

A Pilot Study on PET-to-CT Cross-Modal Translation: Systematic Validation of Brownian Bridge Diffusion Models

Xiaoyang Gu and Tianyu Shi *

School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110004, China

* Correspondence: ty.shi@sylu.edu.cn

How To Cite: Gu, X.; Shi, T. A Pilot Study on PET-to-CT Cross-Modal Translation: Systematic Validation of Brownian Bridge Diffusion Models. *AI Medicine* 2026, 3(2), 9. <https://doi.org/10.53941/aim.2026.100009>

Received: 25 June 2026

Revised: 21 August 2026

Accepted: 2 September 2026

Published: 20 September 2026

Abstract: Cross-modal PET-to-CT translation is an important problem in medical image computing. PET encodes the functional distribution of a radiotracer, whereas CT encodes anatomical electron density, and the two are separated by a substantial modality gap. Existing approaches have well-known limitations: generative adversarial networks (GANs) suffer from unstable training and anatomical artifacts, while standard denoising diffusion probabilistic models (DDPMs) converge slowly because the reverse process starts from pure noise. In this work, we systematically evaluate the Brownian Bridge diffusion model (BBDM) for prostate-specific membrane antigen (PSMA) PET-to-CT translation under a patient-disjoint experimental setting. BBDM constructs a stochastic bridge between the source and target domains, so that the reverse diffusion trajectory is initialized from the PET source endpoint rather than from pure Gaussian noise. The processed dataset was partitioned at the patient level into 287 training patients, 32 validation patients, and 35 independent test patients. Quantitative evaluation was performed on 3741 test slices using complementary image-quality metrics, and statistical significance was assessed using patient-level MAE values. Compared with the matched DDPM baseline using the same UNet backbone and training budget, BBDM achieved substantially lower reconstruction error, supporting the utility of source-anchored bridge diffusion relative to conventional noise-initialized diffusion for PET-to-CT synthesis. After adding SelfRDB-L1 as a recent diffusion-bridge baseline, BBDM remained competitive across the broader metric set, although SelfRDB-L1 achieved a lower patient-level MAE. These results suggest that BBDM is a competitive bridge-based formulation for PET-conditioned CT synthesis, while further clinical validation is required before diagnostic use.

Keywords: PET-to-CT translation; Brownian Bridge diffusion; cross-modal medical image synthesis

1. Introduction

1.1. Clinical Background and Problem Motivation

PET and CT are the two most important imaging modalities in nuclear medicine and radiology. With picomolar sensitivity, PET maps the three-dimensional in vivo distribution of a radiotracer and provides molecular-level functional information for tumor staging and treatment monitoring. CT, in turn, provides anatomical structure at sub-millimeter spatial resolution and quantifies tissue physical density in Hounsfield units (HU) [1]. The two modalities are highly complementary. PET/CT combines functional information from PET with anatomical information from CT and has become an important imaging modality in oncology [1]. Accurate



attenuation correction and attenuation map generation are essential technical components for quantitative PET/CT imaging [2,3].

However, CT acquisition contributes substantially to the radiation burden of hybrid PET/CT examinations, and the associated infrastructure requirements may limit its adoption in some clinical settings [4,5]. Cross-modal translation from PET to CT—the direct generation of a synthetic CT from PET data—offers a route to lowering these costs.

From a clinical standpoint, the development of CT-reduced or CT-free PET workflows based on synthetic anatomical information is motivated by two intertwined needs: reducing cumulative patient radiation exposure and lowering the infrastructure costs of healthcare systems worldwide.

In terms of radiation dosimetry, a standard whole-body PET/CT examination exposes patients to radiation from both the injected radiotracer and the CT acquisition. The CT-related radiation contribution varies substantially according to the acquisition purpose and scanning protocol, with attenuation-correction or localization CT generally requiring different exposure settings from diagnostic CT [4]. This cumulative radiation burden may be particularly relevant for patients requiring repeated examinations. Therefore, if CT-like anatomical information could be reliably generated from PET data in appropriately validated workflows, eliminating the CT acquisition could reduce CT-associated radiation exposure.

In terms of health economics, PET/CT systems require substantial capital investment, together with ongoing costs associated with system maintenance, infrastructure, and radiopharmaceutical management [5]. If sufficiently validated, deep-learning-based PET-to-CT translation may provide a complementary computational approach for estimating CT-like information from PET data. Such a strategy may have potential value in selected imaging scenarios; however, its applicability to attenuation correction, anatomical localization, or standalone PET systems requires dedicated downstream validation.

Potential downstream applications of PET-to-CT synthesis may include attenuation correction, radiotherapy-related imaging, and longitudinal PET assessment. However, image-level similarity alone is insufficient to establish suitability for these applications. Task-specific validation, including quantitative assessment of attenuation-correction accuracy, SUV bias, radiotherapy dose calculation, anatomical fidelity, and clinical interpretation, would be required before synthetic CT could be considered for such downstream use.

In summary, PET-to-CT translation represents a potentially useful methodological direction for cross-modal medical image synthesis, but its clinical value must ultimately be established through task-specific and prospective validation. Motivated by this need, the present study systematically evaluates the performance and mechanistic advantages of the Brownian Bridge diffusion model for this translation task.

1.2. Technical Challenges

PET-to-CT translation is, in essence, a highly challenging cross-modal mapping problem [6]. The PET signal reflects tracer-specific biological uptake patterns; for prostate-specific membrane antigen (PSMA) tracers, many non-target soft-tissue regions may show low or heterogeneous uptake, although physiological uptake can still occur in organs such as the salivary glands, kidneys, and urinary tract. A CT image, by contrast, must delineate substantial density differences across anatomical regions, including high-density cortical bone, lower-density trabecular bone, soft tissue, lung parenchyma, and air. These structures span a wide range of CT attenuation values, making accurate anatomical reconstruction particularly challenging. No simple deterministic mapping exists between the functional PET signal and anatomical density across most regions (Figure 1a–c). Therefore, the deep-learning model must learn a complex cross-modal relationship between tracer-dependent functional information and CT-derived anatomical density.

1.3. Limitations of Existing Methods and the Rationale of This Study

Current work on this task follows two main technical routes. The GAN route, represented by Pix2Pix [7] and MedGAN [8], pursues visual realism through the adversarial game between a generator and a discriminator. GAN-based methods can produce images with rich texture and sharp appearance, but their adversarial training optimizes for fooling the discriminator, which is fundamentally at odds with the pixel-level accuracy required in medical image translation. GAN-based methods can produce visually plausible synthetic CT images while still exhibiting substantial local HU errors, particularly in high-contrast anatomical regions. Such HU inaccuracies may propagate to downstream quantitative tasks such as PET attenuation correction, indicating that visual realism alone is insufficient for evaluating synthetic CT quality.

The diffusion-model route, represented by DDPM [9] and Palette [10], transforms pure random noise into a structured image through iterative denoising. Its advantages lie in stable training and comprehensive coverage of generative modes. Its structural limitation, however, is that the reverse diffusion is initialized from pure Gaussian

noise $x_T \sim N(0, I)$. In the early denoising stages (t close to T), the current sample x_t carries almost no structural information about the target CT. Specifically, under the linear β -schedule adopted in our DDPM baseline ($\beta_1 = 10^{-4}$, $\beta_T = 0.02$, $T = 1000$), the cumulative signal-retention coefficient progressively decreases during forward diffusion. For example, in a high-noise region around $t \approx 670$, the retained target-signal amplitude is only approximately 10%. Therefore, during the corresponding early reverse-sampling stage, the model must recover target-domain structure from a strongly noise-dominated state while simultaneously incorporating the PET condition. This observation is intended only to provide an intuitive interpretation of the DDPM trajectory under the adopted noise schedule and should not be interpreted as a quantitative partition of the reverse process, computational effort, or model capacity.

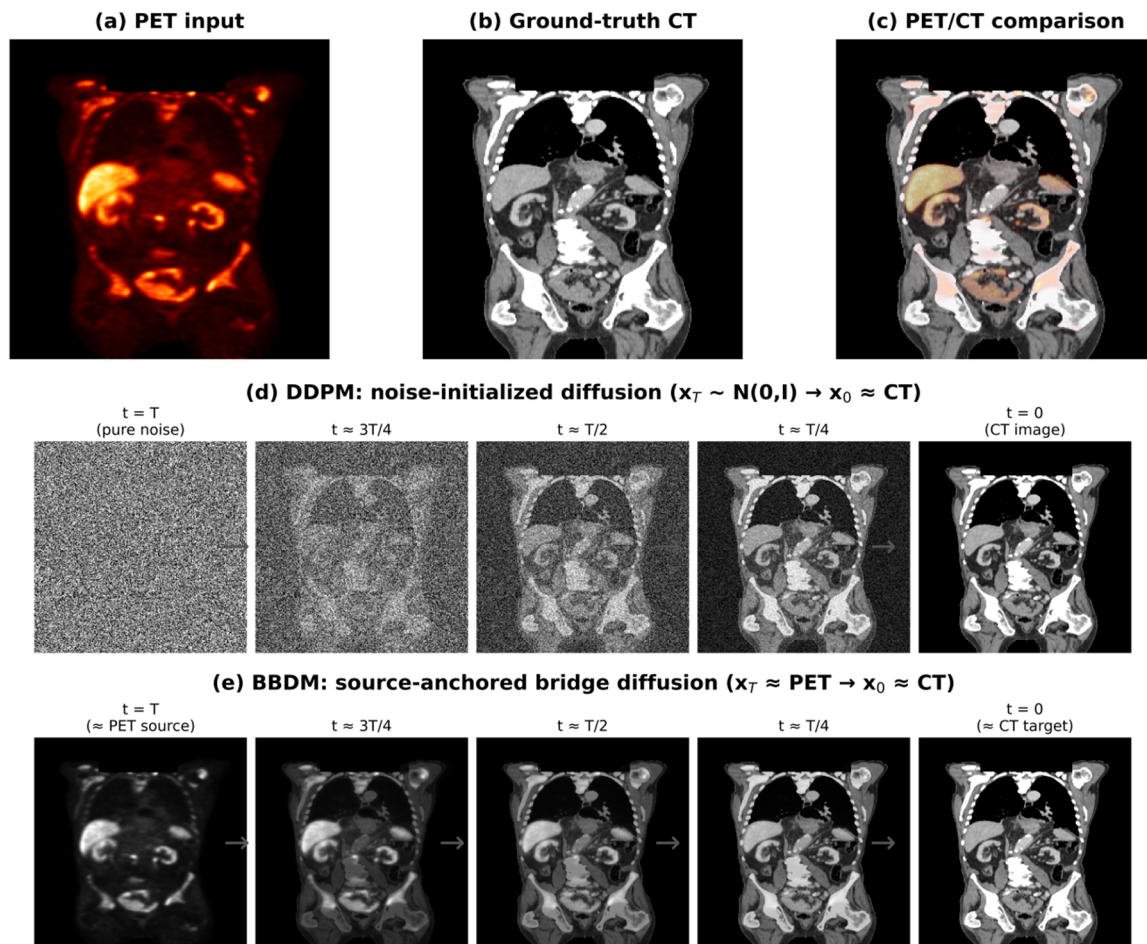


Figure 1. Conceptual overview of PET-to-CT translation using conventional diffusion and Brownian-bridge diffusion. Panels (a–c) show a representative paired PET/CT example from the processed dataset: (a) PET input, (b) spatially corresponding ground-truth CT, and (c) PET/CT comparison illustrating the modality gap between tracer-dependent functional uptake and anatomical density. Panel (d) schematically illustrates the noise-initialized trajectory of DDPM, whereas panel (e) schematically illustrates the source-anchored bridge trajectory of BBDM initialized from the PET source endpoint. Panels (d,e) provide qualitative illustrations of the difference between noise-to-image diffusion and source-to-target bridge diffusion and should not be interpreted as a quantitative decomposition of sampling steps, computational cost, or model capacity allocation.

