Tsinghua Science and Technology
Doi: 10.26599/TST.2026.90100027

Research Article

Supervise-assisted Multi-modality Fusion Diffusion Model for PET Restoration

Yingkai Zhang†, Shuang Chen†, Ye Tian, Yunyi Gao, Jianyong Jiang, and Ying Fu (✉)

Abstract Positron emission tomography (PET) offers powerful functional imaging but involves radiation exposure. Efforts to reduce this exposure by lowering the radiotracer dose or scan time can degrade image quality. While using magnetic resonance (MR) images with clearer anatomical information to restore standard-dose PET (SPET) from low-dose PET (LPET) is a promising approach, it faces challenges with the inconsistencies in the structure and texture of multi-modality fusion, as well as the mismatch in out-of-distribution (OOD) data. In this paper, we propose a supervise-assisted multi-modality fusion diffusion model (MFdiff) for addressing these challenges for high-quality PET restoration. Firstly, to fully utilize auxiliary MR images without introducing extraneous details in the restored image, a multi-modality feature fusion module is designed to learn an optimized fusion feature. Secondly, using the fusion feature as an additional condition, high-quality SPET images are iteratively generated based on the diffusion model. Furthermore, we introduce a two-stage supervise-assisted learning strategy that harnesses both generalized priors from simulated in-distribution datasets and specific priors tailored to in-vivo OOD data. Experiments demonstrate that the proposed MFdiff effectively restores high-quality SPET images from multi-modality inputs and outperforms state-of-the-art methods both qualitatively and quantitatively.

Keywords Positron emission tomography (PET), Diffusion Models, Image Restoration, Multi-modality

• Yingkai Zhang, Yunyi Gao, and Ying Fu are with the Beijing Institute of Technology, Beijing, 100081, China (e-mail: zhangyingkai@bit.edu.cn; yiiclass@qq.com; fuying@bit.edu.cn).
• Shuang Chen is with the Research and Development Center of Agricultural Bank of China, Beijing, 100073, China (e-mail: chenshuangxd@foxmail.com).
• Ye Tian is with the Peking University, Beijing, 100871, China (e-mail: tye@pku.edu.cn).
• Jianyong Jiang is with the Beijing Normal University, Beijing, 100872, China (e-mail: jianyong@bnu.edu.cn).
† Equal Contribution.
(✉) To whom correspondence should be addressed.
Manuscript received: 2025-07-07; revised: 2026-01-05; accepted: 2026-02-06

1 Introduction

Positron emission tomography (PET) is a powerful molecular imaging technique, crucial in clinical diagnosis and scientific research. It is extensively employed for diagnosing and staging cancer, screening for cardiovascular diseases, researching brain function and neurological diseases, and in drug development [1–3]. Although PET is a highly versatile functional imaging modality, the substantial radiation exposure from PET tracers limits its use and raises significant health concerns. Researchers follow the principle of "as low as reasonably achievable" (ALARA) [4] when injecting radiotracers, attempting to minimize radiation doses to patients. However, a lower injection dose significantly reduces the number of acquired photon counts, resulting in a lower signal-to-noise ratio (SNR) and compromising PET images, which could potentially impact the final clinical diagnosis [5]. Two examples of paired low-dose PET (LPET) and standard-dose PET (SPET) from phantom and in-vivo datasets are shown in Fig. 1(a-b).

In recent years, benefiting from the continuous development and advancement of algorithms in the field of computer vision [6–11], various deep learning-based methods have been widely applied in the field of medical image processing [12–15]. They employ end-to-end learning to transform low-dose (LD) into standard-dose (SD) PET images, to restore high-quality PET images. Prior work has primarily focused on LD/SD PET image pairs, but the information available from LD PET images is limited, which presents a bottleneck in SD PET image restoration. The workflow is depicted in I of Fig. 1(c). With the rising demand for early and accurate disease diagnosis, advanced medical imaging devices such as PET/MR scanners have been widely used in clinical examinations [16]. As observed in Fig. 1(a-b), MR images provide clear anatomical details and superior tissue differentiation, potentially enhancing PET restoration quality. Consequently, several approaches [17, 18] have attempted to supply neural networks with anatomical information from simultaneously acquired MR images via additional input channels to aid PET restoration. Compared to using LPET images alone, multi-modality input has proven to significantly enhance the restoration quality of SPET images, as illustrated in II-III of Fig. 1(c). However, sig-

Just Accepted

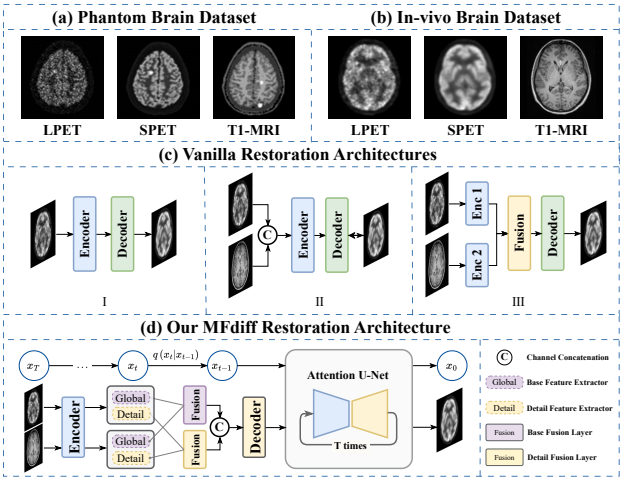

Fig. 1: (a-b) Two examples of T1-MRI, LPET, and the corresponding SPET images from the phantom and in-vivo dataset. (c) The categorization of existing PET restoration methods. (d) Our MFdiff restoration method.

nificant data deviations between modalities may result in the inadvertent retention of anatomical and structural information in the restored SPET images, leading to undesirable residual artifacts [19]. Therefore, we aim to effectively utilize deep semantic features from anatomical guidance images to enhance image quality while carefully avoiding the introduction of extraneous patterns and structural details.

From the perspective of medical image acquisition, large datasets are often unavailable publicly due to concerns related to patient privacy, ethics, and data sharing protocols. In medical imaging, particularly with PET scans, concerns about harmful ionizing radiation make it nearly impossible to acquire significant numbers of paired low-dose and standard-dose data from volunteers, resulting in small-scale PET datasets. Moreover, the complexity of medical imaging processes means that even slight variations in settings can affect the characteristics of acquired PET images, leading to out-of-distribution (OOD) data. Factors such as the type and dosage of radiotracers, scan time, hardware, and anatomical changes influence these outcomes [20]. Data-driven deep learning methods, typically trained in specific scenarios, perform well on in-distribution (ID) data. However, they often face a performance bottleneck due to limited information and cannot directly adapt to out-of-distribution (OOD) data, resulting in comparatively low image quality that challenges the maintenance of clinically applicable performance. Due to (i) limited dataset size and difficulty in accessing sufficient LD/SD data pairs, (ii) performance degradation of dataset-specific models in handling OOD data, enhancing the generalization capability of models has become a critical challenge in medical imaging applications. Consequently, we aim to maintain high restoration quality across specific clinical acquisition variations (e.g., single center, scanner, and tracer) without relying on large-scale paired clinical datasets for each specific scenario.

In this paper, we propose a novel supervise-assisted multi-modality fusion diffusion model (MFdiff) with a multi-modality feature fusion module to restore SPET images from LPET and corresponding MR images acquired through joint PET/MR imaging. The framework is illustrated in Fig. 1(d) and Fig. 2. The multi-modality feature fusion module primarily consists of a dual-branch intra-modality learning (IML) module and a cross-modality aggregation (CMA) module. Initially, the IML module extracts global and detailed features based on the global attention and local convolution from each modality. Subsequently, the CMA module learns an optimized fusion feature representation, which enhances restoration quality and minimizes the impact of modality mismatch. Furthermore, we adopt a two-stage supervise-assisted learning strategy that circumvents the need for large-scale paired data in the target clinical domain. Our model is trained on a substantial corpus of paired synthetic data and fine-tuned on a limited amount of domain-specific in vivo data. This strategy allows the network to learn generalized priors from simulated in-distribution datasets while tailoring specific priors to in vivo OOD data. Consequently, the model demonstrates improved adaptability to acquisition-based OOD shifts, thereby enhancing transfer learning efficiency. The contributions of our work are as follows:

(1) We present a conditional diffusion restoration model based on fusion of the PET/MR to iteratively generate high-quality SPET images.
(2) We propose a multi-modality feature fusion module to fully extract and utilize both the specific and shared information of LPET and MR images. We adopt a two-stage supervise-assisted learning strategy with generalized priors and specific priors for high-quality PET restoration.
(3) Extensive experiments demonstrate that our proposed method outperforms existing state-of-the-art approaches qualitatively and quantitatively.

The remainder of this paper is organized as follows. Section II reviews related work. The proposed method is detailed in Section III. We introduce the simulated and in-vivo data along with experimental settings in Section IV, experimental results in Section V, and conclusion in Section VI.

