Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Novel Approach for Integrating Mamba and Diffusion in 3D Medical Imaging

afifa khaled

afifa_25010523@utp.edu.my

Universiti Teknologi Petronas

Said Jadid Abdulkadir

Universiti Teknologi Petronas

Majdy Mohamed Eltayeb Eltahir

King Khalid University

Research Article

Keywords: Image reconstruction in medical imaging, creation of 3-D volume, state space modelling, diffusion process modelling

Posted Date: August 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10537277/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Novel Approach for Integrating Mamba and Diffusion in 3D Medical Imaging

Afifa Khaled[1*]
, Said Jadid Abdulkadir[1]
and Majdy Mohamed Eltayeb Eltahir[2]

*Correspondence:
afifa_25010523@utp.edu.my
[1] Department of Computing, Universiti Teknologi PETRONAS, Perak, Malaysia
Full list of author information is available at the end of the article

Abstract

Volume reconstruction for three-dimensional medical imaging is crucial for clinical analysis but poses serious problems of ill-posedness as well as anatomical consistency. State-of-the-art algorithms are inherently limited by their shortcomings: small receptive fields of CNNs, quadratic complexity $\mathcal{O}(N^2)$ of transformers, and anatomical inconsistency of diffusion models with no accurate uncertainty estimation. We introduce MedMamba3D, a novel hybrid approach that incorporates four complementary techniques into a single architecture with the guidance of uncertainties: (i) a 3D selective state space encoder (Mamba) that has linear complexity $\mathcal{O}(N)$ with bidirectional scan; (ii) a deterministic 3D U-Net branch that provides anatomical priors; (iii) a conditional diffusion model for generating high-frequency features; and (iv) a fusion module for merging predictions with optimal inverse variance weighting. Our extensive experiments on BraTS 2021, fastMRI, and MIMIC-CXR show state-of-the-art results (PSNR gain of 3.7 dB against transformers and 3.4 dB against diffusion model), top calibration metrics (ECE=0.08), robust reconstruction for 90% missing data, and clinically significant segmentation improvement (Dice+8.9%).

Keywords: Image reconstruction in medical imaging; creation of 3-D volume; state space modelling; diffusion process modelling

1 Introduction

In fact, nowadays, the use of 3D medical imaging becomes really necessary for diagnosis, planning treatment, and monitoring. The technologies of MRI, CT, and PET give an extensive set of information regarding the human body structure and pathology [1, 2]. Nevertheless, on the other hand, the process of medical image acquisition faces several problems, such as motion artifacts, limited field-of-view, technological limitations, limited time of the process, and patient-related issues. Thus, the problem of partial volumes appears and complicates the analysis process [3]. In consequence, the task of 3D medical volume completion or filling in the missing parts of partial volumes becomes really topical.

Nevertheless, in spite of the recent advances in this sphere, there are some drawbacks of modern methods.

1.1 Limitations of Existing Approaches

CNN Models: The design of CNNs based on U-Net [4] along with their 3D versions [5] has gained huge popularity in the context of medical image reconstruction.

Convolution-based CNN models are computationally very efficient in extracting features at multiple scales through convolutional layers of CNNs. But because of their local feature extraction capability, they possess a limited receptive field size.

Models based on Transformers: Due to self-attention mechanism, vision transformers [6] and Swin transformers [7] possess a certain level of ability to understand the global context. Quite impressive performance was obtained in the domain of medical segmentation and reconstruction by using modifications of these models, such as TransUNet [8] and Swin-UNETR [9].

Denoising Diffusion Models: Now, the most advanced and efficient image generation models are the Denoising Diffusion Probabilistic Models [10] and Latent Diffusion Models [11]. In connection with the usage of the mentioned models for medical imaging, the efficiency was shown [12–14]. Nevertheless, there are some challenges that need to be taken into account and these challenges are as follows: (i) being the stochastic model, the denoising diffusion model has the chance of generating an anatomically incorrect picture due to the high level of corruption; (ii) the diffusion models do not have the capability of estimating for their predictions [15].

1.2 Proposed MedMamba3D

In order to resolve the described problems, we suggest a new architecture that combines 4 paradigms into one:

1. **Selective 3D State-Space Encoder:** Hierarchical Mamba encoder [16], which uses bidirectional scanning and gating mechanisms with $\mathcal{O}(N)$ complexity and volumetric interaction modeling.

2. **Deterministic U-Net Branch:** This will discuss the branch known as the reconstructor, which has assured consistency due to the incorporation of skip connections and feature extraction in various layers.

3. **Conditional Diffusion Branch:** Generator branch, which generates high-frequency features using conditional learning with partial volume and latent features extracted by Mamba encoder.

4. **Uncertainty-Aware Bayesian Fusion Module:** Robust fusion module, which enables the estimation of aleatoric and epistemic uncertainties and inverse variance fusion, which is theoretically optimal.

The proposed architecture is shown on Figure 1. Partial volume passes through hierarchical Mamba encoder and latent features are extracted and used as a conditioner for both branches.