The Brownian Bridge diffusion model (BBDM) [11] provides a source-anchored formulation for the above problem (Figure 1e). Rather than defining the diffusion process as a one-way degradation from the target image to pure noise, BBDM constructs a stochastic bridge between the source and target domains, with the paired PET and CT images serving as the two endpoints. Its forward process is defined as $x_t = (1 - m_t)x_0 + m_t y + \sigma_t \varepsilon$, where m_t increases from 0 to 1. Under this formulation, the reverse process is initialized from a state close to the PET source image y , rather than from uninformative Gaussian noise. This source-anchored trajectory provides a qualitative mechanism for retaining source-domain information during PET-conditioned CT synthesis.

It should be emphasized that this schematic is intended as a conceptual illustration of the difference between noise-to-image diffusion and source-to-target bridge diffusion. It should not be interpreted as a quantitative

decomposition of sampling steps, computational cost, or model capacity allocation. Although BBDM provides a source-anchored formulation for paired image-to-image translation, its application to PET-to-CT synthesis remains less explored. In this direction, the discrepancy between the source modality, functional PET, and the target modality, anatomical CT, is substantially larger than that between different sequences within the same modality or between different contrast-enhancement phases.

1.4. Contributions of This Work

This study systematically evaluates BBDM for PSMA PET-to-CT translation under a patient-disjoint experimental setting. The contributions are summarized as follows.

First, we formulate PET-to-CT translation as a source-anchored bridge-based conditional synthesis problem and evaluate BBDM on a processed PSMA PET/CT dataset split at the patient level. The dataset contains 287 training patients with 29,543 paired slices, 32 validation patients with 3903 paired slices, and 35 independent test patients with 3741 paired slices.

Second, we provide a controlled comparison between BBDM and a conventional DDPM baseline under the same 22.2M-parameter UNet architecture, the same patient-disjoint training set, and the same training budget. This comparison is used to assess the effect of the diffusion formulation under matched experimental conditions.

Third, we expand the comparison set by adding SelfRDB-L1 as a recent diffusion-bridge baseline and evaluate all methods using complementary quantitative metrics, including MAE, RMSE, PSNR, SSIM, MS-SSIM, and LPIPS. The revised results do not claim uniform superiority of BBDM; instead, they show that BBDM remains competitive with the recent bridge-based baseline while substantially outperforming the matched DDPM baseline.

Finally, we analyze qualitative outputs, spatial error distributions, and intermediate sampling trajectories to characterize the behavior and limitations of bridge-based PET-to-CT synthesis.

2. Related Work

2.1. Cross-Modal Medical Image Translation

Cross-modal translation of medical images has long been dominated by MRI-to-CT conversion and vice versa [12,13], with some work also targeting synthesis across MRI pulse sequences [14]. Translation in the PET-to-CT direction has received comparatively little attention. Previous deep-learning-based approaches have explored PET-derived attenuation correction and synthetic anatomical information generation using convolutional neural networks [6]. These methods demonstrate the feasibility of learning PET-to-CT-like mappings; however, pixel-wise reconstruction losses alone may favor overly smooth predictions when the source modality does not uniquely determine high-frequency target-domain details, such as trabecular bone structure. The MedGAN of Armanious et al. [8] cascades several U-Net blocks as the generator and adds a VGG-feature perceptual loss and an SSIM-based structural loss, and is effective across several medical image translation tasks including PET-to-CT. However, GAN-based synthesis may be sensitive to adversarial training instability and implementation details, which can affect the reproducibility and numerical reliability of HU-level outputs in settings that require precise pixel values for attenuation correction. To date, no systematic, controlled comparison of different diffusion paradigms—in particular standard DDPM versus the Brownian Bridge process—on the PET-to-CT task exists.

Hu et al. [15] explored PET-to-CT synthesis from non-attenuation-corrected PET using a Wasserstein GAN-based framework.

More broadly, image-to-image translation has evolved beyond paired convolutional GANs. MUNIT [16] introduced a multimodal unsupervised formulation that decomposes images into a domain-invariant content representation and domain-specific style codes, enabling diverse translation without paired observations. For medical imaging, SynDiff [17] combines a conditional diffusion process with adversarial projections and cycle-consistent training to support unpaired translation, with demonstrations on multi-contrast MRI and MRI-to-CT synthesis. More recently, I2I-Mamba [18] uses selective state-space modeling within a hybrid convolutional architecture to capture long-range contextual dependencies efficiently and has been evaluated on multi-contrast MRI and MRI-to-CT synthesis. These approaches represent complementary developments in unsupervised representation disentanglement, diffusion-based generation, and state-space modeling. They broaden the architectural landscape of medical image translation, although their pairing assumptions and imaging protocols differ from the patient-disjoint paired PSMA PET-to-CT setting considered here.

2.2. GAN-Based Image Translation Methods

Pix2Pix [7], pairing a U-Net generator with a PatchGAN discriminator and jointly optimizing L1 and adversarial losses, established the paradigm of conditional GANs for paired image-to-image (I2I) translation. In medical imaging, Pix2Pix has been widely used for tasks such as CT denoising and MR-to-CT synthesis. The key innovation of MedGAN [8] is the adoption, as in style transfer, of intermediate-layer feature matching of a pretrained VGG-19 network as the perceptual-loss target, and of SSIM as a structural-similarity-loss target, which markedly improves the textural fidelity specific to medical images. Upadhyay et al. [19] proposed uncertainty-guided progressive GANs that incorporate aleatoric uncertainty as attention maps during progressive training for medical image translation.

Although GAN-based methods have undeniable advantages in perceptual quality, applying them to medical image translation-whose core metric is per-pixel HU accuracy-poses fundamental obstacles. The objective optimized by a GAN generator corresponds, at the optimum, to making the distribution of generated samples indistinguishable from that of real samples in the discriminator's feature space. This criterion does not entail that each pixel of the synthetic CT approaches the corresponding pixel of the ground truth. A GAN can therefore produce visually plausible CT slices while retaining substantial local HU deviations from the ground truth, particularly in high-contrast anatomical regions. Such errors may adversely affect downstream quantitative applications, including PET attenuation correction, where accurate tissue-density estimation is important.

2.3. Diffusion-Model-Based Image Translation Methods

The DDPM of Ho et al. [9] models image generation as an iterative denoising process from noise, via the inverse transition kernel of a parameterized Markov chain. Palette [10] extends DDPM to conditional generation: at each denoising step, the source image is concatenated with the current noisy estimate along the channel dimension and fed into the UNet. SDEdit [20] uses the SDE formulation of diffusion models for more flexible guided image editing and synthesis. The chief advantages of diffusion models over GANs are training stability and mode coverage: their denoising likelihood-based objective avoids the adversarial min-max optimization that often makes GAN training unstable.

Recent medical image translation studies have also adapted diffusion models beyond conventional noise-to-image denoising. SynDiff [17], for example, couples conditional diffusion with adversarial projections in the reverse process and uses a cycle-consistent architecture for unpaired training, demonstrating diffusion-based translation across multi-contrast MRI and MRI-to-CT tasks. This line of work illustrates the broader flexibility of diffusion formulations for medical image translation, while the present study focuses specifically on a paired source-to-target Brownian-bridge formulation for PET-to-CT synthesis.

However, for the specific subtask of I2I translation, standard DDPM starts the reverse process from a strongly noise-dominated state. At large t , the intermediate state x_t contains little target-domain structural information. Although the source image y is explicitly provided as a conditioning input throughout reverse sampling, the model must recover target-domain structure from this highly corrupted intermediate state. In contrast, Brownian-bridge diffusion initializes the reverse trajectory from a source-anchored state, providing a structurally more informative starting point for cross-modal translation. The experimental analysis in Section 4.7 and the mechanistic discussion in Section 5.1 of this paper provide empirical evidence and a theoretical explanation for this phenomenon.

2.4. Brownian Bridge Diffusion Model and Its Medical Applications

The BBDM of Li et al. [11] addresses the above structural bottleneck of DDPM by reformulating the forward process as a stochastic bridge between two endpoints. Its essence is to change the diffusion origin from pure noise to the source-domain image, so that the initial state of the reverse process inherently carries prior information about the target anatomy. BBDM has shown better quantitative performance than both GANs and DDPMs under equal training budgets on several natural-image benchmarks, including CelebA facial attribute editing and Edges2Shoes edge-to-photo translation.

In medical imaging, BBDM has recently been extended by several groups: Slice-Consistent BBDM [21] introduces a trajectory-alignment mechanism across 2D slices for 3D volumetric brain CT-to-MRI translation; HiFi-BBrg [22] incorporates the reconstruction mapping of a conditional GAN into the bridge training pipeline for deterministic medical image translation; and SC-BBDM [23] adds a slice-consistency constraint for synthesizing contrast-enhanced chest CT to preserve volumetric continuity. None of these works, however, addresses the PET-to-CT direction, where the gap in image properties between the source modality (functional PET) and the target modality (anatomical CT) is far larger than that between sequences within a modality or between contrast phases, placing markedly higher demands on the cross-modal modeling capability of the bridge process.

More recently, diffusion-bridge models have been further developed for image-to-image translation. SelfRDB [24] is a representative recent bridge-based formulation that improves residual bridge diffusion for paired image translation. Because no task-specific SelfRDB configuration is available for PSMA PET-to-CT translation, we adapted the publicly available SelfRDB implementation to this task and trained an L1-adapted SelfRDB baseline under the same preprocessing and evaluation protocol as the other methods. Its inclusion provides a stronger contemporary bridge-based comparison and helps avoid evaluating BBDM only against GANs and conventional DDPM. A structured comparison of these representative methods is summarized in Table 1.