Just Accepted

2 Related Work

In PET scanning, it is customary to decrease the injection dose of radioactive tracer or reduce the scanning time in order to reduce radiation exposure or increase throughput. These adjustments inevitably lead to increased noise in PET images. Various methods have been proposed to restore SPET images from LPET images, which can be broadly classified into two categories: (i) within-restoration denoising techniques and (ii) post-restoration denoising techniques.

2.1 Within-Restoration Denoising Techniques

Within-restoration denoising techniques aim to restore PET images from raw data in list-mode or sinogram domain while simultaneously suppressing noise. To achieve this goal, model-based image restoration methods typically rely on the forward model of the PET system and the designed priors to construct the objective function. For instance, Chun *et al.* [21] proposed using block-based non-local regularization and employing alternating direction method of multipliers (ADMM) as the solver. Häggström *et al.* [13] proposed a direct deep learning based restoration method, named DeepPET, which uses an end-to-end network to learn the mapping of sinogram data to PET images without any imaging system information. It forces the network to learn complex mappings between different domains, which is beyond the range of its abilities, thus harming the performance. To tackle these issues, Mehranian *et al.* [22] proposed a forward-backward splitting algorithm that unfolds networks into maximum a posteriori (MAP) PET image restoration and utilizes MR images as additional input to guide PET denoising. These within-restoration denoising techniques require access to raw list-mode or sinogram data, which is vendor-specific and frequently excluded from large open-access datasets, thereby limiting their practical applications.

2.2 Post-Restoration Denoising Techniques

Compared to within-restoration denoising techniques, post-restoration methods that focus on the image domain can benefit from the support of most medical imaging platforms and can be easily and flexibly integrated into existing restoration pipelines. Besides, rapid advancements in image restoration tasks [23, 24] within the natural image domain, such as denoising [25–28], have significantly accelerated the development of medical image processing (PET). Initially, filtering-based methods [29] were used to enhance PET image quality. However, these methods are mainly focused on image denoising rather than predicting the SPET image from the LPET image. Xiang *et al.* [17] were the first to apply CNN to the field of PET restoration. They proposed an auto-context strategy to recover SPET images directly from both LPET images and the corresponding MR images. However, methods based on encoder-decoder CNN are typically trained for specific tasks, which limits the generalization of the model across various settings. Furthermore, images processed by CNN frequently exhibit issues such as blurred edges and over-smoothing.

In recent years, generative adversarial networks (GAN) [30] have been widely used in the field of image restoration. Compared with methods based on encoder-decoder CNN, they generate images by learning the data distribution, making them more adaptable when dealing with various imaging scenarios. Ouyang *et al.* [31] presented a GAN-based restoration approach for ultra-low-dose amyloid PET images. Geng *et al.* [32] proposed a content-noise complementary learning strategy, utilizing two predictors to learn content and noise complementarily, and implemented based on GAN. While GAN-based methods can rapidly generate high-quality samples, they often suffer from poor mode coverage, lack of sample diversity, and training instability, leading to issues like mode collapse, vanishing gradients, and convergence difficulties.

Fortunately, the diffusion model has been proposed to address the limitations of GAN-based methods, offering high sampling quality, extensive pattern coverage, and sample diversity. It excels in learning complex data distributions, such as in medical imaging [33–35], *etc.* Previous methods often directly utilize raw multimodal data for image processing. For instance, Gong *et al.* [33] employed Denoising Diffusion Probabilistic Models (DDPM), using PET and MR images as conditional inputs for denoising. Yoon *et al.* [34] utilized a Score-based Residual Diffusion Model to learn the residual between paired images. A significant drawback of these direct approaches is that the forward and reverse diffusion processes can inadvertently introduce irrelevant artifacts or lose critical details, potentially compromising clinical diagnostic accuracy. Moreover, the generalizability of diffusion models to OOD PET data from diverse acquisition environments and the effective utilization of such OOD data are critical challenges that must be addressed to develop robust and clinically viable PET restoration methods.

3 Supervise-assisted Multi-modality Fusion Diffusion Model

The proposed MFdiff is depicted in Fig. 2(a), including a multi-modality feature fusion module and a conditional diffusion restoration module. Suppose that given the LPET $x_P \in \mathbb{R}^{H\times W}$ and the reference MRI $x_M \in \mathbb{R}^{H\times W}$. H, W

Just Accepted

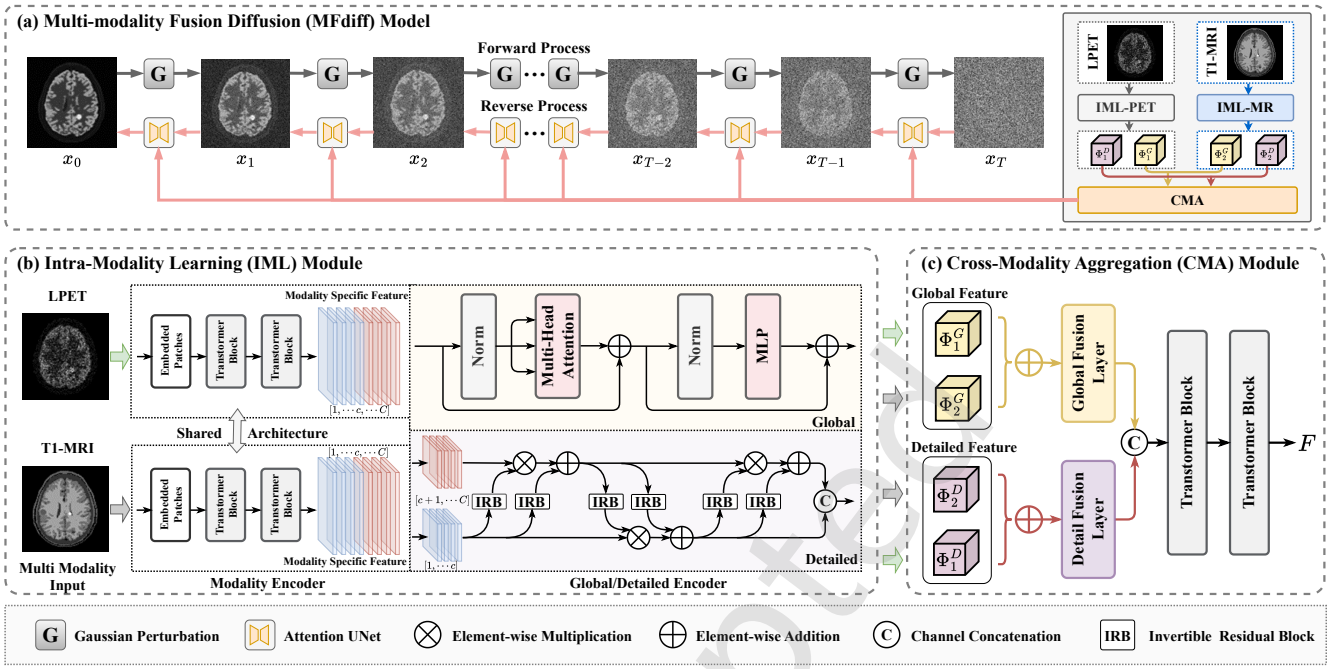

Fig. 2: Fig. 2 Overview of our supervise-assisted multi-modality fusion diffusion model (MFdiff). (a) provides the framework of MFdiff, which includes two main components: a conditional diffusion restoration module and a multi-modality feature fusion module. The details of the latter module are illustrated in (b) intra-modality learning module with modality ecoder and global/detailed encoder, and (c) cross-modality aggregation module.

are the number of rows and columns. We aim to restore SPET x_S by:

$$x_S = f_\theta(x_P, x_M), \tag{1}$$

where $x_S \in \mathbb{R}^{H\times W}$ and $f_\theta(\cdot)$ denotes the multi-modality fusion diffusion model (MFdiff). Specifically, the multi-modality feature fusion module (detailed in Sec. 3.1) aims to extract and fuse cross-modality image features by the local and global feature perception capabilities of CNN and Transformer. The conditional diffusion restoration module (detailed in Sec. 3.2) aims to restore high-quality SPET images based on diffusion model. In Sec. 3.3, we introduce a two-stage supervise-assisted learning strategy to address the lack of real paired data, and OOD data frequently encountered in real clinical settings.

3.1 Multi-modality Feature Fusion Module

Data disparities exist between PET and MR modalities, accompanied by challenges in effective message interaction. Existing methods, which often directly concatenate data, fail to capitalize on the information from both modalities and may even lead to the underutilization or damage of LPET image. To address these challenges in inter-modal data interaction, we propose the multi-modality feature fusion module shown in Fig. 2(a), which comprises an Intra-Modality Learning (IML) module and a Cross-Modality Aggregation (CMA) module.