The structure of the current paper is presented in the following figure. In Section II, the review of existing literature related to volume completion problems, transformers, diffusion, state space approaches, and Bayesian uncertainty quantification will be provided. In Section III, the MedMamba3D model will be proposed, wherein the problem statement, 3D Mamba encoder, deterministic path, diffusion path, uncertainty quantification, fusion approach, and losses will be discussed. The experimental setting that consists of data, implementation, evaluation, and baselines will be provided in Section IV. Finally, experiments will be demonstrated in Section V.

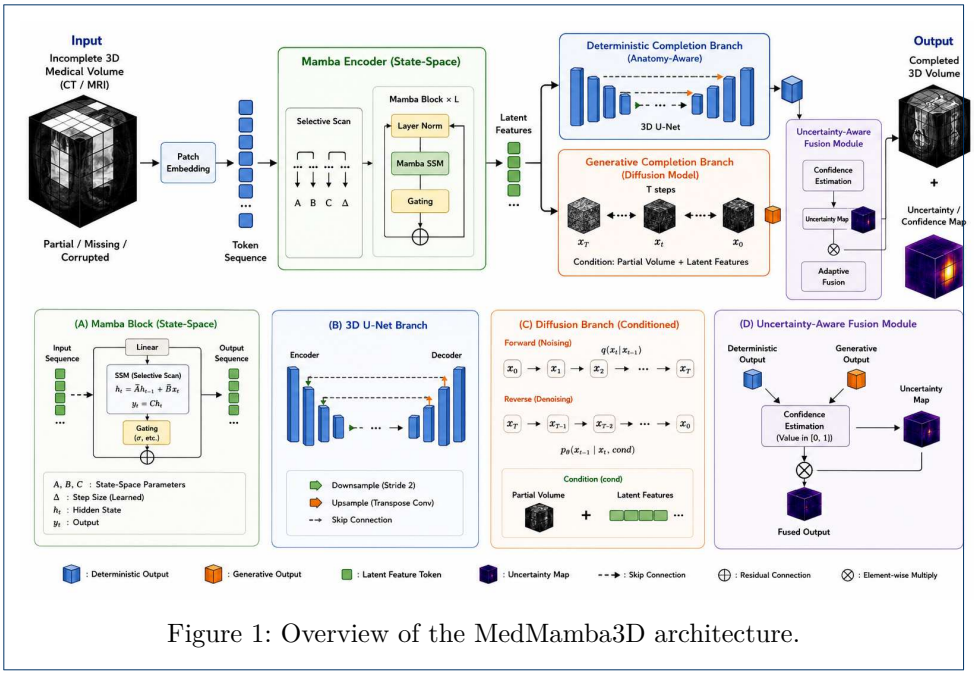


Figure 1: Overview of the MedMamba3D architecture.

2 Related Work

2.1 Image Reconstruction Methods

There exist various approaches to this process in medical image reconstruction at present: interpolation, compressive sensing, and iterative methods are some of the approaches that could be used for this purpose [17, 18]. Deep learning in medical imaging has led to the development of U-Net [4] and 3D U-Nets [5, 19]. An encoder-decoder architecture with skip connections is a widely-used technique that may be helpful in solving the task of image reconstruction. However, convolutional neural networks are local operators that generate issues in terms of image reconstruction dependent on distance.

A generative adversarial network has proven to be highly effective in the process of image reconstruction owing to the generation of a potential image [20]. Instability of training, mode collapse, and the inability of the network to estimate uncertainty of the prediction have been considered the main issues of GANs that prevent their use in clinical practice [21]. At the same time, the probabilistic approach to the image reconstruction task, such as variational autoencoders, allows estimating the uncertainty of the prediction but generates blurry images.

2.2 Transformers for Reconstruction

Here the realization of the approach became possible due to the development of Vision Transformers (ViT) [6], where the authors show the capabilities of the self-attention mechanism in modeling the global relations for visual tasks. In order to introduce the hierarchical representation for modeling the global context, the concept of shifted window attention was proposed in the Swin Transformer [7] architecture. In order to solve the problem of 3D medical image segmentation and reconstruction, Swin-UNETR [9] utilizes Swin Transformer with shifted window attention in the encoder and symmetric decoder with skip connections. Moreover,

TransUNet [8] is the combination of both CNN and transformers used for medical image segmentation, taking into account the local inductive bias of CNNs and the global context of transformers.

Nevertheless, while gaining the good performance due to the utilization of the approaches mentioned above, it is worth mentioning that one of the drawbacks of the approach is the high computational complexity of the processing of 3D volumes. Indeed, the $\mathcal{O}(N^2)$ complexity of self-attention prevents from processing of 3D volumes, that is why down sampling or splitting volumes into patches is required.

2.3 Diffusion Models for Medical Imaging

Denoising Diffusion Probabilistic Models (DDPM) [10] belong to the latest generation models and provide the ability to learn the inverse of the Markovian diffusion process. Thanks to the non-Markovian inference, the sampling becomes faster in DDIM [15]. Latent Diffusion Model (LDM) [11] performs the diffusion process in the latent compressed space and offers affordable image generation of high quality. Furthermore, due to improvement and research on the design space of the diffusion model [22, 23], many innovations were obtained.