Table 1. Literature review summary of representative image-to-image translation and diffusion-bridge methods.

Category	Representative Methods	Main Idea	Relevance to PET-to-CT Translation	Limitation in This Study
GAN-based image translation	Pix2Pix, MedGAN	Learn direct source-to-target mapping with adversarial and reconstruction losses	Common baselines for medical image synthesis	May suffer from unstable training, artifacts, or limited HU accuracy
Conventional diffusion	DDPM	Generate target images through iterative denoising from Gaussian noise	Strong generative baseline and matched ablation reference	Does not explicitly model a source-to-target bridge
Brownian-bridge diffusion	BBDM	Construct a stochastic bridge between source and target domains	Source-anchored formulation suitable for paired PET-to-CT synthesis	Should be evaluated against recent bridge variants
Recent bridge-based I2I diffusion	SelfRDB-L1	Adapt residual bridge diffusion for paired image translation	Provides a stronger contemporary bridge-based baseline	No standardized PSMA PET-to-CT configuration has been publicly established
General/recent I2I architectures	MUNIT, SynDiff, I2I-Mamba	Unpaired content–style disentanglement, adversarial diffusion, and selective state-space modeling	Broadens the methodological context beyond GAN, DDPM, and diffusion-bridge formulations	Different pairing assumptions and/or imaging protocols; not direct matched PET-to-CT baselines

Overall, this summary places the present study within the broader landscape of medical image translation while motivating the inclusion of both conventional baselines and a recent diffusion-bridge baseline in the experimental comparison.

3. Method

3.1. Problem Formulation

Given a source-domain image $y \in R^{H \times W}$ (PET, normalized to $[-1, 1]$) and a target-domain image $x_0 \in R^{H \times W}$ (CT, normalized to $[-1, 1]$), the cross-modal PET-to-CT translation task aims to learn a conditional distribution $p(x_0|y)$ that approximates the paired CT distribution conditioned on the observed PET image. Importantly, because PET and CT encode different physical properties, this conditional mapping is intrinsically underdetermined, and the PET observation may not uniquely specify all anatomical structures visible in CT. The core difficulty of the task lies in the fact that y and x_0 originate from different physical imaging mechanisms, and the signal distribution in y —that is, tracer-specific uptake and x_0 —that is, electron density—lack a simple monotonic mapping relationship across most anatomical regions.

3.2. Brownian Bridge Forward Diffusion Process

Let T denote the total number of diffusion steps. Define the bridge ratio parameter:

$$m_t = \frac{t}{T}, \quad t \in \{1, 2, \dots, T\} \quad (1)$$

The Brownian Bridge forward diffusion process is defined as:

$$x_t = (1 - m_t) \cdot x_0 + m_t \cdot y + \sigma_t \cdot \varepsilon, \quad \varepsilon \sim N(0, I) \quad (2)$$

where the bridge variance is:

$$\sigma_t^2 = 2 \cdot v \cdot m_t \cdot (1 - m_t) \quad (3)$$

where $v = 1.0$ (denoted max_var) is a hyperparameter controlling the overall noise magnitude. Equation (2) defines a continuous trajectory from x_0 ($t = 0, m_0 = 0$) to y ($t = T, m_T \approx 1$). Figure 2 illustrates the overall architecture of this process schematically. As schematically illustrated in Figure 1d,e, this bridge construction causes the reverse

process to start naturally from the PET source endpoint, which provides a source-anchored initialization for PET-conditioned CT synthesis, rather than from the uninformative pure noise used in standard DDPM.

The factor 2 in the bridge-variance formula (3) originates from the mathematical construction of the Brownian bridge: it ensures that at m (the midpoint of the diffusion), the variance reaches its maximum value of $v/2$, so that the diffusion path retains low noise near the two end points (ensuring structural fidelity) while introducing moderate stochasticity at the midpoint to facilitate the learning of multimodal distributions. Compared with the variance schedule of DDPM, α_t the bridge variance possesses a key structural property- $\sigma_T^2 \approx 0$ (when $m_T, 1 - m_T \approx 0$), which means that $x_T \approx y$, i.e., the endpoint of the diffusion naturally “stops” near the source-domain image rather than near pure noise. The implication of this property is elaborated in Section 3.5.

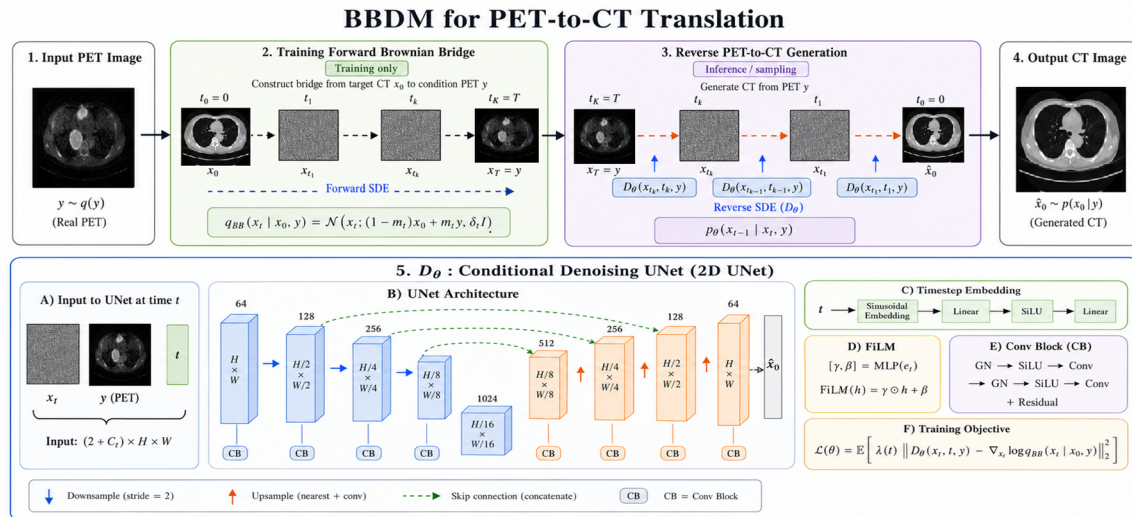


Figure 2. The Brownian Bridge diffusion-based framework for cross-modal PET→CT translation. PET and CT images serve directly as the endpoints of the diffusion process in pixel space; BBDM constructs a stochastic bridge between the two domains-the forward process perturbs CT toward PET, and the reverse process starts from PET and recovers CT through a conditional UNet denoiser, all performed in normalized pixel space.

3.3. Bridge Gradient Prediction and Training Objective

The denoising network $f_\theta(\cdot)$ takes the current state x_t , the source image y , and the time step t as inputs and outputs a prediction of the “bridge gradient”. This bridge gradient is defined as:

$$G_t = m_t \cdot (y - x_0) + \sigma_t \cdot \varepsilon \quad (4)$$

Note that G_t , unlike the prediction target ε of DDPM, is not pure noise but rather a weighted combination of two components: $m_t(y - x_0)$ encodes the structural residual between PET and CT at diffusion progress m_t , and $\sigma_t \cdot \varepsilon$ encodes the random-noise component. When the diffusion is early and m_t is small, G_t is dominated by noise; when the diffusion is late and m_t approaches 1, G_t is dominated by the structural residual. This time-dependent target construction endows the training signal with different information densities at different stages of the diffusion, helping the network learn modality-mapping relationships at different levels and across different time scales.

The training objective is to minimize the L1 prediction error:

$$\mathcal{L}(\theta) = E_{t, x_0, y, \varepsilon} \left[\| f_\theta(x_t, y, t) - G_t \|_1 \right] \quad (5)$$

where t is sampled uniformly from $\{1, \dots, T\}$, and the noise term is sampled from a standard normal distribution. We adopt the L1 loss rather than the L2 loss because CT synthesis requires accurate preservation of voxel-level HU values and anatomical transitions. Compared with the quadratic penalty imposed by L2 loss, L1 loss is less dominated by large local errors and is generally less prone to excessive smoothing. It may therefore be more suitable for maintaining HU transitions at anatomical boundaries.

3.4. Conditional UNet Denoiser Architecture

The denoising network f_θ adopts an encoder-decoder UNet architecture [11]. The input is the concatenation of x_t and y along the channel dimension (2 channels), and the output is a single-channel bridge-gradient prediction of the same size as x_t . The network consists of three parts: an encoder, a bottleneck, and a decoder.

The encoder contains four successive downsampling operations, reducing the spatial resolution from 256×256 to 128×128 , 64×64 , 32×32 , and finally 16×16 . The feature-channel dimensions progressively increase from 64 to 128, 256, and 512, followed by a 1024-channel bottleneck at 16×16 resolution, consistent with the architecture illustrated in Figure 2. Each convolution block (CB) adopts a post-activation residual structure consisting of 3×3 convolution $\rightarrow$ GroupNorm (8 groups) $\rightarrow$ SiLU $\rightarrow$ FiLM time-step modulation $\rightarrow$ 3×3 convolution $\rightarrow$ GroupNorm $\rightarrow$ SiLU, with a 1×1 convolution used in the residual branch when the channel dimensions differ. The time step t is encoded using sinusoidal positional encoding and subsequently transformed by a two-layer MLP. The resulting embedding is used to generate the scale and bias parameters for FiLM modulation, enabling each convolution block to adapt its feature representation according to the current diffusion time step.

Self-attention is applied at the 16×16 bottleneck resolution to capture long-range spatial dependencies. In CT synthesis, this mechanism facilitates the modeling of global anatomical configurations and relative spatial relationships among structures. The attention layer follows a standard self-attention formulation: queries (Q), keys (K), and values (V) are projected from the normalized feature maps, and the resulting attention weights are normalized using Softmax before being applied to V.

The decoder is mirror-symmetric to the encoder; through skip connections, the feature maps at each level of the encoder are concatenated along the channel dimension with the corresponding feature maps of the decoder and then fed into the residual blocks. Upsampling is performed via 4×4 transposed convolutions (stride 2). The output layer is GroupNorm $\rightarrow$ SiLU $\rightarrow$ 3×3 convolution.

The network has approximately 22.2M parameters, with `base_channels` = 64, channel multipliers `channel_mults` = [1, 2, 4, 8], and a time-embedding dimension of `time_dim` = 256. The encoder feature dimensions are 64, 128, 256, and 512 channels, followed by a 1024-channel bottleneck at the 16×16 resolution, where standard self-attention is applied. This configuration strikes a practical balance between model capacity and training/inference efficiency.