3.1.1 Intra-Modality Learning (IML) Module

In PET image denoising, anatomical information from the corresponding MR image is expected to enhance the quality of denoised SPET images. However, it is crucial to ensure that detailed information from MR image, which mismatches with PET, is not introduced. Therefore, it is necessary to exploit the potential correlation between the two modalities while capturing modality-specific information, to enhance the features of the target modality and suppress mismatched features from the auxiliary modalities. To achieve this goal, we propose the IML module, as shown in Fig. 2(b). Firstly, to extract modality specific features from LPET x_P and MRI x_M, we adopt the 'Modality Encoder' [36]. It consists of two independent Transformer encoders with the same shared architecture but different parameters, and can be expressed as:

$$\begin{aligned} F_P &= f_{\theta_p}(x_P), \\ F_M &= f_{\theta_m}(x_M), \end{aligned} \tag{2}$$

where f_{θ_p} and f_{θ_m} are two modality-specific networks.

Subsequently, to enhance the global features of PET and

Just Accepted

MR images while distinguishing their respective detailed features, we divide PET and MR images into PET global feature $G_P \in \mathbb{R}^{C\times H\times W}$, PET detailed feature $D_P \in \mathbb{R}^{C\times H\times W}$, MR global feature $G_M \in \mathbb{R}^{C\times H\times W}$, and MR detailed feature $D_M \in \mathbb{R}^{C\times H\times W}$. By reinforcing the consistency of high-level global features between PET and MR, and differentiating the detailed information of each modality, we ensure that the final fused features effectively assist the generation of SPET images. To achieve this, we propose a dual-branch feature extractor ('Global/Detailed Encoder') to further extract global and detailed features from the shallow features of LPET and MRI. The global features G_P and G_M are extracted as:

$$\begin{aligned} G_i' &= MHA(LN(F_i)) + F_i, \\ G_i &= MLP(LN(G_i')) + G_i', \end{aligned} \tag{3}$$

where $i = \{P, M\}$ represents the input modality. MHA denotes the multi-head attention, MLP is the multilayer perceptron [36], and LN is the layer normalization. The global branch based on Transformer with spatial self-attention extracts global features G_P and G_M from the shallow features $F_P \in \mathbb{R}^{C\times H\times W}$ and $F_M \in \mathbb{R}^{C\times H\times W}$ of each modality. Specifically, the global feature G_P from the low-quality LPET image exhibits a relatively blurry structure and edges, while the global feature G_M from the high-quality MR image captures clear and precise global anatomical information.

As PET image applications are sensitive to fine details, we strive for lossless detail feature extraction to preserve as much valuable information from the original LPET and MR images as possible. Due to the capability of the invertible neural network [37] to achieve precise inverse mapping from output to input, thus ensuring information integrity, it is employed as a lossless feature extraction block to further extract detailed features D_P and D_M from the shallow features of both modalities. The detail feature extraction branch initially divides the shallow features $F_i \in \mathbb{R}^{C\times H\times W}$ along the channel dimension into two parts, $F_{i,f} \in \mathbb{R}^{C/2\times H\times W}$ and $F_{i,l} \in \mathbb{R}^{C/2\times H\times W}$. It then computes the final detailed features through three affine coupling layers, with the first layer being expressed as:

$$\begin{aligned} D_{i,f}^1 &= F_{i,f} \odot \exp\left(IRB\left(F_{i,l}\right)\right) + IRB\left(F_{i,l}\right), \\ D_{i,l}^1 &= F_{i,l}, \end{aligned} \tag{4}$$

where $IRB(\cdot)$ is the invertible residual block used in MobileNetV2 [38], and $\odot$ denotes the Hadamard product. After three layers of transformation, D_i is obtained by concatenating two sets of detailed features along the channel dimension. The design of the independent Detailed Encoder prevents the direct transfer of modality-specific textures. The shared-weight encoders could force different modalities into a common latent space, potentially causing the high-frequency texture of the MRI to overwrite the lower-resolution PET signal. Our independent parameterization ($f_{\theta_p} \neq f_{\theta_m}$) ensures that the detailed features D_P and D_M retain their distinct modality characteristics. This allows the subsequent Cross-Modality Aggregation module to selectively filter inconsistent anatomical edges (pseudo-structures) that appear in the MR image but do not correspond to metabolic activity in the PET image.

3.1.2 *Cross-Modality Aggregation (CMA) Module*

To integrate the high-level features extracted by modality-specific modules, we input the global and detailed features from both the target and auxiliary modalities into a cross-modality aggregation module shown in Fig. 2(c). To enhance the deep connections between PET and MR images, we employ a dual-fusion strategy. Specifically, we adopt a global fusion layer and a detail fusion layer to process the corresponding features from both modalities. This process yields the fused global feature $G_{fuse} \in \mathbb{R}^{C\times H\times W}$ and the fused detailed feature $D_{fuse} \in \mathbb{R}^{C\times H\times W}$. The architecture of the global fusion layer mirrors that of the global feature extraction branch, while the detail fusion layer aligns with the detail feature extraction branch. Subsequently, the decoder network, which includes two Transformer blocks, maps high-level features back to the original input size and outputs the final fused feature $F \in \mathbb{R}^{C\times H\times W}$.

3.2 Conditional Diffusion Restoration Module

Once the extracted fusion feature has been obtained, we aim to utilize it for SPET image restoration. Due to the powerful generative capability of diffusion models, *e.g.*, DDPM [39], to learn complex data distributions and sample effectively to produce high-quality images, we adopt it to restore high-quality SPET images. We inject the fusion feature F (termed as x_{fusion}) into the proposed conditional diffusion restoration module as a condition. By incorporating fusion features from LPET and MR images as conditional constraints, the diffusion model is able to generate complex details in SPET images more accurately.

3.2.1 *Forward Process*

Assuming the target data $x_0 \sim q\left(x_0\right)$, the forward diffusion process can be considered as a Markov chain. By incrementally adding Gaussian noise to the target data as follow, a sequence of latent noise variables $x_1, x_2, ...x_t$ can be generated:

$$q\left(x_1, .., x_T \mid x_0\right) = \prod_{t=1}^{T} q\left(x_t \mid x_{t-1}\right), \forall t \in \{1, \ldots, T\} \tag{5}$$

Just Accepted

where $q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$ represents a single-step Markov process:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t} \cdot x_{t-1}, \beta_t \cdot I\right), \quad (6)$$

where t represents the number of diffusion steps, $\{\beta_t\}_{t=1}^{T}$ denotes the variance schedule of Gaussian noise across these steps, with $0 < \beta_1 < \beta_2 < ... < \beta_t < 1$. This schedule is used to control the amount of noise added at each step. For a predefined variance schedule, when the number of diffusion steps T is sufficiently large, the resulting distribution $q(x_T) \sim \mathcal{N}(x_T; 0, I)$ becomes stationary.

As defined on a Markov chain, a key characteristic of the diffusion process is the ability to directly sample data x_t at any time step t based on the original data x_0. By defining $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=0}^{t} \alpha_i$, the sampling process can be expressed as:

$$\begin{aligned} q(x_t \mid x_0) &= \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t} x_0, (1-\bar{\alpha}_t) \cdot \mathbf{I}\right), \\ x_t &= \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \end{aligned} \quad (7)$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

3.2.2 *Reverse Process*

The reverse process is the inverse of the forward process, which generates target data by gradually denoising noisy samples. To enhance the restoration of SPET images using an auxiliary MR image, we integrate fusion features as additional conditions into the inverse process. As shown in Fig. 2, alongside the noisy sample x_t and time step t are simultaneously input into the denoising attention UNet at each step. To fully leverage the information from both the original LPET x_P and the fusion feature x_{fusion} with the MR guided, we also supply these components at each step to guide the denoising process. It alters the noise prediction model from $\epsilon_\theta(x_t, t)$ to $\epsilon_\theta(x_t, t, x_P, x_{fusion})$, adding two inputs as conditions. The fusion feature x_{fusion} (denoted as F in Fig. 2(c)), the LPET image x_P, and the noisy state x_t are concatenated along the channel dimension to form the input to the first layer of the diffusion U-Net. For an input sample x_t with t steps noise added, the objective of the noise prediction model is to minimize the distance between the predicted noise and actual added noise. Then, with the trained noise prediction model $\epsilon_{\hat{\theta}}(x_t, t, x_P, x_{fusion})$, each refinement step in the reverse process is:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left[x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_{\hat{\theta}}(x_t, t, x_P, x_{fusion})\right] + \sigma_t z, \quad (8)$$

where $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

3.3 Two-stage Supervise-assisted Learning

Acquiring PET images poses significant challenges, often resulting in small clinical PET datasets. Variability in scanning times, tracer types, injection doses, and methods customized for individual patient needs contribute to the generation of out-of-distribution (OOD) data. Deep learning approaches that train and test under uniform conditions excel on test sets mirroring the training data distribution but falter when confronted with real-world clinical variations. These methods struggle with low-dose PET images and simulations, leading to notable performance degradation when encountering diverse OOD data.