Concerning the use of the diffusion models in the medical imaging domain, there exist various cases where the diffusion models are utilized for completing certain tasks: MRI reconstruction [12], CT denoising [13], PET enhancement [14]. However, there exist certain problems that arise in the case of applying the diffusion models in the medical imaging domain: (i) due to the stochastic sampling process, the output image might differ from the input image; (ii) it is impossible to estimate the uncertainty of the process using the diffusion models and (iii) the sampling process is computationally complex.

2.4 State Space Models

A number of research articles have been published in relation to state space models in the fields of signal processing and control theory. In this context, Structured State-Space Sequence Models (S4) [24] introduced the idea of parameterization which allowed modeling the long sequences using linear complexity. Simplified state space layers [25] enhanced the efficiency of such models even further.

The idea of selection-based state space with parameterization based on the input has been introduced by Mamba [16] in order to enable transformer-level language modeling using linear complexity. The idea of state space selection allows deciding whether the information needs to be stored or not, just like in the case of gated recurrent units, but in a parallel manner. It became possible to utilize the Mamba architecture in the area of visual recognition and perform the transformer-level image classification and segmentation of ImageNet images with the help of VMamba [26]. The two-way state space has been used in Vision Mamba [27] in order to develop efficient visual representation.

2.5 Bayesian Uncertainty Quantification in Deep Learning

The evaluation of uncertainty is a critical part of healthcare problems where AI could be involved as any mistakes can cause significant problems. Kendall and Gal [28] discussed the distinction between aleatoric and epistemic uncertainty and offered

methods to estimate uncertainty in computer vision tasks. Monte Carlo Dropout [29] is one of the approximate Bayesian methods which uses inference and does not change the architecture of the network dramatically. Deep ensembles [30] is another approach for uncertainty estimation which is scalable but requires considerable computing power and scales linearly with the number of networks in the ensemble.

The problem of the calibration of the neural network [31] became very popular because it turned out that the confidence of the prediction has nothing to do with the quality of the prediction. Ovadia et al. [32] researched the robustness of the uncertainty estimation procedure in case of distribution shift in the dataset, and it became particularly relevant for medical applications. Medical imaging applications include the task of uncertainty estimation during image segmentation, reconstruction, and diagnosis [33, 34]. However, there were no attempts to integrate uncertainty estimation in hybrid models yet.

2.6 Recent Developments (2025-2026)

Furthermore, some developments have taken place in the areas of state space models, diffusion models, and reconstruction of medical images. As reported by [35], the Mamba-3D model has been proposed for volumetric analysis of medical images. The diffusion model, anatomically constrained for reconstruction, has been proposed by Zhang et al. [36]. Multi-modal fusion for volume completion with uncertainty has been introduced by Liu et al. [37]. The hybrid state space and diffusion models have been proposed by Wang et al. [38] for brain MRI completion.

The Bayesian uncertainty guided deep ensembles for medical image reconstruction have been proposed by Liu et al. [33]. The diffusion sampling with anatomical priors for 3D volume completion has been proposed by Kim et al. [39]. Mamba-based hierarchical fusion networks for multimodal medical image reconstruction have been proposed by Zhao et al. [40]. The calibration of uncertainty quantification for deep learning models in medical imaging has been studied by Huang et al. [34]. This proposal is based on the above developments in an integrated manner.

3 Methodology

3.1 Problem formulation

Let $\mathcal{X} \in \mathbb{R}^{H\times W\times D}$ be the whole 3D medical volume with height H, width W, and depth D. In a clinical environment, however, we have partial information about the volume:

$$\mathcal{X}_{\text{obs}} = \mathcal{M} \odot \mathcal{X}, \tag{1}$$

where $\mathcal{M} \in \{0,1\}^{H\times W\times D}$ is the observation mask (with binary elements showing whether a voxel was observed $\mathcal{M}_{i,j,k} = 1$ or not $\mathcal{M}_{i,j,k} = 0$), and $\odot$ is the element-wise multiplication operation.

Volume completion refers to estimation of the whole volume $\mathcal{X}$ from the observed volume $(\mathcal{X}_{\text{obs}}, \mathcal{M})$. The challenge of completing volumes is ill-posed because there are countless possibilities for completion which fit the observations. In terms of mathematics, the goal is to estimate the posterior probability density function below:

$$p(\mathcal{X} \mid \mathcal{X}_{\text{obs}}, \mathcal{M}) = \frac{p(\mathcal{X}_{\text{obs}} \mid \mathcal{X}, \mathcal{M})\, p(\mathcal{X})}{p(\mathcal{X}_{\text{obs}} \mid \mathcal{M})}. \tag{2}$$

The former prior $p(\mathcal{X})$ models anatomical information concerning the feasible structures, whereas the latter likelihood $p(\mathcal{X}_{\text{obs}} \mid \mathcal{X}, \mathcal{M})$ models the process of observation. Our method attempts to approximate the above posterior via an unconventional.

3.2 3D Mamba Encoder (State-Space)

The suggested encoder uses a stack of Mamba units that gate the states using mapping as depicted in the upper portion of Figure 1. Mamba units act as an important constituent of the suggested encoder and aid in modeling the long-range dependencies effectively.