3.5. Reverse Sampling Process

During inference, the reverse process is initialized from $x_T = y$. At step i (decreasing from T to 1), the network predicts $f_\theta(x_t, y, t)$, from which the current estimate of x_0 is recovered:

$$\hat{x}_0 = x_t - f_\theta(x_t, y, t) \quad (6)$$

This formula follows from a rearrangement of Equation (4): when the network predicts $f_\theta \approx G_t = m_t(y - x_0) + \sigma_t \varepsilon$, $x_t - G_t \approx x_t - m_t(y - x_0) - \sigma_t \varepsilon = (1 - m_t)x_0 + m_t y + \sigma_t \varepsilon - m_t(y - x_0) - \sigma_t \varepsilon = x_0$.

Given $\hat{x}_0$, the next-step state x_{t-1} is computed via the bridge reverse transition kernel:

$$x_{t-1} = (1 - m_{t-1}) \cdot \hat{x}_0 + m_{t-1} \cdot y + \frac{\sqrt{\sigma_{t-1}^2 - \frac{\sigma_t^2(1 - m_{t-1})^2}{(1 - m_t)^2}} \cdot (x_t - (1 - m_t)\hat{x}_0 - m_t y)}{\sigma_t} \quad (7)$$

This formula is the analytical form of the standard Brownian Bridge reverse transition, derived from the properties of Gaussian conditional distributions [11]. When $m_{t-1} \rightarrow 0$, $x_{t-1} \rightarrow \hat{x}_0$; when $m_{t-1} \rightarrow 1$, $x_{t-1} \rightarrow y$. The deterministic component (the mean term) of the transition is a weighted combination of $\hat{x}_0$ and y , and the stochastic component (the noise term) is scaled by the bridge-variance residual.

To accelerate inference, a DDIM-style reparameterization [25] is adopted to compress the $T = 1000$ training steps into 200 sampling steps. Under this reparameterization framework, the sampling step size Δ can take integer values greater than 1, skipping multiple steps per sampling step. The parameter η controls the stochasticity of sampling: $\eta = 0$ corresponds to deterministic DDIM (the same input always yields the same output), $\eta = 1$ recovers the fully stochastic DDPM-equivalent process. Unless otherwise noted, all quantitative results in this paper use $\eta = 0$ (deterministic sampling) to ensure reproducible inference and to minimize sampling-induced variability during subsequent evaluation.

3.6. Analysis of the Advantages of the Conditioning Mechanism

The conditioning mechanism of BBDM differs fundamentally from that of a standard conditional DDPM, a difference that stems directly from the mathematical properties of the bridge construction. In the forward process of a standard conditional DDPM, $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\varepsilon$ degrades to pure noise at $t = T$, so during the early stages of reverse denoising (t close to T), x_t the target-domain structural information contained in x_t tends to zero. Although the conditional image y is provided to the UNet throughout reverse sampling, the intermediate state x_t

itself remains strongly noise-dominated at large t , increasing the difficulty of recovering target-domain structure during the early reverse trajectory.

In BBDM, the above situation is improved. Because $x_t = (1 - m_t)x_0 + m_t y + \sigma_t \varepsilon$, as the diffusion approaches its endpoint ($m_t \approx 1$), $x_t \approx y$, x_T naturally approaches the source image, and the network estimates CT in a feature space close to PET from the very first denoising step. As the reverse process proceeds from $t = T$ to $t = 0$, the bridge state gradually transitions from being close to the source PET image toward the target CT image. Consequently, the intermediate states remain anchored to source-domain information throughout the reverse trajectory. Compared with conventional DDPM, this formulation may reduce the burden of recovering target-domain structure from strongly noise-dominated states and may facilitate the progressive incorporation of cross-modal conditioning information. The experimental analysis in Section 4.7 provides empirical observations consistent with this interpretation.

3.7. Training Procedure

Training uses the AdamW optimizer (learning rate 1×10^{-4} , weight decay 1×10^{-4} , $\beta_1 = 0.9$, $\beta_2 = 0.999$), with a batch size of 32 for a total of 200 epochs. Each training step comprises the following procedure: (i) randomly sample a batch of PET-CT image pairs from the training set; (ii) for each pair, sample a time step t uniformly from $\{1, \dots, T\}$ and sample noise ε ; (iii) construct x_t ; (iv) compute the bridge gradient target G_t ; (v) compute the L1 loss (Equation (5)) and backpropagate; (vi) apply gradient clipping with a norm of 1.0; (vii) update the network parameters (see Algorithm 1). The complete 200-epoch training takes approximately 6 h on a single NVIDIA RTX 5090 GPU (32 GB of memory).

Algorithm 1: BBDM Training Procedure

Input: training set $D = \{(y_i, x_{0i})\}$ of paired PET-CT images, total diffusion steps T and network f_θ

```

1:   for epoch = 1 to 200 do
2:     for each minibatch  $\{(y, x_0)\} \sim D$  do
3:       Sample time step  $t \sim \text{Uniform}(\{1, \dots, T\})$ 
4:       Sample noise  $\varepsilon \sim N(0, I)$ 
5:       Compute bridge ratio  $m_t \leftarrow t/T$ 
6:       Compute bridge variance  $\sigma_t \leftarrow \sqrt{2 \cdot v \cdot m_t(1 - m_t)}$  ( $v = 1.0$ )
7:       Construct forward diffusion state  $x_t \leftarrow (1 - m_t) \cdot x_0 + m_t \cdot y + \sigma_t \varepsilon$ 
8:       Compute bridge gradient target  $G_t \leftarrow m_t(y - x_0) + \sigma_t \varepsilon$ 
9:       Compute L1 loss  $L \leftarrow \|f_\theta(x_t, y, t) - G_t\|_1$ 
10:      Backpropagate  $L$ 
11:      Apply gradient clipping with a maximum norm of 1.0
12:      Update network parameters  $\theta$  using AdamW
13:    end for
14:  end for
15:  return  $\theta$ 

```

4. Experiments

4.1. Dataset Construction and Preprocessing

This study uses the PSMA-PET-CT-Lesions dataset [26] publicly released by The Cancer Imaging Archive (TCIA). The data were collected at LMU University Hospital in Munich, Germany, between 2014 and 2022, and comprise 597 whole-body PSMA PET/CT studies from 378 patients with suspected or diagnosed prostate carcinoma. During preprocessing, 24 patients were excluded because their data could not be retained as valid paired PET/CT samples after the preprocessing and quality-control pipeline, resulting in a final cohort of 354 patients. The original data are stored as axial slices in DICOM format. To construct training data for the cross-modal generative model, the following preprocessing pipeline converts them into pairs of two-dimensional coronal slices: (i) extract the pixel spacing, slice position, and image orientation from the DICOM headers; (ii) reconstruct a three-dimensional volume from the consecutive axial slices according to the DICOM spatial metadata, and uniformly resample the voxel spacing to $1.0 \text{ mm} \times 1.0 \text{ mm} \times 1.0 \text{ mm}$ to ensure spatial-scale consistency across slices of different orientations (trilinear interpolation is used for CT to keep tissue edges continuous and smooth); (iii) generate coronal two-dimensional slices from the resampled three-dimensional volume as the training samples for the cross-modal generative model—the coronal rather than the axial plane is chosen because coronal slices of whole-body PET/CT scans preserve the longitudinal anatomical continuity of the body more completely, and a single slice covers a larger anatomical extent, which helps the model learn the overall correspondence of whole-

body structures; (iv) automatically pair the PET and CT coronal slices according to the spatial coordinates in the DICOM metadata; and (v) resample all images to a uniform size of 256×256 pixels via bilinear interpolation.

$$HU = I_{\text{raw}} \times \text{RescaleSlope} + \text{RescaleIntercept} \quad (8)$$

where RescaleSlope (DICOM tag 0028, 1053) and RescaleIntercept (0028, 1052) are determined by the scanner manufacturer during calibration. The converted HU values were clipped to $[-1000, 1000]$ to reduce the influence of extreme intensity values, including those caused by high-density foreign objects (e.g., metal implants), and to maintain a consistent intensity range for model training. The clipped CT values were then normalized to $[-1, 1]$ by division by 1000. (vi) The storage unit of the PET voxel values is determined according to the DICOM tag Units (0054, 1001). If the values are stored in BQML (becquerels per milliliter), they are converted to standardized uptake values (SUV) via the following formula:

$$SUV = \frac{A_{\text{voxel}} \times W \times 1000}{D_{\text{inj}}} \quad (9)$$

where A_{voxel} is the voxel radioactivity concentration (Bq/mL), W is the patient's body weight (kg), and D_{inj} is the injected dose (Bq); the factor of 1000 converts body weight from kilograms to grams. If the PET data are already stored in SUV, the original values are used directly. The converted SUV values are clipped to $[0, 20]$ to limit the influence of extremely high-intensity signals, such as injection-site activity or urinary bladder accumulation, and are then normalized to $[-1, 1]$ via the linear mapping $I_{\text{norm}} = I_{\text{SUV}}/10 - 1$.

The processed dataset was partitioned at the patient level into 287 training patients (29,543 paired slices), 32 validation patients (3903 paired slices), and 35 test patients (3741 paired slices). All slices from a given patient were assigned exclusively to one subset, with no patient overlap among the training, validation, and test sets, thereby preventing subject-level data leakage.

4.2. Evaluation Metrics

Following standard practice in medical image synthesis, we adopt six complementary quantitative metrics to characterize synthetic-CT quality from pixel-wise, structural, and perceptual perspectives. Each metric is computed per image on all 3741 test slices and reported as mean $\pm$ standard deviation for descriptive evaluation. For inferential statistical testing, slice-level errors are further aggregated at the patient level to account for subject-level dependence.

MAE (Mean Absolute Error, in HU): defined as the mean of the per-pixel absolute deviation between the synthetic CT and the ground-truth CT. MAE is the most direct metric for measuring the pixel-level accuracy of the synthetic CT relative to the ground truth.

RMSE (Root Mean Square Error, in HU): defined as the square root of the mean squared pixel-wise difference between the synthetic CT and the ground-truth CT. Compared with MAE, RMSE assigns a larger penalty to large local errors and is therefore useful for evaluating high-error regions.

PSNR (Peak Signal-to-Noise Ratio, in dB): defined as $20 \cdot \log_{10}(\text{DR}/\text{RMSE})$, where $\text{DR} = 2000$ HU is the dynamic range of the data $[-1000, 1000]$ HU. PSNR is sensitive to global signal fidelity but insensitive to small shifts in local structure.

SSIM (Structural Similarity Index Measure) [27]: computes the perceptual similarity between the synthetic CT and the ground-truth CT along three dimensions-luminance, contrast, and structure. SSIM computes local statistics including (mean, variance, and covariance) via a sliding window and is therefore more sensitive to the preservation of local image structure.

MS-SSIM (Multi-Scale Structural Similarity Index Measure) [28]: extends SSIM by computing structural similarity across multiple image scales. It provides a more comprehensive assessment of structural consistency by considering both fine-scale local details and coarse-scale image structures.