To address these challenges, we introduce a two-stage supervise-assisted learning strategy with the loss in Eq. (9). (**i**) In the first stage, *i.e.*, **external stage**, the model MFdiff is trained on a large-scale simulated phantom dataset. We leverage the simulated data to learn multimodal feature fusion and high-quality PET image generation, thereby avoiding the high costs associated with extensive clinical data usage. Using the diffusion model, we supervise the learning of noise at each prediction step. The loss function can be expressed as:

$$\mathcal{L} = E_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t, x_P, x_{fusion})\|^2\right]. \quad (9)$$

Thus, we derive generalized priors from the large-scale simulated dataset and initialize the model parameters. (**ii**) In the second stage, *i.e.*, **internal stage**, we further train the initialized model on a small set of clinically realistic in-vivo data to learn specific priors for in-vivo OOD data. The supervised training approach in this stage aligns with Eq. (9), allowing us to further discern specific differences through OOD data, which enhances its adaptability across various scenarios with minimal practical overhead.

Learning details. During the training in the external stage, the model processes data from the simulated dataset, comprising 1000 samples from 20 brain phantoms, distributed across training, validation, and test sets at ratios of 0.8, 0.15, and 0.05, respectively. In the internal stage, we conduct training on a small cohort of clinical data, encompassing varied scanning times and injection dosages.

4 Experiments

In this section, we first introduce our experimental settings, including the comparison methods, evaluation metrics, and implementation details. We then introduce the simulated phantom dataset and three in-vivo out-of-distribution datasets used for experiments.

Just Accepted

4.1 Experimental Settings

4.1.1 Comparison Methods

In order to verify the superiority of our method, the following 6 state-of-the-art image restoration or denoising methods are implemented for comparison, including M-UNet [40], FB-SEM [22], EA-GAN [41], Hi-Net [42], CNCL [32], and CSRD [34]. Following the work [43], we implement the networks that are originally set to 2.5D or 3D in the comparison methods as 2D networks (changing the input dimension to match our slice-based approach). We also add MR images as additional input, ensuring all models have access to the same multi-modal information. For all comparison methods, we adopt the hyperparameters (*e.g.*, loss weighting) mentioned in the respective papers and train the models with the same settings (*e.g.*, iterations, learning rates, and batch sizes) as ours.

4.1.2 Evaluation Metrics

To evaluate and compare the performance of different methods comprehensively, three typical evaluation metrics are used, including peak signal to noise ratio (PSNR) [44], structural similarity (SSIM) [45] and normalized mean square error (NMSE) [46]. The larger PSNR and SSIM values the better the restored images are, while the smaller values of NMSE indicate better restoration.

4.1.3 Implementation Details

The proposed method is implemented through the Pytorch framework, trained with an NVIDIA GeForce RTX 3090 GPU. The denoising network used in the experiment originated from the UNet architecture with time embedding modules and self-attention layers. During network training, the size of input LPET and auxiliary MR image are both 192×192, the batch size is 4, and the number of time step is 1000. If not mentioned otherwise, the same hyperparameters as those in the work [39] are used. We train the model for 50,000 iterations in the external stage and 20,000 iterations in the internal stage, using the AdamW optimizer [47] with learning rate of 0.0001.

4.2 Simulated Phantom Data

We leverage the BrainWeb [48, 49], which has 20 3D brain phantoms to construct the simulated dataset. For PET data, by modeling the acquisition geometry of the Siemens Biograph mMR, the simulated PET data has a voxel size of $2.08 \times 2.08 \times 2.03 mm^3$ and an image resolution of $344 \times 344 \times 127$. For each 3D brain phantom, we select 10 noncontinuous slices in the middle and perform data augmentation. Each slice is rotated along the axial direction by 5 random angles in the range 0-15, resulting in 1000 data slices. For each data slice, a high-definition SD sinogram and a low-definition LD sinogram are generated. For SD sinograms, we adopt 1×10^{10} counts and point spread function modeling using 2.5 mm full-width at half-maximum (FWHM) Gaussian kernels. The OSEM algorithm [50] is then used to restore SPET images, and the number of iterations and subsets are set to 15 and 10, respectively. Similarly, for LD sinograms, randomly select count levels between 9×10^7 and 12×10^7, with 4 mm FWHM Gaussian kernels. To restore SPET images, the OSEM algorithm is used with 10 iterations and 14 subsets. The split of 1000 slices from 20 brain phantoms is performed at the subject (phantom) level before slicing. This ensures that slices from the same phantom do not appear in both training and testing, preventing subject-level leakage.

4.3 In-Vivo Data

To verify and compare the generalization performance of all methods, we adopt in-vivo out-of-distribution (OOD) datasets in three different imaging scenarios, including (1) variation in imaging time, (2) variation in radiotracer injection dose, and (3) variation in radiotracer administration protocol. The following details the acquisition, data preprocessing, and data augmentation of OOD datasets.

4.3.1 OOD Data with Variation in Imaging Time

For the OOD data-1, we utilize an open-access dataset[51] acquired from a Siemens (Erlangen) Biograph 3-Tesla molecular MR (mMR) scanner at Monash Biomedical Imaging in Melbourne, Australia. Participants include 27 healthy volunteers aged 18-23 years. They are perfused with ^{18}F-Flurodeoxyglucose (^{18}F-[FDG]), with average dose 233MBq and at a rate of 36 mL/hour. Scans take 95 minutes. Non-functional MR scans are performed 30 minutes after perfusion, corresponding to T1-MR images with FOV = $256 \times 256\ mm^2$, voxel size = $1 \times 1 \times 1\ mm^3$, 176 slices. Afterwards, 3600 s of list-mode PET data for each subject starting at the 30-minute time point are merged into 225 3D sinogram frames, with a single PET frame originating from a scan of 16 s. The 3D PET image is restored through the OSEM algorithm with 3 iterations and 21 subsets, and then a 5 mm FWHM Gaussian post-processing filter is applied to each 3D volume, and all 3D PET volumes are spliced into a 4D volume with a size of $344 \times 344 \times 127 \times 225$ in time series.

The SPM12 toolbox* is used to register the PET images to the T1-MR images, and then all image pixel intensities are normalized to range of $[0, 1]$. For each subject, we select 10 spaced slices in the third dimension from 130 to 150, and

*https://www.fil.ion.ucl.ac.uk/spm/software/spm12/

Just Accepted

construct SPET images using a 30-minute scan. By reducing the acquisition time, LPET images with two settings of 2min scan and single 16s scan are constructed. For each slice, the same data augmentation method as simulated data is used.

4.3.2 *OOD Data with Variation in Radiotracer Injection Dose*

The MR/PET data acquisition protocol for OOD data-2 is the same as that for OOD data-1, *i.e.*, using the same mMR scanner, radiotracer, and imaging settings, *etc.*, but with different injection doses. For the LD-PET data, the average injected dose is 225 MBq and average PSNR is 23.74 dB, and for the ultra low-dose PET (uLD-PET) data, the average injected dose is 167 MBq and average PSNR is 21.35 dB. The data preprocessing and data augmentation methods of OOD data-2 are consistent with those used for OOD data-1.

4.3.3 *OOD Data with Variation in Radiotracer Administration Protocol*

The data acquisition protocol for OOD data-3 is the same as that for OOD data-1 and data-2. However, the radiotracer administration protocol varies, incorporating three forms: bolus, constant infusion, and hybrid 50% bolus-50% infusion. For all protocols, the bolus or infusion is administered concurrently with the start of the PET scan, and the infusion rate is set as 36 mL/hour. For both infusion and hybrid bolus-infusion administration, the infusion is discontinued at 55 minutes. For additional details on data acquisition and procedures, refer to work [52]. We select 300 pairs of typical data (6 subjects) for all in-vivo OOD data tested in our experiments, including 50 pairs (1 subject) for OOD data-1, 100 pairs (2 subjects) for OOD data-2, and 150 pairs (3 subjects) for OOD data-3. For each type of data, we adopt 250 pairs of data (5 subjects) for internal stage training and validation (10%).

5 Results

5.1 Simulation Study

Table 1 presents the quantitative results on the phantom brain dataset in the form of $u \pm v$, where u and v represent the average value and the standard deviation value of each evaluation metric, respectively. From Table 1, it can be observed that our MFdiff outperforms all compared methods in PSNR, SSIM, and NMSE. This indicates that MFdiff can effectively utilize information from different modalities to achieve high-quality SPET image restoration.

In Fig. 3, we select two representative slices with lesions for visual comparison, and display the restored SPET images along with error maps by different methods shown in Figs. 3(a-b). Notably, the upper circular lesion (indicated by the green arrow in Fig. 3(a)) represents a critical test case for MR-induced artifact suppression. In this region, the anatomical MRI (first column) displays healthy tissue structure, which contradicts the metabolic lesion present in the ground-truth SPET. Comparison methods like M-UNet and FBSEM rely on direct concatenation or naive fusion, which inadvertently transfers healthy MRI textures into the PET image. This interference causes the blurring or partial removal of lesions, effectively acting as a false negative. In contrast, MFdiff correctly preserves the shape and intensity of lesions. This validates that the IML module effectively disentangles the modalities, utilizing MRI for global anatomical guidance without hallucinating MRI-specific structures where they are metabolically inconsistent.

Table 1: Quantitative results (Mean±Standard Deviation) associated with state-of-the-art PET image restoration methods in terms of PSNR, SSIM and NMSE on Phantom brain dataset. **Bold** and Underline indicate the best and second-best method for each quantitative evaluation, respectively.