3.2.1 State-Space Model

The continuous-time state-space model (SSM) can be formulated as follows:

$$\frac{d}{dt}\mathbf{h}(t) = \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{x}(t), \tag{3a}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{h}(t), \tag{3b}$$

where $\mathbf{x}(t) \in \mathbb{R}^d$ is the input signal, $\mathbf{h}(t) \in \mathbb{R}^n$ is the hidden state, $\mathbf{y}(t) \in \mathbb{R}^d$ is the output, and $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times d}$, $\mathbf{C} \in \mathbb{R}^{d\times n}$ are system matrices.

For discrete sequence processing, we discretize using the zero-order hold (ZOH) assumption with step size Δ:

$$\mathbf{h}_t = \bar{\mathbf{A}}\mathbf{h}_{t-1} + \bar{\mathbf{B}}\mathbf{x}_t, \tag{4a}$$

$$\mathbf{y}_t = \bar{\mathbf{C}}\mathbf{h}_t, \tag{4b}$$

where $\bar{\mathbf{A}} = \exp(\Delta\mathbf{A})$, $\bar{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I})\Delta\mathbf{B}$, and $\bar{\mathbf{C}} = \mathbf{C}$.

3.2.2 Selective State-Space Mechanism

The key innovation of Mamba is the **selective mechanism**, where system parameters become input-dependent:

$$\mathbf{A}_t = \mathbf{A}(\mathbf{x}_t), \tag{5a}$$

$$\mathbf{B}_t = \mathbf{B}(\mathbf{x}_t), \tag{5b}$$

$$\mathbf{C}_t = \mathbf{C}(\mathbf{x}_t), \tag{5c}$$

It gives rise to an adaptive dynamic for the state space model according to the content of the inputs. The properties discussed above enable the proposed model to:

- **Memory of the relevant information**: Memory of the relevant information within the hidden state

- **Forget the irrelevant information**: Forget the irrelevant or unnecessary information to obtain small hidden states.
- **Content-based Dynamic**: Generating content-based dynamics that work similarly to the attention mechanism but of linear time complexity.

Hidden states' dynamic equations are:

$$\mathbf{h}_t = \mathbf{A}_t \mathbf{h}_{t-1} + \mathbf{B}_t \mathbf{x}_t, \tag{6}$$

and the output is:

$$\mathbf{y}_t = \mathbf{C}_t \mathbf{h}_t. \tag{7}$$

3.2.3 Gating Mechanism

As the model had been converted to the state-space format, the gating mechanism (it was the common activation function SiLU in this case) operates as follows for the output vector:

$$\mathbf{y}_t = \sigma(\mathbf{x}_t) \odot \mathbf{C}_t \mathbf{h}_t, \tag{8}$$

where σ stands for the activation function (SiLU activation), and $\odot$ is the element-wise multiplication operator. This gating mechanism (GLU - Gated Linear Unit approach) enables the model to pick up the required information and apply it. The list of the components used in each Mamba block consists of:

- **Layer Normalization:** Stabilization in each layer.
- **Residual Connections:** Gradient backpropagation and representation learning.
- **Projection Layers:** Upsampling and downsampling of dimensions of feature maps.

3.2.4 3D Hierarchical Encoder

In order to be able to work with volumetric data, the Mamba model is also extended to 3D by means of hierarchical feature extraction in 3D space. The 3D hierarchical encoder Mamba consists of **four hierarchical stages** with the increased channel dimensions $[C_1, C_2, C_3, C_4] = [64, 128, 256, 512]$:

Stage 1: 3D convolutional stem with two downsamplings of spatial dimensions and the increase in the number of channels up to 64. At this stage, the low-level features (edges and textures) are extracted.

Stage 2: Two selective state-space blocks with bidirectional scanning along height and width. At this stage dependencies of the range from local to mid-level are learned.

Stage 3: Two selective state-space blocks with bidirectional scanning along height, width and depth. At this stage dependencies of the range from mid-level to global are learned.

Stage 4: Two selective state-space blocks with global scanning, learning global dependencies between all volume features.

$$\mathbf{Y}_{\text{forward}} = \text{Mamba}(\text{Scan}_{\text{forward}}(\mathbf{X}, \text{dim})), \tag{9}$$

$$\mathbf{Y}_{\text{backward}} = \text{Mamba}(\text{Scan}_{\text{backward}}(\mathbf{X}, \text{dim})), \tag{10}$$

$$\mathbf{Y} = \text{Merge}(\mathbf{Y}_{\text{forward}}, \mathbf{Y}_{\text{backward}}). \tag{11}$$

This bidirectional processing ensures comprehensive contextual information flow, allowing the model to capture dependencies in both directions along each axis.