LPIPS (Learned Perceptual Image Patch Similarity) [29]: the L2 distance based on the intermediate-layer features of a pretrained deep network (AlexNet). Unlike SSIM, which is based on shallow statistics, LPIPS captures the similarity of high-level semantic features and is closer to the perception of image quality by the human visual system.

The quantitative image-quality metrics were computed at the slice level across all 3741 test slices and are reported as mean $\pm$ standard deviation for descriptive comparison. Because multiple slices originated from the same patient, individual slices were not treated as statistically independent observations for inferential analysis. For each method, slice-level MAE values were first averaged within each of the 35 test patients to obtain one patient-level MAE value per subject. Pairwise comparisons between BBDM and each comparison method were subsequently

performed using two-tailed paired t-tests on these patient-level MAE values. Bonferroni correction was applied to the four primary pairwise comparisons, resulting in an adjusted significance threshold of $p < 0.0125$.

4.3. Comparison Methods

To ensure a comprehensive, fair, and reproducible comparison, we select the following methods for systematic evaluation. The preprocessing pipeline is identical across all methods.

Pix2Pix [7]: a conditional GAN baseline. The generator is a U-Net and the discriminator is a 70×70 PatchGAN. The L1 loss and the adversarial loss are jointly optimized. Trained for 200 epochs with the Adam optimizer ($\text{lr} = 2 \times 10^{-4}$).

MedGAN [8]: a cascaded GAN baseline. The generator is a CasNet of 6 cascaded U-Net blocks, and the discriminator is a PatchGAN. The loss combines an adversarial loss, an L1 loss, a VGG-19 perceptual loss computed at conv1_2, conv2_2, conv3_3, and conv4_3, and an SSIM structural loss. Trained for 200 epochs.

DDPM ablation baseline: A standard conditional diffusion model was implemented in-house as a controlled ablation baseline. DDPM uses the same 22.2M-parameter UNet architecture, training data, number of training epochs, and batch size as BBDM. The forward diffusion process follows the standard DDPM formulation:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{(1 - \bar{\alpha}_t)} \epsilon \quad (10)$$

a linear noise schedule from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$ is used. The network predicts the noise term ϵ using an L1 loss, and reverse sampling is performed using DDIM [25] with 200 sampling steps and $\eta = 0$. The network architecture, training data, training budget, and evaluation protocol were matched between DDPM and BBDM; the primary methodological difference lies in the diffusion formulation and its corresponding prediction target. This design provides a controlled comparison between conventional noise-initialized diffusion and Brownian-bridge diffusion.

SelfRDB-L1 [24]: a recent bridge-based image-to-image translation baseline adapted from the publicly available SelfRDB implementation. Because no universally established or task-specific SelfRDB configuration is available for PSMA PET-to-CT translation, the original framework requires task-dependent adaptation. To improve comparability with BBDM, whose training objective in the present study is based on an L1 reconstruction loss, we therefore adopted an L1-based SelfRDB variant, denoted as SelfRDB-L1. SelfRDB-L1 was trained using the same preprocessing pipeline and evaluation protocol as the other methods.

Table 2 summarizes, in matrix form, the characteristics of each method along four dimensions: method type, diffusion formulation, translation trajectory, and conditioning. Table 3 lists the training hyperparameters aligned item by item between BBDM and DDPM.

Table 2. Method characteristic comparison.

Method	Type	Diffusion Formulation	Translation Trajectory	Conditioning
Pix2Pix	GAN	Adversarial training	PET $\rightarrow$ CT	PET source image
MedGAN	GAN	Adversarial training	PET $\rightarrow$ CT	PET source image
DDPM	Diffusion	Standard DDPM	Noise $\rightarrow$ CT	PET source image
SelfRDB-L1	Diffusion	Residual bridge diffusion	PET $\rightarrow$ CT	PET source image
BBDM (ours)	Diffusion	Brownian Bridge diffusion	PET $\rightarrow$ CT	PET source image

Table 3. Ablation control variables.

Hyperparameter	BBDM	DDPM
UNet base_channels	64	64
channel_mults	[1, 2, 4, 8]	[1, 2, 4, 8]
Parameters	22.2M	22.2M
Batch size	32	32
epoch	200	200
Optimizer	AdamW	AdamW
Learning rate	1×10^{-4}	1×10^{-4}
T	1000	1000
Inference steps	200 (DDIM)	200 (DDIM)
Diffusion schedule	$m_t = t/T$ (bridge)	$\beta: 1 \times 10^{-4} \rightarrow 0.02$
max_var	1.0	-
GPU	RTX 5090	RTX 5090

4.4. Main Quantitative Results

Table 4 summarizes the quantitative results of all methods on six complementary image-quality metrics, including MAE, RMSE, PSNR, SSIM, MS-SSIM, and LPIPS. The values in Table 4 are reported as slice-level mean $\pm$ standard deviation across the 3741 test slices and are used for descriptive comparison. Statistical significance was assessed separately at the patient level to account for within-subject dependence.

Among the evaluated methods, SelfRDB-L1 achieved the lowest MAE and RMSE and the highest SSIM, whereas BBDM achieved slightly higher PSNR and MS-SSIM. MedGAN achieved the lowest LPIPS, indicating that perceptual similarity and HU-level reconstruction accuracy are not fully aligned in this PET-to-CT synthesis task. Therefore, the revised results do not support a claim that BBDM uniformly outperforms all comparison methods. Instead, BBDM remains competitive across the broader quantitative metric set and shows a substantial advantage over the matched DDPM baseline.

Table 4. Quantitative comparison.

Method	MAE↓(HU)	RMSE↓(HU)	PSNR↑(dB)	SSIM↑	MS-SSIM↑	LPIPS↓
Pix2Pix	42.40 $\pm$ 7.55	90.50 $\pm$ 14.73	26.99 $\pm$ 1.34	0.7491 $\pm$ 0.0463	0.9174 $\pm$ 0.0206	0.0734 $\pm$ 0.0196
MedGAN	25.90 $\pm$ 4.85	50.54 $\pm$ 11.31	32.12 $\pm$ 1.65	0.8740 $\pm$ 0.0277	0.9696 $\pm$ 0.0105	0.0301 $\pm$ 0.0120
DDPM	33.23 $\pm$ 9.26	66.01 $\pm$ 25.03	29.94 $\pm$ 2.03	0.8354 $\pm$ 0.0322	0.9560 $\pm$ 0.0195	0.0507 $\pm$ 0.0199
SelfRDB-L1	21.56 $\pm$ 4.42	44.88 $\pm$ 11.52	33.26 $\pm$ 1.77	0.9038 $\pm$ 0.0203	0.9753 $\pm$ 0.0076	0.0536 $\pm$ 0.0133
BBDM	21.88 $\pm$ 6.56	44.90 $\pm$ 19.19	33.36 $\pm$ 2.25	0.9020 $\pm$ 0.0249	0.9754 $\pm$ 0.0121	0.0537 $\pm$ 0.0155

Note: Bold values indicate the best-performing result for each metric.

To account for within-subject dependence, statistical significance was further assessed using patient-level MAE values. Across the 35 independent test patients, the patient-level MAE was 22.24 $\pm$ 1.87 HU for BBDM, 21.81 $\pm$ 1.82 HU for SelfRDB-L1, 33.57 $\pm$ 2.79 HU for DDPM, 26.88 $\pm$ 2.22 HU for MedGAN, and 44.23 $\pm$ 4.09 HU for Pix2Pix. BBDM achieved significantly lower patient-level MAE than DDPM ($t(34) = -36.8$, $p < 0.0001$), MedGAN ($t(34) = -36.6$, $p < 0.0001$), and Pix2Pix ($t(34) = -47.4$, $p < 0.0001$). SelfRDB-L1 yielded a lower patient-level MAE than BBDM (21.81 vs. 22.24 HU; $t(34) = 3.77$, $p = 0.00065$). All four comparisons remained statistically significant after Bonferroni correction. Although the patient-level MAE difference between the two bridge-based methods reached statistical significance, the absolute difference was limited to 0.43 HU. Together with their closely matched performance across complementary image-quality metrics, these results indicate that BBDM remains competitive with the recent diffusion-bridge baseline.

Because DDPM and BBDM were compared under a matched experimental setting, including the same 22.2M-parameter UNet architecture, the same 29,543 training pairs, and the same training budget of 200 epochs with a batch size of 32, this comparison provides a controlled assessment of the effect of the diffusion formulation. Under this matched setting, the substantially lower MAE achieved by BBDM suggests that the source-anchored Brownian-bridge formulation is more effective than conventional Gaussian-noise-to-image diffusion for PET-to-CT translation in our experimental configuration.

4.5. Qualitative Comparison

Figure 3 presents qualitative PET-to-CT synthesis results for four representative test cases. Each row shows one test case, and the columns are organized as PET input, ground-truth CT, BBDM, SelfRDB-L1, DDPM, MedGAN, and Pix2Pix. The PET images are displayed using a heatmap colormap with an SUV window of [0, 15], while all CT images are displayed with a unified soft-tissue window of [−160, 240] HU to ensure visual comparability across methods.

Overall, Pix2Pix and MedGAN show more visible local artifacts and discontinuities, particularly near tissue boundaries and high-contrast anatomical interfaces. DDPM preserves coarse anatomical contours but exhibits more pronounced over-smoothing, with blurred transitions between cortical bone, trabecular bone, and surrounding soft tissue. In contrast, the two bridge-based methods, BBDM and SelfRDB-L1, produce sharper anatomical boundaries and more coherent soft-tissue contrast than DDPM.

Visual differences between BBDM and SelfRDB-L1 are relatively subtle in these representative cases, which is consistent with their closely matched quantitative performance across several image-quality metrics. However, such visual correspondence should not be interpreted as evidence that CT anatomy is uniquely determined by PET, because PET-to-CT translation remains an underdetermined conditional synthesis problem.