Method	PSNR (dB) ↑	SSIM ↑	NMSE ↓
LPET	21.13 ± 2.71	0.839 ± 0.041	0.176 ± 0.055
M-UNet [40]	31.35 ± 2.59	0.981 ± 0.007	0.018 ± 0.009
FBSEM [22]	34.93 ± 2.03	0.959 ± 0.009	0.039 ± 0.021
EA-GAN [41]	32.09 ± 2.10	0.962 ± 0.036	0.015 ± 0.009
Hi-Net [42]	32.65 ± 2.15	0.981 ± 0.008	0.013 ± 0.006
CNCL [32]	33.32 ± 1.78	0.984 ± 0.004	0.016 ± 0.009
CSRD [34]	37.28 ± 1.94	0.990 ± 0.006	0.009 ±0.005
MFdiff (Ours)	**37.93 ± 1.85**	**0.991 ± 0.008**	**0.008 ± 0.004**

5.2 Clinical Study

Clinical real data often exhibit significant distributional differences due to varied operational practices, so the ability of models to adapt to such distributionally inconsistent data, *i.e.*, generalization, is an important basis for measuring the value of its practical application. We conduct experiments on in-vivo OOD data generated in three different situations with the model trained on simulated datasets and further on a modest amount of clinical real data. In clinical study, FBSEM is excluded from comparison because it requires sinogram domain data, which is not present in these datasets.

Tables 2-4 illustrate the quantitative restoration results across changes in imaging time, radiotracer injection dose, and radiotracer administration protocol on three in-vivo OOD datasets. The results demonstrate that our MFdiff consistently achieves higher quantitative scores compared to other methods in all three scenarios. Specifically, for the 2-minute scan setting in Table 2, M-UNet, EA-GAN, Hi-Net, and CNCL enhance the visual quality of restoration compared to low-dose PET images, as depicted in Fig. 4. Despite these visual

Just Accepted

Table 2: Quantitative results (Mean±Standard Deviation) associated with state-of-the-art PET image restoration methods in terms of PSNR, SSIM and NMSE on OOD Data-1.

Method	16s scanning-PET			2min scanning-PET		
	PSNR (dB) ↑	SSIM ↑	NMSE ↓	PSNR (dB) ↑	SSIM ↑	NMSE ↓
LPET	25.01 ± 1.21	0.928 ± 0.004	0.037 ± 0.009	34.65 ± 1.24	0.987 ± 0.001	0.004 ± 0.002
M-UNet [40]	27.04 ± 1.86	0.949 ± 0.004	0.013 ± 0.006	32.57 ± 0.92	0.977 ± 0.001	0.003 ± 0.001
EA-GAN [41]	27.16 ± 1.68	0.970 ± 0.003	0.019 ± 0.007	33.83 ± 1.14	0.949 ± 0.008	0.005 ± 0.001
Hi-Net [42]	27.95 ± 1.86	0.970 ± 0.002	0.020 ± 0.009	30.02 ± 1.81	0.985 ± 0.001	0.013 ± 0.006
CNCL [32]	28.77 ± 1.54	0.974 ± 0.002	0.014 ± 0.005	34.36 ± 1.24	0.976 ± 0.007	0.004 ± 0.001
CSRD [34]	29.90 ± 1.39	0.973 ± 0.014	0.012 ± 0.004	36.05 ± 0.98	0.987 ± 0.005	**0.002 ± 0.001**
MFdiff (Ours)	**30.67 ± 1.37**	**0.974 ± 0.011**	**0.010 ± 0.003**	**36.97 ± 0.98**	**0.989 ± 0.005**	**0.002 ± 0.001**

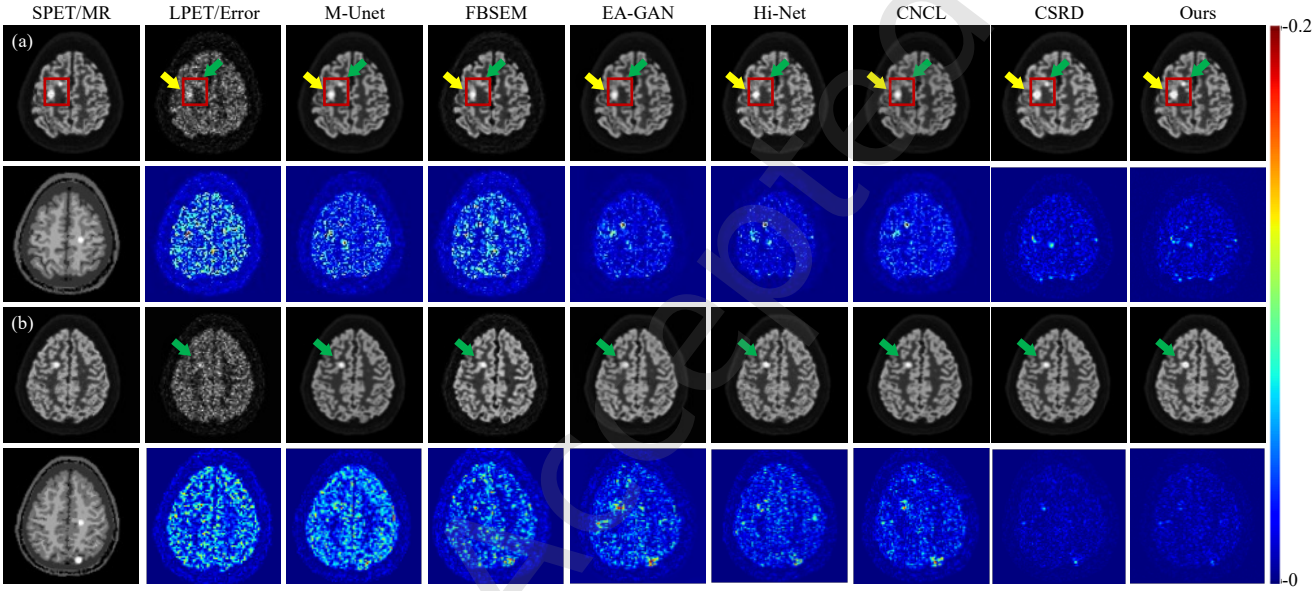

Fig. 3: Visual comparison results of the restored PET along with error maps by different methods on phantom dataset.

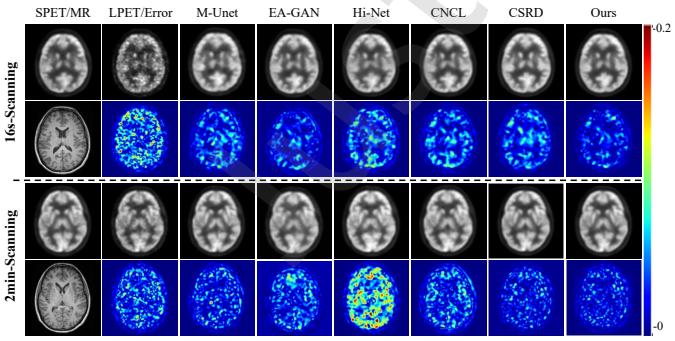

Fig. 4: Visual comparison along with error maps of different methods on OOD data-1 with variation in imaging time.

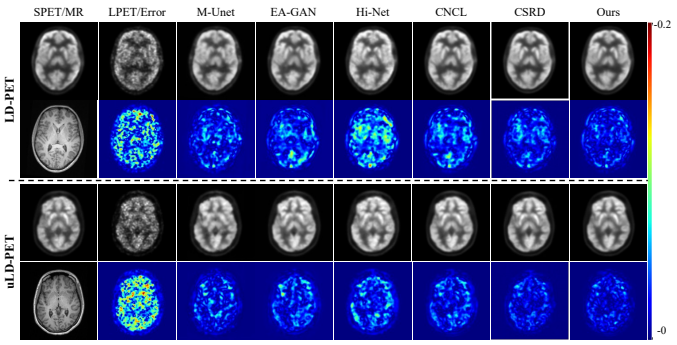

Fig. 5: Visual comparison along with error maps of different methods on OOD data-2 with variation in radiotracer injection dose.

enhancements, the quantitative performance indices for these methods show some deterioration compared to LPET images. Notably, the average PSNR for MFdiff method under 2-minute setting improves from 34.65 dB to 36.97 dB by 0.92 dB over the second-best method (CSRD), demonstrating the adaptability. Figs. 4-6 show the qualitative results of all methods on three in-vivo OOD datasets, respectively. The results illustrate that MFdiff predicts missing information in LPET images more accurately, resulting in restored images with clearer edges, structures, and rich texture details. From the error maps, it can be observed that MFdiff achieves

Not Accepted

Table 3: Quantitative results (Mean±Standard Deviation) associated with state-of-the-art PET image restoration methods in terms of PSNR, SSIM and NMSE on OOD Data-2.