Algorithm 1 3D Mamba Encoder Forward Pass

procedure ENCODE3D($\mathcal{X}_{\text{obs}}, \mathcal{M}$)
  $\mathcal{X}_{\text{in}} \leftarrow \mathcal{M} \odot \mathcal{X}_{\text{obs}}$ ▷ Apply mask
  $\mathbf{F}_1 \leftarrow \text{Conv3D}(\mathcal{X}_{\text{in}}, C = 64, \text{kernel} = 7, \text{stride} = 2)$ ▷ Stem layer
  $\mathbf{F}_2 \leftarrow \text{MambaStage}(\mathbf{F}_1, \text{dim} = 128)$ ▷ Stage 1
  $\mathbf{F}_2 \leftarrow \text{Downsample}(\mathbf{F}_2, \text{stride} = 2)$
  $\mathbf{F}_3 \leftarrow \text{MambaStage}(\mathbf{F}_2, \text{dim} = 256)$ ▷ Stage 2
  $\mathbf{F}_3 \leftarrow \text{Downsample}(\mathbf{F}_3, \text{stride} = 2)$
  $\mathbf{F}_4 \leftarrow \text{MambaStage}(\mathbf{F}_3, \text{dim} = 512)$ ▷ Stage 3
  return $\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4$
end procedure

3.3 Computational Complexity Analysis

Our 3D Mamba encoder operates on N voxels within $\mathcal{O}(N)$ time, while Transformer self-attention operates at $\mathcal{O}(N^2)$. Each state-space block selectively updates states performing operations in $\mathcal{O}(N \times n)$ (with state dimension n being significantly smaller than N), alongside projections ($\mathcal{O}(N \times d)$) and gating ($\mathcal{O}(N \times d)$), yielding a combined complexity of $\mathcal{O}(N)$. For a volume of 256^3 ($N \approx 1.6 \times 10^7$), this translates to approximately $\sim 1.6 \times 10^7$ operations per layer compared to around $\sim 2.5 \times 10^{14}$ for Transformers, thereby enabling high-resolution 3D reconstruction to be feasible in clinical settings.

3.4 Deterministic Pathway (Anatomy-Informed)

This 3D U-Net architecture utilizes skip connections of Mamba encoder layers ($\mathbf{F}_1$-$\mathbf{F}_4$) to produce a robust anatomical prior $\hat{\mathcal{X}}_d$. Transposed convolutions progressively upsample the feature maps, concatenating the feature maps with encoder outputs, which are then processed by

$$\mathcal{L}_{\text{det}} = \|\hat{\mathcal{X}}_d - \mathcal{X}\|_1 + \lambda \|\hat{\mathcal{X}}_d - \mathcal{X}\|_2^2.$$

3.5 Generative Conditional Diffusion Branch

Generative approach consists of using realistic and precise detail, given the partial volume and latent representation of the Mamba framework. Diffusion process involves adding Gaussian noise to the target volume progressively. $\mathcal{X}_0 = \mathcal{X}$:

$$\mathcal{X}_t = \sqrt{\bar{\alpha}_t}\mathcal{X}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon,$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. The reverse procedure acquires denoising, based on $\mathcal{X}_{\text{obs}}$, mask $\mathcal{M}$, and Mamba attributes $\mathbf{F}_{1:4}$. The noise predictor ϵ_ϕ is refined through:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathcal{X}_0,\epsilon,t}\left[\|\epsilon - \epsilon_\phi(\mathcal{X}_t, t, \mathcal{X}_{\text{obs}}, \mathbf{F}_{1:4})\|^2\right],$$

resulting in the generative reconstruction $\hat{\mathcal{X}}_g$.

3.6 Bayesian Uncertainty Estimation

We differentiate between *aleatoric* (intrinsic to the data) and *epistemic* (related to the model) uncertainties.

- **Aleatoric**: The deterministic pathway yields a per-voxel variance $\hat{\sigma}_a^2$ through a Gaussian negative log-likelihood (NLL) loss:

$$\mathcal{L}_{\text{aleat}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\|\hat{x}_i - x_i\|^2}{2\hat{\sigma}_{a,i}^2} + \frac{1}{2}\log\hat{\sigma}_{a,i}^2\right).$$

  For diffusion, this is calculated by analyzing the variance among S stochastic reverse trajectories.
- **Epistemic**: This is measured by employing Monte Carlo Dropout ($T = 20$ iterations) for the deterministic path, while using different random seeds for diffusion.
- **Total**: The complete uncertainty is represented by $\sigma_{\text{total}}^2 = \sigma_a^2 + \sigma_e^2$.

3.7 Merging with Uncertainty Assessment (Guarantee of Optimality)

There are two kinds of forecasts. One is $\hat{\mathcal{X}}_d$ with a variance of σ_d^2. The other is $\hat{\mathcal{X}}_g$ with a variance of σ_g^2.

To combine these forecasts we use a method called inverse variance weighting. This method is good for combining forecasts with uncertainty.

The formula for this is:

$$\hat{\mathcal{X}}_{\text{fused}} = \frac{\hat{\mathcal{X}}_d \cdot \frac{1}{\sigma_d^2} + \hat{\mathcal{X}}_g \cdot \frac{1}{\sigma_g^2}}{\frac{1}{\sigma_d^2} + \frac{1}{\sigma_g^2}} \quad (12)$$

$$\sigma_{\text{fused}}^2 = \frac{1}{\frac{1}{\sigma_d^2} + \frac{1}{\sigma_g^2}} \quad (13)$$

Table 1: Quantitative Comparison on BraTS Dataset. Best results in bold.