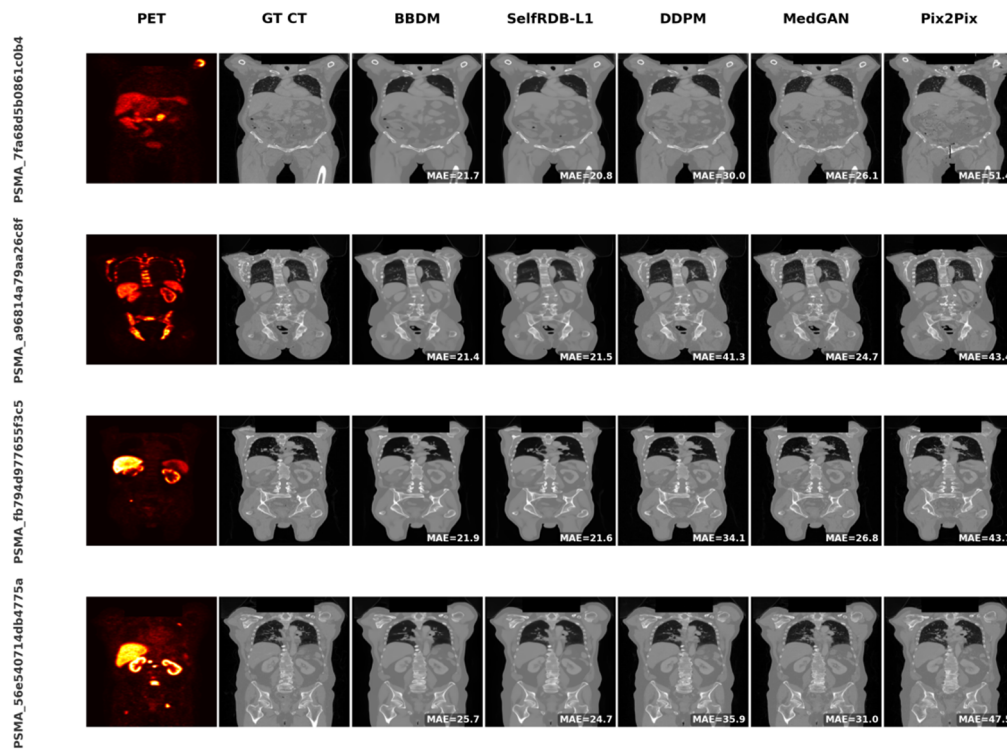


Figure 3. Qualitative comparison of PET-to-CT synthesis results. Each row shows one representative test case. Columns show the PET input, ground-truth CT, and synthetic CT images generated by BBDM, SelfRDB-L1, DDPM, MedGAN, and Pix2Pix. PET images are displayed using a heatmap colormap with an SUV window of [0, 15], and CT images are displayed with a unified soft-tissue window of [−160, 240] HU. The MAE value shown in each synthetic CT panel denotes the slice-level MAE for that example.

4.6. Spatial Distribution Analysis of Error Maps

Qualitative visual comparison is inevitably influenced by observer interpretation. To provide a more localized assessment of reconstruction errors, Figure 4 presents absolute error maps, defined as $|\text{synthetic CT} - \text{ground-truth CT}|$, for the same representative cases shown in Figure 3. Error maps are shown for BBDM, SelfRDB-L1, DDPM, MedGAN, and Pix2Pix using a unified colorbar range of 0–200 HU. Brighter regions indicate larger HU reconstruction errors.

The error maps show clear differences in the spatial distribution of reconstruction errors. Pix2Pix tends to produce localized high-error regions, particularly around high-contrast anatomical boundaries. MedGAN shows fewer extreme localized artifacts than Pix2Pix, but residual errors remain visible in soft-tissue and boundary regions. DDPM exhibits broader moderate errors across anatomical regions, consistent with the over-smoothed reconstructions observed in Figure 3.

In contrast, the two bridge-based methods, BBDM and SelfRDB-L1, show lower and more spatially confined error distributions than DDPM, with residual errors mainly concentrated near tissue boundaries and regions with strong anatomical contrast. The difference between BBDM and SelfRDB-L1 is visually modest, which is consistent with their close quantitative performance in Table 4. These findings suggest that bridge-based diffusion methods provide a more favorable balance between global HU reconstruction accuracy and anatomical boundary preservation than conventional DDPM.

4.7. Controlled Comparison and Mechanistic Analysis of the Bridge Process

4.7.1. Visualization of Intermediate Processes

The quantitative comparison between BBDM and DDPM provides a controlled comparison because the two models use the same 22.2M-parameter UNet backbone, the same patient-disjoint training set, and the same training budget. As shown in Table 4, BBDM achieves a lower slice-level MAE than DDPM (21.88 HU vs. 33.23 HU), showing that the source-anchored Brownian-bridge formulation achieves lower reconstruction error than conventional DDPM under the matched experimental setting.

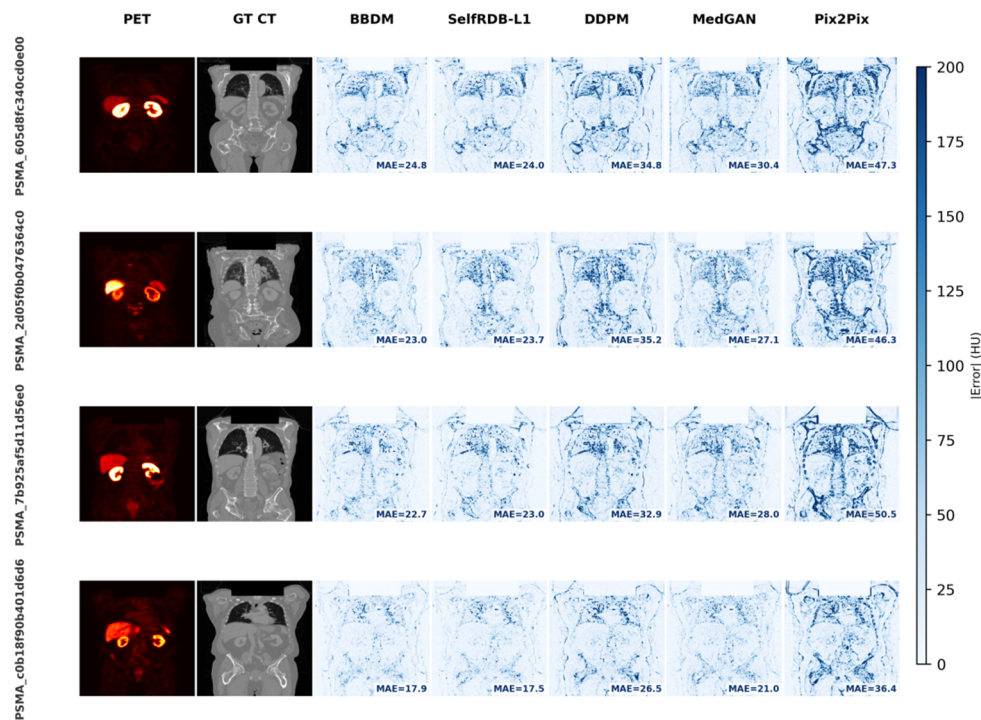


Figure 4. Absolute error maps between synthetic CT and ground-truth CT for the same representative cases as in Figure 3. Error maps are shown for BBDM, SelfRDB-L1, DDPM, MedGAN, and Pix2Pix using a unified colorbar range of 0–200 HU. Brighter regions indicate larger HU reconstruction errors.

Figure 5 further provides a descriptive slice-level visualization of the MAE distribution between BBDM and DDPM. Each point represents one test slice, with DDPM MAE on the horizontal axis and BBDM MAE on the vertical axis. Points below the diagonal indicate lower slice-level MAE for BBDM than for DDPM. This scatter plot is intended as a descriptive visualization of the error distribution, whereas statistical significance was assessed separately using patient-level MAE values to account for within-subject dependence.

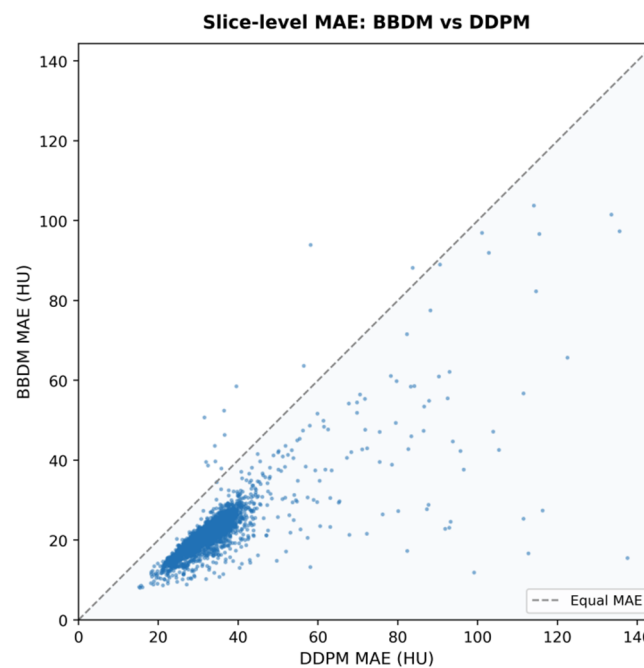


Figure 5. Slice-level MAE scatter comparison between BBDM and DDPM. Each point represents one test slice, with DDPM MAE on the horizontal axis and BBDM MAE on the vertical axis. The dashed diagonal line denotes equal MAE between the two methods. Points below the diagonal indicate lower slice-level MAE for BBDM. Statistical significance was assessed separately using patient-level MAE values.

Figure 6 visualizes representative reverse sampling trajectories for DDPM and BBDM using the same test case. In conventional DDPM, the reverse process is initialized from Gaussian noise and progressively denoises toward the target CT image. In contrast, BBDM is initialized from the PET source endpoint and follows a source-anchored bridge trajectory toward the CT target domain.

The visualization illustrates a qualitative difference between the two diffusion formulations. DDPM remains noise-dominated during the early sampling stages and gradually recovers coarse anatomical structure, whereas BBDM starts from a source image that already contains patient-specific spatial information. As a result, anatomical structures are visually recognizable earlier along the BBDM trajectory. However, this figure is intended as a qualitative visualization of the sampling process rather than a quantitative decomposition of model behavior. The percentages indicate relative sampling progress and should not be interpreted as computational allocation, model capacity allocation, or distinct functional stages.

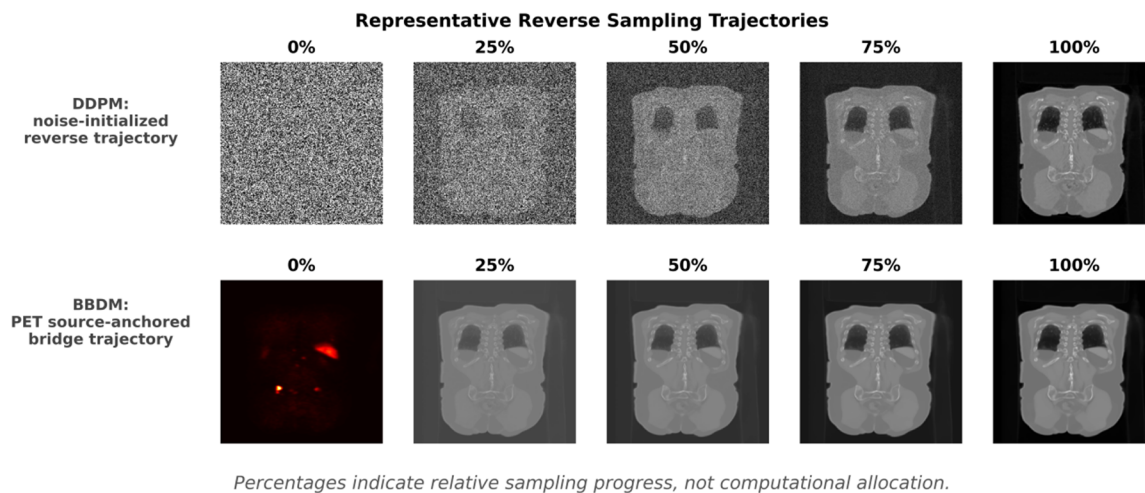


Figure 6. Representative reverse sampling trajectories of DDPM and BBDM. The upper row shows DDPM sampling initialized from Gaussian noise, whereas the lower row shows BBDM sampling initialized from the PET source endpoint. Percentages indicate relative sampling progress along the reverse trajectory and should not be interpreted as a quantitative decomposition of computational cost or model capacity allocation.

