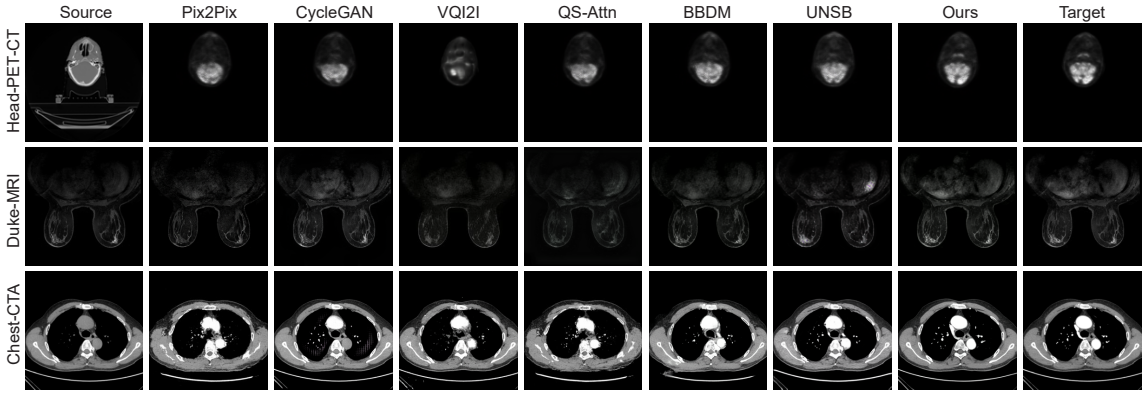

Figure 3: Qualitative comparison with state-of-the-art baselines on three datasets with various A2MIS tasks.

	Head-Neck-PET-CT CT → PET			Chest-CT-CTA CT → CTA		
	MAE	PSNR	SSIM	MAE	PSNR	SSIM
RLBCD	1.97	25.97	91.26	6.51	24.92	85.86
w/o SAM	2.21	25.45	90.02	6.85	24.51	85.47
w/o CMA	2.88	24.85	88.14	8.47	23.08	83.78
w/o HGF	3.20	24.56	87.49	9.28	21.50	80.26
w/o LCS	2.36	25.18	89.55	6.68	24.56	84.99
w/o RDB	3.85	23.91	86.06	12.70	19.31	72.80
w/o EDB	4.01	23.51	85.23	11.83	19.88	73.56

Table 2: Ablation study of designed components in RLBCD.

- **RLBCD without the Residual Branch (w/o RDB)**: This variant removes the residual branch, representing the original Brownian bridge diffusion model (BBDM).
- **RLBCD without the Enhanced Branch (w/o EDB)**: This variant removes the enhanced branch and aims to predict the residuals of two images for translation.

Table 2 reports the performance of RLBCD and its six variant s on Head-Neck-PET-CT and Chest-CT-CTA datasets. It can be observed that all variants of RLBCD produce the decreased performance, demonstrating that all components contribute to A2MIS.

4.6 Influence of Latent Consistency Score

We investigate the relationship between the latent consistency score (LCS) and the quality of generated images. For each anatomical image, we employ RLBCD to generate ten corresponding metabolic images. We then categorized these images into five groups based on the normalized LCS values and calculate the SSIM metric for each group. As illustrated in Fig. 4, as the LCS increases, the quality of the corresponding generated images also improves. The findings underscore the importance of the LCS as a valuable metric for evaluating the quality of generated samples, providing insights into the effectiveness of our generative models.

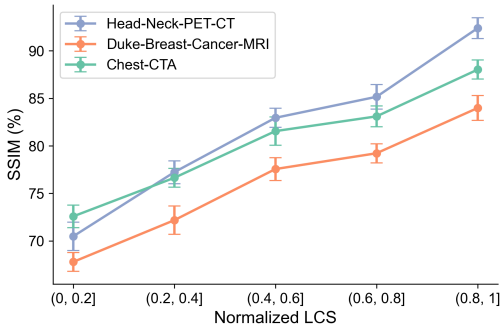

Figure 4: Relation between LCS and the quality of sampled images.

4.7 External Validation

Medical Images typically exhibit diversity across different data sources due to imaging protocols, patient demographics, and the variability in disease presentation. To assess the generalization ability of the models, we conduct external validation using two independent datasets. Specifically, we directly apply the models trained on the Duke-Breast-Cancer-MRI dataset to test its performance on the Breast-MRI-NACT-PiloT dataset for synthesizing post-contrast MRI from pre-

	HUASHANCT-PET CT → PET		
	MAE (Voxel) ↓	PSNR (dB) ↑	SSIM (%) ↑
Pix2Pix	13.15	18.35	60.82
CycleGAN	15.16	17.81	73.02
VQI2I	11.54	19.55	48.75
QS-Attn	15.98	18.28	53.65
BBDM	10.43	19.16	67.49
UNSB	9.23	20.10	69.71
RLBCD	**6.53**	**22.15**	**78.85**

Table 3: Comparison of RLBCD and baselines for PET generation from CT on an external dataset.