Method	LD-PET			uLD-PET		
	PSNR (dB) ↑	SSIM ↑	NMSE ↓	PSNR (dB) ↑	SSIM ↑	NMSE ↓
LPET	21.35 ± 1.16	0.880 ± 0.008	0.084 ± 0.021	23.74 ± 0.92	0.905 ± 0.008	0.046 ± 0.009
M-UNet [40]	27.89 ± 0.65	0.945 ± 0.003	0.009 ± 0.002	26.22 ± 1.57	0.936 ± 0.010	0.013 ± 0.005
EA-GAN [41]	27.33 ± 0.85	0.941 ± 0.004	0.017 ± 0.003	27.88 ± 1.35	0.933 ± 0.004	0.016 ± 0.005
Hi-Net [42]	24.67 ± 1.02	0.948 ± 0.012	0.038 ± 0.010	27.80 ± 1.31	0.964 ± 0.003	0.020 ± 0.007
CNCL [32]	26.35 ± 1.10	0.954 ± 0.004	0.020 ± 0.004	27.34 ± 1.54	0.956 ± 0.004	0.018 ± 0.006
CSRD [34]	28.48 ± 1.29	**0.966 ± 0.012**	0.015 ± 0.004	29.69 ± 0.99	0.965 ± 0.009	**0.009 ± 0.002**
MFdiff (Ours)	**28.96 ± 1.25**	0.965 ± 0.012	**0.014 ± 0.004**	**30.37 ± 1.02**	**0.972 ± 0.005**	0.010 ± 0.002

Table 4: Quantitative results (Mean±Standard Deviation) associated with state-of-the-art PET image restoration methods in terms of PSNR, SSIM and NMSE on OOD Data-3.

Method	Bolus (B)			Infusion (I)			Bolus + Infusion (B/I)		
	PSNR (dB) ↑	SSIM ↑	NMSE ↓	PSNR (dB) ↑	SSIM ↑	NMSE ↓	PSNR (dB) ↑	SSIM ↑	NMSE ↓
LPET	19.92 ± 0.56	0.867 ± 0.004	0.097 ± 0.012	20.03 ± 1.35	0.886 ± 0.008	0.119 ± 0.030	24.45 ± 2.44	0.940 ± 0.011	0.092 ± 0.036
M-UNet [40]	27.36 ± 1.05	0.892 ± 0.039	0.018 ± 0.005	30.03 ± 1.17	0.893 ± 0.053	0.012 ± 0.003	30.33 ± 2.04	0.899 ± 0.077	0.028 ± 0.027
EA-GAN [41]	27.47 ± 1.30	0.878 ± 0.053	0.018 ± 0.006	29.88 ± 0.68	0.853 ± 0.077	0.012 ± 0.003	28.34 ± 2.08	0.916 ± 0.034	0.017 ± 0.009
Hi-Net [42]	27.32 ± 1.26	0.929 ± 0.036	0.019 ± 0.010	29.76 ± 0.82	0.950 ± 0.014	0.013 ± 0.003	26.45 ± 3.27	0.901 ± 0.057	0.127 ± 0.201
CNCL [32]	27.89 ± 0.80	0.898 ± 0.027	0.015 ± 0.003	29.32 ± 1.17	0.888 ± 0.040	0.014 ± 0.003	27.35 ± 1.68	0.932 ± 0.022	0.019 ± 0.006
CSRD [34]	27.88 ± 0.91	0.939 ± 0.017	0.015 ± 0.004	28.86 ± 1.07	0.960 ± 0.014	0.015 ± 0.004	29.13 ± 1.07	0.941 ± 0.014	**0.017 ± 0.005**
MFdiff (Ours)	**28.23 ± 0.71**	**0.940 ± 0.009**	**0.014 ± 0.003**	**30.41 ± 0.93**	**0.967 ± 0.009**	**0.010 ± 0.002**	**31.92 ± 2.35**	**0.949 ± 0.032**	0.019 ± 0.016

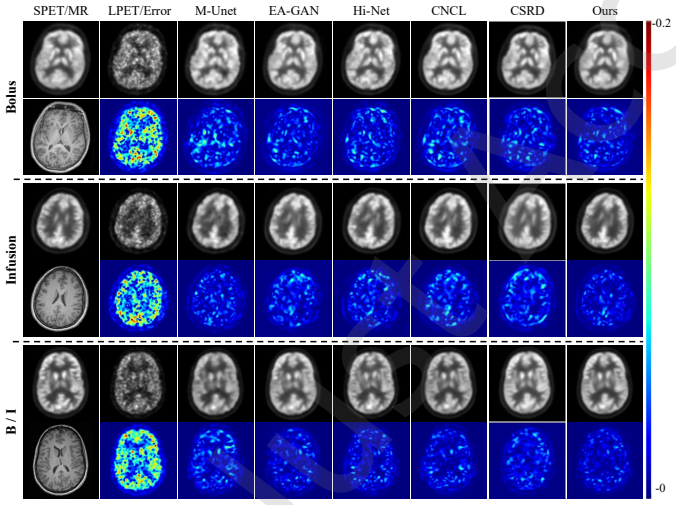

Fig. 6: Visual comparison along with error maps of different methods on OOD data-3 with variation in radiotracer administration protocols.

the smallest restoration errors across all in-vivo OOD data scenarios, indicating that the restored images closely match those of LPET images.

5.3 Ablation Study

The proposed MFdiff comprises several key components. We conduct the following ablation experiments to evaluate the contribution of each module separately: (1) UNet: Using UNet [53] as the restoration network with single LPET image as input; (2) MUNet: Introducing multi-modalily input and employing direct concatenation as the fusion strategy; (3-5) Replacing naive fusion strategy with the multi-modality feature fusion module consisting only of global feature extraction branch, only of detail feature extraction branch and the complete dual-branch, denoted as GMUNet, DMUNet, and CMUNet, respectively; (6) Employing the proposed MFdiff with conditional denoising diffusion probabilistic model (CDDPM) as the restoration module for SPET restoration. Table 5 shows the quantitative evaluation results for the restored SPET images when comparing our full model with its various ablated versions.

(i) Multi-modality input. Compared to UNet, MUNet adds MR image as additional input to guide the SPET restoration. The naive concatenation fusion strategy alone increases PSNR by 6.96 dB, SSIM by 0.057, and decreases NMSE by 0.032, indicating that the effectiveness of utilizing multi-modality inputs for this task. **(ii) Feature extraction module.** To effectively extract and aggregate complementary information from two modalities, we propose a multi-modality feature fusion module. To investigate the effect of this module as a whole and its crucial component, we compare four ablated versions: MUNet, GMUNet, DMUNet, and CMUNet. To ensure the network parameters of different versions are similar, GMUNet replaces the detail feature extraction branch with a global feature extraction branch, and similarly, DMUNet replaces the global branch with a detail branch. The results in Table 5

Just Accepted

Table 5: Quantitative results (Mean±Standard Deviation) of the ablation study over the Phantom sataset.

Method	Multi-input	Global Feature	Detailed Feature	CDDPM	PSNR (dB) ↑	SSIM ↑	NMSE ↓
UNet	-	-	-	-	26.39 ± 1.39	0.924 ± 0.009	0.050 ± 0.009
MUNet	✓	-	-	-	33.35 ± 2.59	0.981 ± 0.007	0.018 ± 0.009
GMUNet	✓	✓	-	-	36.39 ± 2.59	0.985 ± 0.006	0.009 ± 0.005
DMUNet	✓	-	✓	-	35.93 ± 1.56	0.987 ± 0.005	0.009 ± 0.004
CMUNet	✓	✓	✓	-	36.51 ± 2.59	0.984 ± 0.011	0.009 ± 0.005
MFdiff (Ours)	✓	✓	✓	✓	**37.93 ± 1.85**	**0.991 ± 0.008**	**0.008 ± 0.004**

Table 6: Quantitative Results of the Ablation Study of two-stage supervise-assisted learning strategy on OOD data-1. External

Settings	Learning	PSNR (dB) ↑	SSIM ↑	NMSE ↓
16s Scanning PET	External	21.91 ± 3.05	0.718 ± 0.208	0.107 ± 0.140
	Internal	29.81 ± 1.74	**0.975 ± 0.008**	0.013 ± 0.005
	External+Internal	**30.67 ± 1.37**	0.974 ± 0.011	**0.010 ± 0.003**
2min Scanning PET	External	31.35 ± 2.59	0.981 ± 0.007	0.018 ± 0.009
	Internal	35.63 ± 0.95	0.984 ± 0.009	0.003 ± 0.001
	External+Internal	**36.97 ± 0.98**	**0.989 ± 0.005**	**0.002 ± 0.001**

Table 7: Comparison Results of the single-stage (mixed training) and our two-stage learning strategy on OOD data-1 (2 min scanning PET).