4.7.2. Sampling Efficiency Comparison

Figure 7 shows the sensitivity of BBDM and DDPM to the number of reverse sampling steps. BBDM reached stable reconstruction accuracy after approximately 50 steps, whereas DDPM remained substantially worse across the tested step range. Although the lowest BBDM MAE in this analysis was observed at 50 steps, all main quantitative results in Table 4 were reported using the fixed 200-step sampling protocol to maintain a matched comparison with DDPM. These results suggest that the source-anchored bridge formulation is relatively robust to the number of sampling steps and can maintain competitive reconstruction accuracy even with fewer reverse steps.

4.7.3. Bridge Schedule Ablation

Table 5 further examines the choice of the BBDM bridge hyperparameters. The linear schedule $m_t = t/T$ achieves the best MAE (21.9 HU) among all variants. The sinusoidal schedule-defined as $m_t = \sin\left(\frac{\pi t}{2T}\right)$ which provides a smoother transition as t approaches 0 and T -yields an MAE of 23.8 HU (+1.9 HU). The variance-control parameter $\max_var$ causes the MAE to fluctuate only between 21.9 and 22.7 HU over the range 0.5 to 2.0 (a variation of less than 1 HU), indicating relatively low sensitivity to $\max_var$ within the tested range.

Table 5. Bridge schedule ablation.

BBDM Variant	MAE(HU)
m_t Linear, $\max_var = 1.0$	21.88
m_t Sinusoidal, $\max_var = 1.0$	23.8
m_t Linear, $\max_var = 0.5$	22.7
m_t Linear, $\max_var = 2.0$	22.3

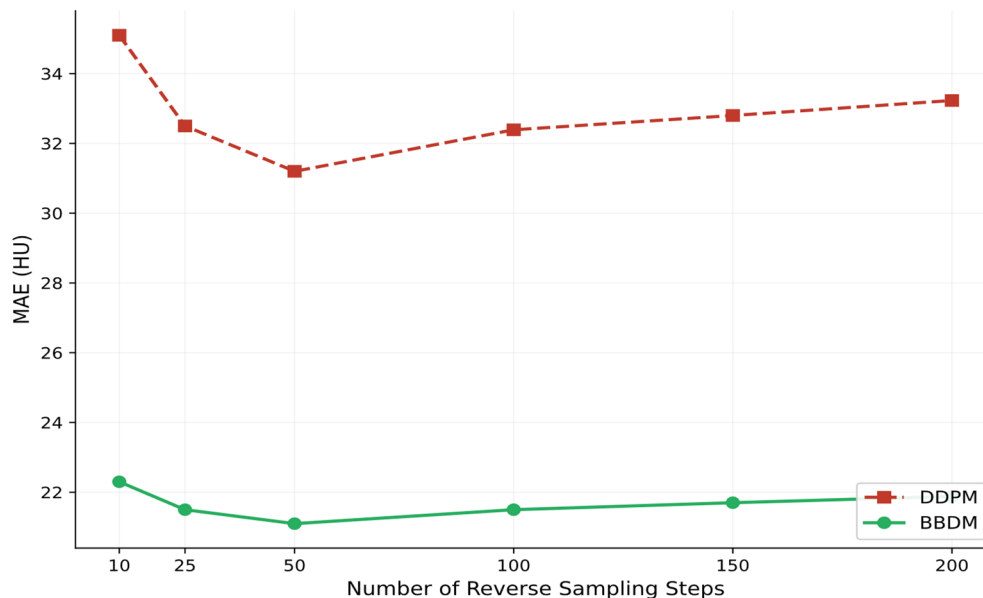


Figure 7. Sampling-step sensitivity comparison between BBDM and DDPM. MAE is shown at different numbers of reverse sampling steps. All main quantitative results in Table 4 were reported using the fixed 200-step sampling protocol for consistency with the matched DDPM comparison.

4.8. Computational Cost Analysis

Table 6 summarizes the computational characteristics of the evaluated methods. The GAN-based methods require only a single forward pass and therefore achieve the shortest inference times, but their MAE values are higher than those of the bridge-based diffusion methods. BBDM and DDPM use the same 22.2M-parameter UNet backbone and the same 200-step DDIM-style sampling procedure, resulting in the same measured inference time of approximately 1.20 s/image. Under this matched setting, BBDM achieves substantially lower MAE than DDPM, supporting the contribution of the source-anchored bridge formulation.

SelfRDB-L1 uses 10 sampling steps and achieves a shorter inference time than BBDM, but it has a larger parameter count of 39.7M. These results indicate that SelfRDB-L1 is a strong and efficient recent bridge-based baseline, while BBDM remains competitive in reconstruction accuracy with fewer parameters.

Table 6. Computational efficiency comparison of the evaluated methods.

Method	Inference Procedure	Sampling Steps/Forward Passes	Inference Time (s/Image)	Parameters (M)	MAE (HU)
BBDM	DDIM-style sampling	200	1.2	22.2	21.88
DDPM	DDIM	200	1.2	22.2	33.23
Pix2Pix	Single forward pass	1	0.02	54.4	42.40
MedGAN	Single forward pass	1	0.03	87.1	25.90
SelfRDB-L1	Bridge-based iterative sampling	10	0.31	39.7	21.56

5. Discussion

5.1. Mechanistic Interpretation of the BBDM–DDPM Difference

The cross-method comparison of LPIPS reveals that perceptual similarity and HU-level reconstruction accuracy are not fully aligned. As shown in Table 4, MedGAN achieved the lowest LPIPS, whereas SelfRDB-L1 and BBDM achieved lower MAE and RMSE. This decoupling indicates that a perceptually favorable texture metric does not necessarily correspond to better HU-level accuracy in PET-to-CT synthesis. Therefore, MAE remains the primary metric for HU-accuracy-sensitive medical image translation, while LPIPS should be interpreted as a complementary perceptual measure.

Synthesizing the analysis in Section 4.7, a mechanistic interpretation can be made as follows. In the conditional DDPM baseline, the denoising network predicts the noise component from the noisy intermediate state x_t while being explicitly conditioned on the PET image y . In the high-noise portion of the DDPM reverse trajectory, x_t contains limited target-domain structural information because the cumulative signal-retention coefficient is small. Nevertheless, the PET image y remains explicitly provided to the conditional denoising

network throughout reverse sampling. As denoising proceeds, coherent anatomical structure gradually emerges while the PET condition continuously contributes to the reconstruction process. Therefore, the transition from coarse structure recovery to PET-conditioned anatomical refinement should be interpreted as a continuous process rather than as two computational stages separated by a specific time-step boundary.

The BBDM formulation mitigates this limitation by initializing the reverse trajectory from the PET source endpoint. Because $x_T \approx y$, the initial bridge state itself already contains source-domain information, rather than being dominated by uninformative Gaussian noise as in conventional DDPM. Thus, in BBDM, the PET information is not only provided as an external conditioning input but is also directly embedded in the evolving bridge state from the beginning of reverse sampling. This source-anchored initialization provides a more direct starting point for PET-conditioned CT reconstruction. The earlier emergence of recognizable anatomical structure in Figure 6 is qualitatively consistent with this mechanism, rather than evidence that individual reverse steps have uniquely defined computational functions.

5.2. Potential Clinical Relevance and Validation Requirements

The MAE of 21.88 HU indicates encouraging image-level agreement between the synthetic CT and the reference CT in the present dataset. However, image-level error alone is insufficient to establish clinical acceptability, because the consequences of synthesis errors depend on their anatomical location and the intended downstream task. The error-map analysis in Section 4.6 further shows that tissue boundaries remain challenging for all evaluated methods, which may be particularly relevant to applications requiring high spatial or quantitative accuracy. Therefore, the present results should be interpreted as evidence of methodological feasibility rather than clinical readiness.

In addition, the deterministic sampling mode used for BBDM (DDIM $\eta = 0$) provides reproducible outputs for a fixed PET input, a property that may be relevant to future longitudinal imaging applications. In oncologic follow-up, the same patient may undergo multiple PET/CT scans to assess treatment response. In this context, reducing inference-time randomness and improving numerical reproducibility may help avoid confounding model-induced variation with true anatomical or density-related changes. Therefore, deterministic sampling provides a practical advantage for longitudinal comparison, although downstream clinical validation remains necessary.

5.3. Limitations and Future Research Directions

This paper has the following limitations, which also point to directions for future work.

First, the pixel-space BBDM operates on 2D slices, which disrupts the inherent 3D volumetric continuity of PET and CT. Although the 2D slice-level MAE has reached 21.88 HU, microscopic inconsistencies may exist between adjacent slices. Extending to 3D-via direct volumetric processing with a 3D UNet or via 2.5D multi-slice inputs-may improve inter-slice consistency and potentially reduce reconstruction error, although at a substantially increased computational cost. This represents an important direction for future work.

Second, the validation of generalizability remains limited. PSMA-targeted tracers exhibit a characteristic biodistribution with physiological uptake in several normal organs, and this distribution differs substantially from that of other tracers such as FDG. Therefore, the generalizability of BBDM across different radiotracers and multi-center datasets requires independent validation. In addition, the PET data in this study are all paired slices from hybrid PET/CT scanners, so the performance of the synthetic CT on standalone PET systems could not be assessed.

Third, the clinical effectiveness of the synthetic CT for downstream tasks such as PET attenuation correction has not yet been evaluated; this is a critical step from method validation to clinical translation. Accordingly, the synthetic CT generated in this study should not be interpreted as a substitute for diagnostic CT, and the present image-level evaluation does not establish clinical equivalence or clinical acceptability.