	Breast-MRI-NACT-PiloT Pre-MRI → Post-MRI		
	MAE (Voxel) ↓	PSNR (dB) ↑	SSIM (%) ↑
Pix2Pix	15.89	20.41	28.98
CycleGAN	11.59	23.69	43.72
VQI2I	17.12	19.41	33.97
QS-Attn	16.88	18.94	55.58
BBDM	12.82	22.83	36.45
UNSB	14.20	21.48	41.76
RLBCD	**8.12**	**24.58**	**60.23**

Table 4: Comparison of RLBCD and baselines for post-contrast MRI generation from pre-contrast MRI on an external dataset.

contrast MRI. Meanwhile, we utilize the models trained on the Head-Neck-PET-CT dataset to test its performance on the HUASHANCT-PET dataset for synthesizing PET from CT. As reported in Tables 3,4, all methods exhibit decreased results in external testing. Despite this, our proposed RLBCD still achieves the best performance across both independent test sets. This indicates that our proposed method maintain a higher level of robustness and adaptability compared to baselines by effectively capturing inter-domain differences.

4.8 Diagnostic Value of Synthetic Images

We further evaluate the diagnostic value of the generated images by diagnosing the status of cardiac amyloidosis on the HUASHANCT-PET dataset. Specifically, we first utilize the generative models to generate PET from CT. Then both the generated PET and the original CT are fed into a ResNet network for the training and testing of cardiac amyloidosis diagnosis. Classification results are presented in Table 5. When using only CT for diagnosis, the AUC and ACC are 43.07% and 69.88%, respectively. In comparison, predictions using the generated PET by each model demonstrate an improvement in diagnosis performance. This enhancement highlights the significance of the A2MIS task, which effectively capture physiological changes of lesions from anatomical images to facilitate early diagnosis. Besides, among all various methods evaluated, our proposed method achieves the superior performance, further verifying its effectiveness in clinical practice.

4.9 Annotation Value of Synthetic Images

We also evaluate the annotation value of the synthesized images by segmenting breast tumors in the Breast-MRI-NACT-PiloT dataset. Specifically, we employ generative models to generate post-contrast MRI from pre-contrast MRI. Subsequently, both the generated post-contrast MRI and the original pre-contrast MRI are fed into a U-Net network for the training and testing of tumor annotation. The segmentation results are reported in Table 6. When using only pre-contrast MRI for tumor segmentation, the DSC and JI metrics are 49.56% and 32.94%, respectively. In comparison, predictions made using the generated post-contrast MRI from each model exhibit an improvement in segmentation performance. This improvement highlights the effectiveness of A2MIS, which also provide valuable information for voxel-level tasks. Moreover, our RLBCD achieves the best performance among all competing baselines, indicating its superiority in facilitating accurate tumor annotation.

	HUASHANCT-PET Cardiac Amyloidosis Classification	
	AUC (%) ↑	ACC (%) ↑
I_A	43.07	69.88
I_A+$\hat{I}_B$ (Pix2Pix)	66.82 (+23.75)	73.17 (+3.29)
I_A+$\hat{I}_B$ (CycleGAN)	49.01 (+5.94)	71.72 (+1.84)
I_A+$\hat{I}_B$ (VQI2I)	56.56 (+13.49)	73.26 (+3.38)
I_A+$\hat{I}_B$ (QS-Attn)	53.52 (+10.45)	72.78 (+2.90)
I_A+$\hat{I}_B$ (BBDM)	62.18 (+19.11)	72.89 (+3.01)
I_A+$\hat{I}_B$ (UNSB)	57.38 (+14.31)	71.43 (+1.55)
I_A+$\hat{I}_B$ (RLBCD)	**69.45 (+26.38)**	**75.02 (+5.14)**

Table 5: Comparison of RLBCD and baselines for disease classification on an external dataset using synthesized images.

	Breast-MRI-NACT-PiloT Breast Cancer Segmentation	
	DSC (%) ↑	JI (%) ↑
I_A	49.56	32.94
I_A+$\hat{I}_B$ (Pix2Pix)	54.84 (+5.28)	37.78 (+4.84)
I_A+$\hat{I}_B$ (CycleGAN)	55.48 (+5.92)	38.39 (+5.45)
I_A+$\hat{I}_B$ (VQI2I)	54.47 (+4.91)	37.43 (+4.49)
I_A+$\hat{I}_B$ (QS-Attn)	51.63 (+2.07)	34.79 (+1.85)
I_A+$\hat{I}_B$ (BBDM)	52.28 (+2.72)	36.19 (+3.25)
I_A+$\hat{I}_B$ (UNSB)	51.94 (+2.38)	35.08 (+2.14)
I_A+$\hat{I}_B$ (RLBCD)	**57.12 (+7.56)**	**39.08 (+6.14)**

Table 6: Comparison of RLBCD and baselines for breast cancer segmentation on an external dataset using synthesized images.