Settings	Learning	PSNR (dB) ↑	SSIM ↑	NMSE ↓
CSRD [34]	Single-Stage	35.17 ± 0.90	0.984 ± 0.002	0.005 ± 0.002
	Two-Stage	36.05 ± 0.98	0.987 ± 0.005	0.002 ± 0.001
MFdiff (Ours)	Single-Stage	35.89 ± 0.85	0.983 ± 0.002	0.003 ± 0.001
	Two-stage	**36.97 ± 0.98**	**0.989 ± 0.005**	**0.002 ± 0.001**

demonstrate that incorporating different versions of the fusion module leads to enhancing PET restoration compared to the concatenation. Moreover, CMUNet outperforms the others, indicating that both global and detail branches are beneficial for enhancing the restoration quality, and they mutually reinforce each other. **(iii) Diffusion module.** We replace the basic UNet with the diffusion module, CDDPM to compose our MFdiff. The results indicate that the CDDPM leads to increases of PSNR from 36.51 dB to 37.93 dB, demonstrating that CDDPM outperforms the commonly used UNet architecture for medical imaging, and exhibiting a stronger ability for restoring high-quality images. **(iv) Learning strategy.** As shown in Table 6, it can be seen that a model only trained with simulation data (external stage) performs poorly on OOD data-1 under various experimental conditions. Using our proposed two-stage supervised-assisted learning strategy can improve the performance, compared to training only with external or internal data, and enhance its ability to generalize to specific out-of-distribution (OOD) data. **(v) Learning Stage.** As shown in Table 7, the 'Single-Stage (Mixed Training)' strategy based on MFdiff achieves a PSNR of 35.89 dB. It is superior to 'Internal Only' (35.63 dB) but inferior to our proposed 'Two-Stage' strategy (36.97 dB). The performance change is also the same on the CSRD model. Therefore, we can see that the large-scale simulated data allows the model to learn stable generalized priors (noise distribution and basic anatomical fusion) without the interference of the domain gap. The second stage then fine-tunes these priors to the specific in-vivo distribution. In contrast, the 'Single-Stage (Mixed Training)' strategy forces the model to simultaneously reconcile the domain shift between simulated physics and real scanner noise. It leads to sub-optimal convergence where the model struggles to prioritize the subtler clinical features against the bulk of simulated data.

6 Conclusion

In this paper, we propose a novel supervise-assisted multi-modality fusion diffusion model (MFdiff) to restore high-quality SPET images from LPET and corresponding MR images. With multi-modality fusion module, specific and complementary features from different modalities can be better extracted and effectively aggregated to generate fusion feature to facilitate SPET restoration. The high-quality SPET image restoration module with the powerful data learning and generation capabilities of diffusion model, utilizes multi-modality fusion feature as additional condition to iteratively generate high-quality SPET images. Additionally, we introduce a two-stage supervision-assisted learning strategy. In the first external stage, the model is trained on a simulated phantom dataset to capture generalized priors. Subsequently, in the second internal stage, it is fine-tuned on a limited amount of in vivo data. It effectively enables the learning of both generalized and domain-specific priors for SPET restoration. This approach reduces reliance on large multi-modality in-vivo datasets for the target domain and improves adaptability of the model to varying acquisition constraints.

The results on simulated dataset and in-vivo OOD datasets show that MFdiff outperforms existing state-of-the-art PET restoration methods in both qualitative and quantitative metrics. It indicates that our method can restore high-quality SPET images more effectively and has better robustness

Just Accepted

against reductions in scan time and dosage. Furthermore, in visual comparisons on the simulated phantom dataset, our method effectively integrates multi-modality information and restores the structure of simulated lesion areas more accurately than competing methods. Our clinical evaluation is limited to healthy volunteers, meaning real pathological uptake was not directly assessed. However, we observe improved structural fidelity in simulated lesions and low error rates in complex healthy anatomy. These results suggest that MFdiff effectively preserves critical details without introducing MR-induced artifacts.

Limitation. Our method does have limitations that require further improvement. First, the in-vivo evaluation is conducted on a limited cohort of 6 healthy subjects using a single scanner and a single tracer. While the model demonstrates data efficiency by fine-tuning on a small internal dataset, the generalization to multi-center data, different PET scanners, or different radiotracers remains to be validated. In the future, we plan to construct additional datasets with diverse settings to explore their impact on the transferability of the pre-trained model. Second, our future work will extend to downstream tasks based on the restored SPET images, such as disease diagnosis and medical image segmentation. Finally, given the time and cost constraints associated with multi-modality acquisition, our forthcoming studies will also include research on missing modalities and unpaired data.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (62331006, 62505006), and the Fundamental Research Funds for the Central Universities.

Reference

[1] A. Alavi, J. W. Kung, and H. Zhuang, Implications of pet based molecular imaging on the current and future practice of medicine, *Seminars in Nuclear Medicine*, vol. 34, no. 1, pp. 56–69, 2004.

[2] J. K. Willmann, N. Van Bruggen, L. M. Dinkelborg, and S. S. Gambhir, Molecular imaging in drug development, *Nature reviews Drug discovery*, vol. 7, no. 7, pp. 591–607, 2008.

[3] S. R. Cherry, T. Jones, J. S. Karp, J. Qi, W. W. Moses, and R. D. Badawi, Total-body pet: maximizing sensitivity to create new opportunities for clinical research and patient care, *Journal of Nuclear Medicine*, vol. 59, no. 1, pp. 3–12, 2018.

[4] S. D. Voss, G. H. Reaman, S. C. Kaste, and T. L. Slovis, The alara concept in pediatric oncology, *Pediatric radiology*, vol. 39, pp. 1142–1146, 2009.

[5] J. Liu, M. Malekzadeh, N. Mirian, T.-A. Song, C. Liu, and J. Dutta, Artificial intelligence-based image enhancement in pet imaging: Noise reduction and resolution enhancement, *PET clinics*, vol. 16, no. 4, pp. 553–576, 2021.

[6] Y. Tian, Y. Fu, and J. Zhang, Transformer-based under-sampled single-pixel imaging, *Chinese Journal of Electronics*, vol. 32, no. 5, pp. 1151–1159, 2023.

[7] T. Zhang, Y. Fu, and J. Zhang, Deep guided attention network for joint denoising and demosaicing in real image, *Chinese Journal of Electronics*, vol. 33, no. 1, pp. 303–312, 2024.

[8] Y. Gao, Q. Liu, L. Gu, and Y. Fu, Grayscale-assisted rgb image conversion from near-infrared images, *Tsinghua Science and Technology*, vol. 30, no. 5, pp. 2215–2226, 2025.

[9] Y. Zhang, Z. Lai, T. Zhang, Y. Fu, and C. Zhou, Unaligned rgb guided hyperspectral image super-resolution with spatial-spectral concordance, *International Journal of Computer Vision*, vol. 133, no. 9, pp. 6590–6610, 2025.

[10] L. Chen, Y. Fu, L. Gu, C. Yan, T. Harada, and G. Huang, Frequency-aware feature fusion for dense image prediction, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 763–10 780, 2024.

[11] Y. Zou, Y. Fu, T. Takatani, and Y. Zheng, Eventhdr: From event to high-speed hdr videos and beyond, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 1, pp. 32–50, 2024.

[12] Q. Huang, Y. Zhou, L. Tao, W. Yu, Y. Zhang, L. Huo, and Z. He, A chan-vese model based on the markov chain for unsupervised medical image segmentation, *Tsinghua Science and Technology*, vol. 26, no. 6, pp. 833–844, 2021.

[13] I. Häggström, C. R. Schmidtlein, G. Campanella, and T. J. Fuchs, Deeppet: A deep encoder–decoder network for directly solving the pet image reconstruction inverse problem, *Medical Image Analysis*, vol. 54, pp. 253–262, 2019.

[14] B. Yang, Y. Yang, Q. Li, D. Lin, Y. Li, J. Zheng, and Y. Cai, Classification of medical image notes for image labeling by using minbert, *Tsinghua Science and Technology*, vol. 28, no. 4, pp. 613–627, 2023.

[15] L. Zhang, K. Zhang, and H. Pan, Sunet++: A deep network with channel attention for small-scale object segmentation on 3d medical images, *Tsinghua Science and Technology*, vol. 28, no. 4, pp. 628–638, 2023.

[16] D. A. Torigian, H. Zaidi, T. C. Kwee, B. Saboury, J. K. Udupa, Z.-H. Cho, and A. Alavi, Pet/mr imaging: technical aspects and potential clinical applications, *Radiology*, vol. 267, no. 1, pp. 26–44, 2013.

[17] L. Xiang, Y. Qiao, D. Nie, L. An, W. Lin, Q. Wang, and D. Shen, Deep auto-context convolutional neural networks for standard-dose pet image estimation from low-dose pet/mri, *Neurocomputing*, vol. 267, pp. 406–416, 2017.

[18] C. O. da Costa-Luis and A. J. Reader, Micro-networks for robust mr-guided low count pet imaging, *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 5, no. 2, pp. 202–212, 2020.

[19] Y. Onishi, F. Hashimoto, K. Ote, H. Ohba, R. Ota,

Just Accepted

E. Yoshikawa, and Y. Ouchi, Anatomical-guided attention enhances unsupervised pet image denoising performance, *Medical Image Analysis*, vol. 74, p. 102226, 2021.