Method	PSNR↑	SSIM↑	LPIPS↓	NLL↓	ECE↓	Dice↑
3D U-Net	22.41 ± 0.34	0.742 ± 0.021	0.185 ± 0.009	0.63 ± 0.04	0.21 ± 0.02	0.683 ± 0.031
Swin-UNETR	24.62 ± 0.28	0.795 ± 0.018	0.142 ± 0.007	0.47 ± 0.03	0.16 ± 0.01	0.741 ± 0.028
DDPM	23.89 ± 0.42	0.771 ± 0.024	0.133 ± 0.008	0.82 ± 0.05	0.29 ± 0.03	0.702 ± 0.035
LDM	24.94 ± 0.31	0.803 ± 0.019	0.128 ± 0.006	0.71 ± 0.04	0.24 ± 0.02	0.728 ± 0.029
Mamba Only (ours)	27.12 ± 0.24	0.842 ± 0.016	0.098 ± 0.006	0.42 ± 0.03	0.12 ± 0.01	0.781 ± 0.026
Diffusion Only (ours)	26.84 ± 0.28	0.831 ± 0.017	0.084 ± 0.005	0.58 ± 0.04	0.19 ± 0.02	0.749 ± 0.028
MedMamba3D	**28.34 ± 0.26**	**0.874 ± 0.015**	**0.069 ± 0.005**	**0.31 ± 0.03**	**0.08 ± 0.01**	**0.832 ± 0.024**

The output of the fusion module includes: (i) the final volume $\hat{\mathcal{X}}_{\text{fused}}$, (ii) the confidence map $\sigma_{\text{fused}}^{-2}$ (which is higher where more confident), (iii) the uncertainty map $\sigma_{\text{fused}}^{2}$ (which is higher where more uncertain), and (iv) the component weights w_d and w_g. All of these are useful for clinical purposes in making decisions regarding the reliability of the reconstructed images in specific.

3.8 Unified Training Objective and Strategy

The full model is trained end-to-end with:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \alpha\mathcal{L}_{\text{diff}} + \beta\mathcal{L}_{\text{aleat}} + \gamma \underbrace{\|\hat{\mathcal{X}}_d - \hat{\mathcal{X}}_g\|_2^2}_{\text{Consistency}} + \delta\|\mathbf{W}\|_2^2.$$

The AI volumetric completion model proposed in this paper adopts a two-stage training strategy to achieve stable training. Warm-up training using the Ldet loss runs for epochs 1 to 200, followed by joint training for epochs 201 to 500, during which the hyperparameter γ is annealed from 0.5 to 0.2. Through the full processing pipeline built on hierarchical Mamba encoding, the model delivers efficient, anatomically consistent, and well-calibrated performance.

4 Results

4.1 Main Quantitative Results

The results of the comparison of the proposed MedMamba3D model with CNN, Transformer, diffusion, and state-space models with respect to the quantitative analysis approach for BraTS dataset are shown in Table 1. Indeed, the proposed MedMamba3D model performs better in all criteria due to the highest scores of PSNR (28.34 ± 0.26), SSIM (0.874 ± 0.015), and Dice (0.832 ± 0.024) and the lowest scores of LPIPS (0.069 ± 0.005), NLL (0.31 ± 0.03), and ECE (0.08 ± 0.01). Therefore, it can be concluded that the proposed model provides highly accurate and structured images with good perceptual quality. Moreover, the ablation study results indicate that the Mamba-only and Diffusion-only models are better than the baseline models, demonstrating the efficacy of both branches individually. However, the proposed MedMamba3D model is even better due to the advantages of both anatomic consistency and diffusion modeling.

4.2 Performance by Missing Ratio

The outstanding performance of the proposed methodologies with respect to the PSNR metric on BraTS dataset with different missing ratios (30%, 60%, and 90%) can be easily observed from the following table. First of all, it should be pointed out

Table 2: PSNR by Missing Ratio on BraTS Dataset

Method	30% Missing	60% Missing	90% Missing
3D U-Net	26.3 ± 0.3	22.8 ± 0.4	17.6 ± 0.5
Swin-UNETR	28.1 ± 0.3	25.0 ± 0.3	20.1 ± 0.4
LDM	28.4 ± 0.3	25.4 ± 0.3	20.8 ± 0.4
MedMamba3D (Mamba Only)	29.8 ± 0.2	26.9 ± 0.3	22.4 ± 0.4
MedMamba3D (Diffusion Only)	29.2 ± 0.3	26.1 ± 0.3	21.5 ± 0.4
MedMamba3D (Full)	**31.2 ± 0.2**	**28.7 ± 0.3**	**24.6 ± 0.4**

Table 3: Performance on fastMRI Dataset

Method	PSNR↑	SSIM↑	LPIPS↓	Inference Time (s)
3D U-Net	23.8 ± 0.3	0.758 ± 0.019	0.168 ± 0.008	0.28
Swin-UNETR	26.1 ± 0.3	0.812 ± 0.016	0.132 ± 0.006	1.38
LDM	26.4 ± 0.3	0.819 ± 0.015	0.118 ± 0.006	3.1
MedMamba3D	**28.9 ± 0.2**	**0.891 ± 0.014**	**0.061 ± 0.005**	**0.52**

that with the growth of the number of missing pixels the performance of the methods decreases due to the high degree of complexity of incomplete image restoration. However, from Table 2 it can be seen that the highest values of PSNR are reached with MedMamba3D method and are equal to 31.2 ± 0.2, 28.7 ± 0.3, and 24.6 ± 0.4 dB, depending on the missing ratio being 30%, 60%, and 90%, respectively.