5 Conclusion

In this work, we propose a residual-guided latent Brownian-bridge co-diffusion network (RLBCD) for A2MIS. By leveraging a residual diffusion branch to capture inter-domain differences, RLBCD is able to maximally reconstruct modality-specific details. A Hybrid-Granularity Fusion is embedded in RLBCD to enhance consistent semantics and complementary information. Besides, we devise LCS that severs as an indicator of the quality of the synthesized images. Comprehensive experiments demonstrate that RLBCD not only surpasses the performance of recent state-of-the-art methods on various datasets and tasks, but provides valuable information for downstream diagnosis and segmentation applications.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under grants W2411054, U21A20521 and 62271178, the Postgraduate Research & Practice Innovation Program of Jiangsu Province KYCX23_2524, National Foreign Expert Project of China under Grant G2023144009L, Zhejiang Provincial Natural Science Foundation of China (LR23F010002), Wuxi Health Commission Precision Medicine Project (J202106), Jiangsu Provincial Six Talent Peaks Project (YY-124), and the construction project of Shanghai Key Laboratory of Molecular Imaging (18DZ2260400).

References

[Abramson *et al.*, 2024] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

[Baranchuk *et al.*, 2021] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

[Batzolis *et al.*, 2021] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021.

[Calabrese *et al.*, 2021] Evan Calabrese, Jeffrey D Rudie, Andreas M Rauschecker, Javier E Villanueva-Meyer, and Soonmee Cha. Feasibility of simulated postcontrast mri of glioblastomas and lower-grade gliomas by using three-dimensional fully convolutional neural networks. *Radiology: Artificial Intelligence*, 3(5):e200276, 2021.

[Chen *et al.*, 2022] Yu-Jie Chen, Shin-I Cheng, Wei-Chen Chiu, Hung-Yu Tseng, and Hsin-Ying Lee. Vector quantized image-to-image translation. In *European Conference on Computer Vision*, pages 440–456. Springer, 2022.

[Choi *et al.*, 2021] Jae Won Choi, Yeon Jin Cho, Ji Young Ha, Seul Bi Lee, Seunghyun Lee, Young Hun Choi, Jung-Eun Cheon, and Woo Sun Kim. Generating synthetic contrast enhancement from non-contrast chest computed tomography using a generative adversarial network. *Scientific reports*, 11(1):20403, 2021.

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

[Faucon *et al.*, 2019] Anne-Laure Faucon, Guillaume Bobrie, and Olivier Clément. Nephrotoxicity of iodinated contrast media: From pathophysiology to prevention strategies. *European Journal of Radiology*, 116:231–241, 2019.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

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

[Hu *et al.*, 2021] Shengye Hu, Baiying Lei, Shuqiang Wang, Yong Wang, Zhiguang Feng, and Yanyan Shen. Bidirectional mapping generative adversarial networks for brain mr to pet synthesis. *IEEE Transactions on Medical Imaging*, 41(1):145–157, 2021.

[Hu *et al.*, 2022] Xueqi Hu, Xinyue Zhou, Qiusheng Huang, Zhengyi Shi, Li Sun, and Qingli Li. Qs-attn: Query-selected attention for contrastive learning in i2i translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18291–18300, 2022.

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

[Jumper *et al.*, 2021] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[Kim *et al.*, 2021] Se Woo Kim, Jung Hoon Kim, Suha Kwak, Minkyo Seo, Changhyun Ryoo, Cheong-Il Shin, Siwon Jang, Jungheum Cho, Young-Hoon Kim, and Kyutae Jeon. The feasibility of deep learning-based synthetic contrast-enhanced ct from nonenhanced ct in emergency department patients with acute abdominal pain. *Scientific reports*, 11(1):20390, 2021.

[Kim *et al.*, 2024] Beomsu Kim, Gihyun Kwon, Kwanyoung Kim, and Jong Chul Ye. Unpaired image-to-image translation via neural schrödinger bridge. In *ICLR*, 2024.

[Kingma and Dhariwal, 2018] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

[Lee *et al.*, 2024] Jeyeon Lee, Brian J Burkett, Hoon-Ki Min, Matthew L Senjem, Ellen Dicks, Nick Corriveau-Lecavalier, Carly T Mester, Heather J Wiste, Emily S Lundt, Melissa E Murray, et al. Synthesizing images of tau pathology from cross-modal neuroimaging using deep learning. *Brain*, 147(3):980–995, 2024.