[20] R. Accorsi, J. S. Karp, and S. Surti, Improved dose regimen in pediatric pet, *Journal of Nuclear Medicine*, vol. 51, no. 2, pp. 293–300, 2010.

[21] S. Y. Chun, Y. K. Dewaraja, and J. A. Fessler, Alternating direction method of multiplier for tomography with nonlocal regularizers, *IEEE Transactions on Medical Imaging*, vol. 33, no. 10, pp. 1960–1968, 2014.

[22] A. Mehranian and A. J. Reader, Model-based deep learning pet image reconstruction using forward–backward splitting expectation–maximization, *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 5, no. 1, pp. 54–64, 2020.

[23] M. Li, Y. Fu, T. Zhang, J. Liu, D. Dou, C. Yan, and Y. Zhang, Latent diffusion enhanced rectangle transformer for hyperspectral image restoration, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[24] H. Li, Z. Wu, R. Shao, T. Zhang, and Y. Fu, Noise calibration and spatial-frequency interactive network for stem image enhancement, in *Proc. of the conference on Computer Vision and Pattern Recognition*, 2025, pp. 21 287–21 296.

[25] Y. Zhang, T. Zhang, J. Nie, and Y. Fu, Real noise decoupling for hyperspectral image denoising, *arXiv preprint arXiv:2511.17196*, 2025.

[26] Y. Zou, Y. Fu, Y. Zhang, T. Zhang, C. Yan, and R. Timofte, Calibration-free raw image denoising via fine-grained noise estimation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[27] T. Zhang, Y. Fu, and J. Zhang, Guided hyperspectral image denoising with realistic data, *International Journal of Computer Vision*, vol. 130, no. 11, pp. 2885–2901, 2022.

[28] K. Wei, Y. Fu, Y. Zheng, and J. Yang, Physics-based noise modeling for extreme low-light photography, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8520–8537, 2021.

[29] J. Yan, J. C.-S. Lim, and D. W. Townsend, Mri-guided brain pet image filtering and partial volume correction, *Physics in medicine & biology*, vol. 60, no. 3, p. 961, 2015.

[30] W. Du and S. Tian, Transformer and gan-based super-resolution reconstruction network for medical images, *Tsinghua Science and Technology*, vol. 29, no. 1, pp. 197–206, 2023.

[31] J. Ouyang, K. T. Chen, E. Gong, J. Pauly, and G. Zaharchuk, Ultra-low-dose pet reconstruction using generative adversarial network with feature matching and task-specific perceptual loss, *Medical Physics*, vol. 46, no. 8, pp. 3555–3564, 2019.

[32] M. Geng, X. Meng, J. Yu, L. Zhu, L. Jin, Z. Jiang, B. Qiu, H. Li, H. Kong, J. Yuan *et al.*, Content-noise complementary learning for medical image denoising, *IEEE Transactions on Medical Imaging*, vol. 41, no. 2, pp. 407–419, 2021.

[33] K. Gong, K. Johnson, G. El Fakhri, Q. Li, and T. Pan, Pet image denoising based on denoising diffusion probabilistic model, *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 51, no. 2, pp. 358–368, 2024.

[34] S. Yoon, M. Tivnan, R. Hu, Y. Wang, Y.-d. Son, D. Wu, X. Li, K. Kim, and Q. Li, Volumetric conditional score-based residual diffusion model for pet/mr denoising, in *Proc. of the Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 754–763.

[35] Y. Shi, A. Abulizi, H. Wang, K. Feng, N. Abudukelimu, Y. Su, and H. Abudukelimu, Diffusion models for medical image computing: a survey, *Tsinghua Science and Technology*, vol. 30, no. 1, pp. 357–383, 2024.

[36] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, Restormer: Efficient transformer for high-resolution image restoration, in *Proc. of the conference on Computer Vision and Pattern Recognition*, 2022, pp. 5728–5739.

[37] L. Dinh, J. Sohl-Dickstein, and S. Bengio, Density estimation using real nvp, in *Proc. Int. Conf. Learn. Represent.*, 2016.

[38] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in *Proc. of the conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[39] J. Ho, A. Jain, and P. Abbeel, Denoising diffusion probabilistic models, *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[40] K. T. Chen, E. Gong, F. B. de Carvalho Macruz, J. Xu, A. Boumis, M. Khalighi, K. L. Poston, S. J. Sha, M. D. Greicius, E. Mormino *et al.*, Ultra–low-dose 18f-florbetaben amyloid pet imaging using deep learning with multi-contrast mri inputs, *Radiology*, vol. 290, no. 3, pp. 649–656, 2019.

[41] B. Yu, L. Zhou, L. Wang, Y. Shi, J. Fripp, and P. Bourgeat, Ea-gans: edge-aware generative adversarial networks for cross-modality mr image synthesis, *IEEE Transactions on Medical Imaging*, vol. 38, no. 7, pp. 1750–1762, 2019.

[42] T. Zhou, H. Fu, G. Chen, J. Shen, and L. Shao, Hi-net: hybrid-fusion network for multi-modal mr image synthesis, *IEEE Transactions on Medical Imaging*, vol. 39, no. 9, pp. 2772–2781, 2020.

[43] Y. Luo, L. Zhou, B. Zhan, Y. Fei, J. Zhou, Y. Wang, and D. Shen, Adaptive rectification based adversarial network with spectrum constraint for high-quality pet image synthesis, *Medical Image Analysis*, vol. 77, p. 102335, 2022.

[44] Q. Huynh-Thu and M. Ghanbari, Scope of validity of psnr in image/video quality assessment, *Electronics Letters*, vol. 44, no. 13, pp. 800–801, 2008.

[45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[46] Z. Wang and A. C. Bovik, Mean squared error: Love it or leave it? a new look at signal fidelity measures, *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[47] D. Kinga, J. B. Adam *et al.*, A method for stochastic optimiza-

Just Accepted

tion, in *International conference on learning representations (ICLR)*, vol. 5, no. 6. San Diego, California;, 2015.

[48] C. A. Cocosco, Brainweb: Online interface to a 3d mri simulated brain database. *NeuroImage*, vol. 5, p. 425, 1997.

[49] D. L. Collins, A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, and A. C. Evans, Design and construction of a realistic digital brain phantom, *IEEE Transactions on Medical Imaging*, vol. 17, no. 3, pp. 463–468, 1998.

[50] H. M. Hudson and R. S. Larkin, Accelerated image reconstruction using ordered subsets of projection data, *IEEE Transactions on Medical Imaging*, vol. 13, no. 4, pp. 601–609, 1994.

[51] S. D. Jamadar, P. G. Ward, T. G. Close, A. Fornito, M. Premaratne, K. O'Brien, D. Stäb, Z. Chen, N. J. Shah, and G. F. Egan, Simultaneous bold-fmri and constant infusion fdg-pet data of the resting human brain, *Scientific Data*, vol. 7, no. 1, p. 363, 2020.

[52] S. D. Jamadar, E. X. Liang, S. Zhong, P. G. Ward, A. Carey, R. McIntyre, Z. Chen, and G. F. Egan, Monash dacra fpet-fmri: A dataset for comparison of radiotracer administration for high temporal resolution functional fdg-pet, *Gigascience*, vol. 11, no. 1, pp. 1–12, 2022.

[53] O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *Proc. of the Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

Yingkai Zhang received the B.S. degree in computer science in 2022 from the Beijing Institute of Technology, where he is currently working toward the Ph.D. degree in computer science and technology. His research interests include low-level vision, image processing, and computational photography.

Shuang Chen received the B.S. degree in computer science in 2021 from Xidian University, and the M.S. degree in computer science in 2024 from the Beijing Institute of Technology. Her research interests include low-level vision, and image processing.

Ye Tian received the B.S. degree from School of Information Science and Engineering, Lanzhou University, Lanzhou, China, in 2017, the M.S. degree from Peking University, Beijing, China, in 2020, and the Ph.D. degree from the School of Information and Electronics, Beijing Institute of Technology, Beijing, China, in 2024. Her current research interests include deep learning, image processing, and computational imaging.

Yunyi Gao received the B.S. degree from Beijing Institute of Technology, China in 2023, where he is currently working toward the M.S. degree in computer science and technology. His research interests focus on computer vision and image processing.

Jianyong Jiang received the B.S. and M.S. degrees in Engineering Physics from Tsinghua University, Beijing, China, in 2010 and 2012, respectively. He obtained the Ph.D. degree in Nuclear Engineering and Management from the University of Tokyo, Tokyo, Japan, in 2015. He is currently an Associate Professor at Beijing Normal University. His research interests encompass nuclear medicine imaging, and artificial intelligence-based methods for image reconstruction.

Ying Fu received the B.S. degree in electronic engineering from Xidian University, Xi'an, China, in 2009, the M.S. degree in automation from Tsinghua University, Beijing, China, in 2012, and the Ph.D. degree in information science and technology from the University of Tokyo, Tokyo, Japan, in 2015. She is currently a professor with the School of Computer Science and Technology, Beijing Institute of Technology. Her research interests include physics-based vision, image and video processing, and computational photography.

Just Accepted