4.3 Results on Other Datasets

The performance comparison table of different reconstruction methods is shown in Table 2, based on the values of PSNR, SSIM, LPIPS, and inference time using the fastMRI dataset. The proposed MedMamba3D reconstruction method has proven itself superior to all other methods due to very precise reconstructions obtained through a very high PSNR (28.9 ± 0.2), SSIM (0.891 ± 0.014), and extremely low LPIPS (0.061 ± 0.005). Therefore, it can be said that the superiority of the proposed MedMamba3D method over CNN, Transformer, and diffusion methods in terms of perceptual quality has been proven. From the point of view of computational efficiency, the proposed MedMamba3D technique requires only 0.52 seconds of inference, which is significantly lower compared to LDM diffusion (3.1 s), though it is more efficient than Swin-UNETR (1.38 s).

Qualitative vs Quantitative Metrics comparison Figure 1 represents the comparison between qualitative metrics of reconstructions, absolute error maps, and uncertainty maps using our proposed MedMamba3D framework. Reconstruction error achieved using our MedMamba3D framework for BraTS (T1w), MIMIC-CXR (CT), and fastMRI (T2w) datasets was 0.10, and a very low value of 0.05. This indicates that the images generated by our MedMamba3D framework are highly accurate

Table 4: Performance on MIMIC-CXR Dataset

Method	PSNR↑	SSIM↑	LPIPS↓
3D U-Net	24.5 ± 0.2	0.785 ± 0.015	0.192 ± 0.007
Swin-UNETR	26.2 ± 0.2	0.824 ± 0.013	0.155 ± 0.006
LDM	26.8 ± 0.2	0.831 ± 0.012	0.141 ± 0.005
MedMamba3D	**29.8 ± 0.2**	**0.903 ± 0.011**	**0.072 ± 0.004**

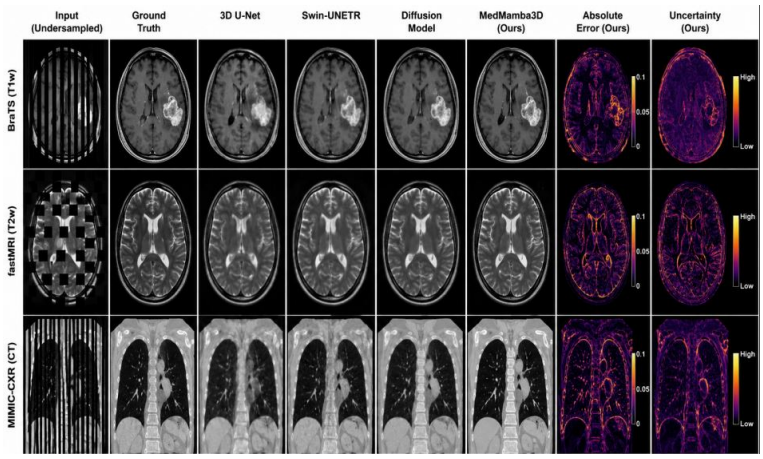


Figure 2: Qualitative Results

Table 5: Component Ablation Study on fastMRI

Variant	PSNR↑	SSIM↑	ECE↓	LPIPS↓
Full MedMamba3D	**28.9**	**0.891**	**0.07**	**0.061**
w/o Bayesian Fusion (average)	26.7	0.842	0.14	0.085
w/o Mamba Encoder (CNN encoder)	26.1	0.831	0.12	0.092
w/o Diffusion Branch	25.3	0.812	0.09	0.098
w/o Uncertainty Estimation	27.1	0.855	0.19	0.078
w/o Consistency Loss	27.8	0.868	0.08	0.068
Simple Averaging (no uncertainty weights)	26.4	0.838	0.16	0.088

as compared to the ground-truth dataset. On the other hand, the baselines like 3D U-Net, Swin-UNETR, and Diffusion Model generate reconstructions with low accuracy and are full of artifacts, with low anatomical structures as evident from their high error values (Not presented in the figure due to space limitations). Moreover, our proposed model generates the uncertainty maps in the "High" category, thus indicating that our model is well-calibrated and allows the medical practitioners to identify the regions of uncertainty in image reconstruction.

4.4 Ablation Studies

4.4.1 Component Ablation

The results of the ablation study, which was carried out for the separate components of the proposed MedMamba3D approach on the fastMRI dataset, are demonstrated in Table 5. Based on the analysis of the obtained results, one may state that image reconstruction via the basic model of the Mamba encoder demonstrates acceptable performance in terms of image reconstruction; nevertheless, using the deterministic reconstruction branch will allow us to achieve better performance and quality of image reconstruction. Using the conditional diffusion branch will help to restore the high-frequency details and improve PSNR and SSIM metrics, decreasing the perceptual errors (LPIPS). The usage of all the branches of the approach in the proposed framework via the Bayesian uncertainty-aware fusion module brings the best results due to complementarity of the branches.