[Li *et al.*, 2023a] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 1952–1961, 2023.

[Li *et al.*, 2023b] Senmao Li, Joost van de Weijer, Fahad Khan, Tao Liu, Linxuan Li, Shiqi Yang, Yaxing Wang, Ming-Ming Cheng, et al. Faster diffusion: Rethinking the role of the encoder for diffusion model inference. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2023.

[Lv *et al.*, 2022] Tianxu Lv, Youqing Wu, Yihang Wang, Yuan Liu, Lihua Li, Chuxia Deng, and Xiang Pan. A hybrid hemodynamic knowledge-powered and feature reconstruction-guided scheme for breast cancer segmentation based on dce-mri. *Medical Image Analysis*, 82:102572, 2022.

[Lv *et al.*, 2023] Tianxu Lv, Yuan Liu, Kai Miao, Lihua Li, and Xiang Pan. Diffusion kinetic model for breast cancer segmentation in incomplete dce-mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 100–109. Springer, 2023.

[Lv *et al.*, 2024] Tianxu Lv, Xiaoyan Hong, Yuan Liu, Kai Miao, Heng Sun, Lihua Li, Chuxia Deng, Chunjuan Jiang, and Xiang Pan. Ai-powered interpretable imaging phenotypes noninvasively characterize tumor microenvironment associated with diverse molecular signatures and survival in breast cancer. *Computer Methods and Programs in Biomedicine*, 243:107857, 2024.

[Lyu *et al.*, 2023] Jinhao Lyu, Ying Fu, Mingliang Yang, Yongqin Xiong, Qi Duan, Caohui Duan, Xueyang Wang, Xinbo Xing, Dong Zhang, Jiaji Lin, et al. Generative adversarial network–based noncontrast ct angiography for aorta and carotid arteries. *Radiology*, 309(2):e230681, 2023.

[Ma *et al.*, 2024] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15762–15772, 2024.

[Newitt and Hylton, 2016] David Newitt and Nola Hylton. Single site breast dce-mri data and segmentations from patients undergoing neoadjuvant chemotherapy. *The Cancer Imaging Archive*, 2016.

[Parmar *et al.*, 2018] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International conference on machine learning*, pages 4055–4064. PMLR, 2018.

[Preechakul *et al.*, 2022] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10619–10629, 2022.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Saha *et al.*, 2021] A. Saha, M. R. Harowicz, L. J. Grimm, J. Weng, E. H. Cain, C. E. Kim, S. V. Ghate, R. Walsh, and M. A. Mazurowski. Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations. *The Cancer Imaging Archive*, 2021.

[Saharia *et al.*, 2022] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

[Sigger *et al.*, 2024] Neetu Sigger, Quoc-Tuan Vien, Sinh Van Nguyen, Gianluca Tozzi, and Tuan Thanh Nguyen. Unveiling the potential of diffusion model-based framework with transformer for hyperspectral image classification. *Scientific Reports*, 14(1):8438, 2024.

[Song *et al.*, 2021] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[Torigian *et al.*, 2007] Drew A Torigian, Steve S Huang, Mohamed Houseni, and Abass Alavi. Functional imaging of cancer with emphasis on molecular techniques. *CA: a cancer journal for clinicians*, 57(4):206–224, 2007.

[Vahdat and Kautz, 2020] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

[Vallières *et al.*, 2017] Martin Vallières, Emily Kay-Rivest, Léo Jean Perrin, Xavier Liem, Christophe Furstoss, Nader Khaouam, Phuc Félix Nguyen-Tan, Chang-Shu Wang, and Khalil Sultanem. Data from head-neck-pet-ct. *The Cancer Imaging Archive*, 2017.

[Vega *et al.*, 2024] Fernando Vega, Abdoljalil Addeh, Aravind Ganesh, Eric E Smith, and M Ethan MacDonald. Image translation for estimating two-dimensional axial amyloid-beta pet from structural mri. *Journal of Magnetic Resonance Imaging*, 59(3):1021–1031, 2024.

[Wang *et al.*, 2022] Yulin Wang, Wenyuan Wu, Yuxin Yang, Haifeng Hu, Shangqian Yu, Xiangjiang Dong, Feng Chen, and Qian Liu. Deep learning-based 3d mri contrast-enhanced synthesis from a 2d noncontrast t2flair sequence. *Medical Physics*, 49(7):4478–4493, 2022.

[Yang *et al.*, 2025] Bo Yang, Zhaohui Jiang, Dong Pan, Haoyang Yu, Gui Gui, and Weihua Gui. Lfdt-fusion: A latent feature-guided diffusion transformer model for general image fusion. *Information Fusion*, 113:102639, 2025.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.