Fourth, PET-to-CT translation is intrinsically underdetermined because PET and CT encode different physical information. PET primarily reflects tracer-specific functional uptake and therefore does not uniquely determine many anatomical structures or tissue-density variations visible on CT. Consequently, a generative model may produce anatomically plausible CT content that is weakly constrained, or potentially unsupported, by the PET observation. This hallucination risk may be particularly relevant in regions with low tracer uptake, small anatomical structures, tissue boundaries, or pathological changes without a distinctive PET signal. Although the paired-CT comparisons and error maps used in this study quantify deviations from the reference CT, they do not establish whether every locally plausible structure in the synthetic CT is directly supported by the PET input. Future work should therefore incorporate uncertainty estimation, region-specific error analysis, and anatomical-consistency assessment to identify and mitigate unsupported synthetic structures before downstream clinical use is considered.

6. Conclusions

This paper presents a systematic study of Brownian Bridge diffusion for PSMA PET-to-CT translation under a patient-disjoint experimental setting. On the independent patient-disjoint test set, BBDM achieved competitive image-level reconstruction performance across complementary quantitative metrics and substantially reduced MAE compared with the matched DDPM baseline under the same UNet architecture, training data, and training budget. After adding SelfRDB-L1 as a recent diffusion-bridge baseline, the revised results no longer support a claim of uniform superiority for BBDM; instead, they indicate that BBDM remains competitive, while SelfRDB-L1 achieves slightly lower patient-level MAE.

The qualitative comparisons, error maps, and sampling-trajectory visualization suggest that source-anchored bridge diffusion provides a useful formulation for PET-conditioned CT synthesis. However, PET-to-CT translation remains an underdetermined conditional synthesis problem, and the present image-level evaluation does not establish clinical equivalence or diagnostic readiness. Future work should focus on 3D volumetric modeling, multi-tracer and multi-center validation, uncertainty estimation, and downstream evaluation for attenuation correction and other clinical tasks.

Author Contributions

X.G.: software, writing—original draft; T.S.: supervision, methodology, writing—review & editing, Funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 62506243), and the Education Department of Liaoning Provincial Youth Scientist Fund (Grant No. LJ212510144013).

Institutional Review Board Statement

Ethical review and approval were waived for this study because only publicly available, de-identified data from The Cancer Imaging Archive (TCIA) were used. The original dataset publication was approved by the Ethics Committee of the Medical Faculty of LMU Munich and the institutional data security and privacy review board.

Informed Consent Statement

Patient consent was waived for this study because only publicly available, de-identified data from The Cancer Imaging Archive (TCIA) were used, and no direct interaction with patients or collection of new clinical data was involved.

Data Availability Statement

The dataset analyzed in this study is publicly available from The Cancer Imaging Archive (TCIA): <https://www.cancerimagingarchive.net/collection/PSMA-PET-CT-Lesions/>. Access to the dataset is provided according to the TCIA data access policies. No new clinical data were generated in this study.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this manuscript, the authors used ChatGPT (OpenAI) for language editing and proofreading assistance. After using this tool, the authors reviewed and revised the content as necessary and take full responsibility for the accuracy and integrity of the published work.

References

1. Beyer, T.; Townsend, D.W.; Brun, T.; et al. A Combined PET/CT Scanner for Clinical Oncology. *J. Nucl. Med.* **2000**, *41*, 1369–1379.
2. Kinahan, P.E.; Townsend, D.W.; Beyer, T.; et al. Attenuation Correction for a Combined 3D PET/CT Scanner. *Med. Phys.* **1998**, *25*, 2046–2053. <https://doi.org/10.1118/1.598392>.
3. Zaidi, H.; Hasegawa, B.H. Determination of the Attenuation Map in Emission Tomography. *J. Nucl. Med.* **2003**, *44*, 291–315.

4. Huang, B.; Law, M.W.M.; Khong, P.L. Whole-Body PET/CT Scanning: Estimation of Radiation Dose and Cancer Risk. *Radiology* **2009**, *251*, 166–174. <https://doi.org/10.1148/radiol.2511081300>.
5. Alessio, A.M.; Kinahan, P.E.; Cheng, P.M.; et al. PET/CT Scanner Instrumentation, Challenges, and Solutions. *Radiol. Clin. North Am.* **2004**, *42*, 1017–1032. <https://doi.org/10.1016/j.rcl.2004.08.001>.
6. Toyonaga, T.; Shao, D.; Shi, L.; et al. Deep Learning-Based Attenuation Correction for Whole-Body PET—A Multi-Tracer Study with ^{18}F -FDG, ^{68}Ga -DOTATATE, and ^{18}F -Fluciclovine. *Eur. J. Nucl. Med. Mol. Imaging* **2022**, *49*, 3086–3097. <https://doi.org/10.1007/s00259-022-05748-2>.
7. Isola, P.; Zhu, J.Y.; Zhou, T.; et al. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134. <https://doi.org/10.1109/cvpr.2017.632>.
8. Armanious, K.; Jiang, C.; Fischer, M.; et al. MedGAN: Medical Image Translation Using GANs. *Comput. Med. Imaging Graph.* **2020**, *79*, 101684. <https://doi.org/10.1016/j.compmedimag.2019.101684>.
9. Ho, J.; Jain, A.N.; Abbeel, P. Denoising Diffusion Probabilistic Models. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; pp. 6840–6851.
10. Saharia, C.; Chan, W.; Chang, H.; et al. Palette: Image-to-Image Diffusion Models. In Proceedings of the ACM SIGGRAPH 2022 Conference Proceedings, Vancouver, BC, Canada, 7–11 August 2022; pp. 1–10. <https://doi.org/10.1145/3528233.3530757>.
11. Li, B.; Xue, K.; Liu, B.; et al. BBDM: Image-to-Image Translation with Brownian Bridge Diffusion Models. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 1952–1961. <https://doi.org/10.1109/cvpr52729.2023.00194>.
12. Wolterink, J.M.; Dinkla, A.M.; Savenije, M.H.F.; et al. Deep MR to CT Synthesis Using Unpaired Data. In *Simulation and Synthesis in Medical Imaging (SASHIMI)*, Québec City, QC, Canada, 10 September 2017; Springer: Cham, Switzerland, 2017; pp. 14–23. https://doi.org/10.1007/978-3-319-68127-6_2.
13. Lei, Y.; Jeong, J.J.; Wang, T.; et al. MRI-Based Pseudo CT Synthesis Using Anatomical Signature and Alternating Random Forest with Iterative Refinement Model. *J. Med. Imaging* **2018**, *5*, 043504. <https://doi.org/10.1117/1.jmi.5.4.043504>.
14. Dar, S.U.H.; Yurt, M.; Karacan, L.; et al. Image Synthesis in Multi-Contrast MRI with Conditional Generative Adversarial Networks. *IEEE Trans. Med. Imaging* **2019**, *38*, 2375–2388. <https://doi.org/10.1109/tmi.2019.2901750>.
15. Hu, Z.; Li, Y.; Zou, S.; et al. Obtaining PET/CT Images from Non-Attenuation Corrected PET Images in a Single PET System Using Wasserstein Generative Adversarial Networks. *Phys. Med. Biol.* **2020**, *65*, 215010. <https://doi.org/10.1088/1361-6560/aba5e9>.
16. Huang, X.; Liu, M.Y.; Belongie, S.; et al. Multimodal Unsupervised Image-to-Image Translation. In *Computer Vision—ECCV 2018, Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018*; Springer: Cham, Switzerland, 2018; pp. 179–196. https://doi.org/10.1007/978-3-030-01219-9_11.
17. Özbey, M.; Dalmaz, O.; Dar, S.U.H.; et al. Unsupervised Medical Image Translation with Adversarial Diffusion Models. *IEEE Trans. Med. Imaging* **2023**, *42*, 3524–3539. <https://doi.org/10.1109/tmi.2023.3290149>.
18. Atli, O.F.; Kabas, B.; Arslan, F.; et al. I2I-Mamba: Multi-Modal Medical Image Synthesis via Selective State Space Modeling. *IEEE Trans. Biomed. Eng.* **2026**, 1–12. <https://doi.org/10.1109/TBME.2026.3698056>.
19. Upadhyay, U.; Chen, Y.; Hepp, T.; et al. Uncertainty-Guided Progressive GANs for Medical Image Translation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021, Proceedings of the 24th International Conference, Strasbourg, France, 27 September–1 October 2021*; Springer: Cham, Switzerland, 2021; pp. 614–624. https://doi.org/10.1007/978-3-030-87199-4_58.
20. Meng, C.; He, Y.; Song, Y.; et al. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In Proceedings of the Tenth International Conference on Learning Representations (ICLR), Online, 25–29 April 2022.
21. Choo, K.; Jun, Y.; Yun, M.; et al. Slice-Consistent 3D Volumetric Brain CT-to-MRI Translation with 2D Brownian Bridge Diffusion Model. In Medical Image Computing and Computer Assisted Intervention (MICCAI) 2024, Marrakesh, Morocco, 6–10 October 2024; Springer: Cham, Switzerland, 2024; pp. 657–667. https://doi.org/10.1007/978-3-031-72104-5_63.
22. He, Q.; Summerfield, N.; Wang, P.; et al. Deterministic Medical Image Translation via High-Fidelity Brownian Bridges. *arXiv* **2025**, arXiv:2503.22531.
23. Shiri, P.; Yi, X.; Mistry, N.P.; et al. Generating Synthetic Contrast-Enhanced Chest CT Images from Non-Contrast Scans Using Slice-Consistent Bridge Diffusion Network. In *Statistical Atlases and Computational Models of the Heart. Regular and CMRxRecon Challenge Papers, Proceedings of the 16th International Workshop, STACOM 2025, Daejeon, South Korea, 27 September 2025*; Springer: Cham, Switzerland, 2026; pp. 160–170. https://doi.org/10.1007/978-3-032-17734-6_16.
24. Arslan, F.; Kabas, B.; Dalmaz, O.; et al. Self-Consistent Recursive Diffusion Bridge for Medical Image Translation. *Med. Image Anal.* **2025**, *106*, 103747. <https://doi.org/10.1016/j.media.2025.103747>.
25. Song, J.; Meng, C.; Ermon, S. Denoising Diffusion Implicit Models. In Proceedings of the International Conference on Learning Representations (ICLR), Online, 3–7 May 2021.

26. Jeblick, K.; Schachtner, B.; Mittermeier, A.; et al. A Whole-Body PSMA-PET/CT Dataset with Manually Annotated Tumor Lesions. *Sci. Data* **2026**, *13*, 1023. <https://doi.org/10.1038/s41597-026-07821-z>.
27. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; et al. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. <https://doi.org/10.1109/tip.2003.819861>.
28. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multi-Scale Structural Similarity for Image Quality Assessment. In Proceedings of the 37th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 9–12 November 2003; IEEE; pp. 1398–1402. <https://doi.org/10.1109/ACSSC.2003.1292216>.
29. Zhang, R.; Isola, P.; Efros, A.A.; et al. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595. <https://doi.org/10.1109/cvpr.2018.00068>.