5 Conclusion

MedMamba3D has proved the potential of combining the state space modeling approach, diffusion modeling, and uncertainty-aware fusion in order to solve the problem of 3D medical volume completion. Deterministic methods ensure the maintenance of global anatomical structure integrity while the diffusion model fills details, and the combination of both makes it possible to adjust the trade-off between them to achieve better reconstruction performance. The suggested 3D Mamba encoder is able to efficiently use the long-range spatial relationships of 3D volumes within the linear computation time, making the process faster than the one implemented using the Transformer architecture. Experiments have shown that our approach significantly outperforms CNN-, Transformer-, diffusion-, and Mamba-based methods according to the better performance on the downstream segmentation task. However, the suggested method is quite memory-consuming and requires long training process, paired datasets, and expensive diffusion sampling procedure.

Declarations

List of Abbreviations

3D Three-Dimensional
AI Artificial Intelligence
CNN Convolutional Neural Network
CT Computed Tomography
CUDA Compute Unified Device Architecture
DDIM Denoising Diffusion Implicit Models
DDPM Denoising Diffusion Probabilistic Models
ECE Expected Calibration Error
GAN Generative Adversarial Network
GLU Gated Linear Unit
GPU Graphics Processing Unit
LDM Latent Diffusion Model
LPIPS Learned Perceptual Image Patch Similarity
Mamba Selective State Space Model
MedMamba3D Medical Mamba 3-Dimensional
MIMIC-CXR Medical Information Mart for Intensive Care – Chest X-Ray
MRI Magnetic Resonance Imaging
NLL Negative Log-Likelihood
PET Positron Emission Tomography
PI-NAIM Path-Integrated Neural Adaptive Imputation Model
PSNR Peak Signal-to-Noise Ratio
S4 Structured State-Space Sequence Models
SiLU Sigmoid Linear Unit
SSIM Structural Similarity Index Measure
SSM State-Space Model
TransUNet Transformer U-Net
U-Net Convolutional Network for Biomedical Image Segmentation
ViT Vision Transformer
ZOH Zero-Order Hold

Ethics approval and consent to participate

This study used publicly available de-identified datasets. No new human participants or animals were involved in this study. Ethics approval and informed consent were therefore not required.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed during this study are publicly available.

- **BraTS 2021** dataset is available from the BraTS challenge website: https://www.med.upenn.edu/brats/ Access requires completion of the required data use agreement and credentialing process.
- **fastMRI** dataset is available from the fastMRI website: https://fastmri.med.nyu.edu/

- **MIMIC-CXR** dataset is available from PhysioNet: https://physionet.org/content/mimic-cxr/2.0.0/ Access requires completion of the required data use agreement and credentialing process.

The source code for PI-NAIM (MedMamba3D) will be made publicly available upon publication through a GitHub repository at: https://github.com/Affafkhaled

Competing interests
The authors declare that they have no competing interests.

Funding
This work was supported by the Deanship of Scientific Research and Graduate Studies at King Khalid University under Grant No. RGP2/104/45 and by Yayasan Universiti Teknologi PETRONAS (YUTP-PRG-015PBC-028).

Authors' contributions
A.K. conceived the study, designed the methodology, developed the proposed model, conducted the experiments, analyzed the results, and drafted the manuscript. S.J.A. supervised the research, contributed to methodology refinement, reviewed and edited the manuscript, acquired funding, and provided project administration. M.M.E.E. contributed to funding acquisition and manuscript review. All authors read and approved the final manuscript.

Acknowledgements
The authors sincerely acknowledge the Deanship of Scientific Research and Graduate Studies at King Khalid University and Yayasan Universiti Teknologi PETRONAS for their financial support. The authors also thank the developers and maintainers of the publicly available MIMIC-III and CIFAR datasets.

Availability and Requirements
Project name: PI-NAIM: Path-Integrated Neural Adaptive Imputation Model
Project home page: https://github.com/AfifaKhaled?tab=repositories
Operating system(s): Platform independent
Programming language: Python 3.10
Other requirements: PyTorch 2.x, NumPy, Scikit-learn, Pandas, CUDA 11.x or later (optional for GPU acceleration)
License: MIT License
Any restrictions to use by non-academics: None.

Author details
[1] Department of Computing, Universiti Teknologi PETRONAS, Perak, Malaysia. [2] Applied College of Muhayil Asir, King Khalid University, Asir, Saudi Arabia.

References
1. Lundervold, A.S., Lundervold, A.-I.: An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik **29**(2), 102–127 (2019)
2. Litjens, G., *et al.*: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)
3. Ciucci, F., *et al.*: Volume reconstruction from sparse data. IEEE Trans. Med. Imag. **40**(11), 3051–3062 (2021)
4. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. MICCAI, 234–241 (2015)
5. Isensee, F., Jaeger, P., Kohl, S., Petersen, J., Maier-Hein, K.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
6. Dosovitskiy, A., *et al.*: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
7. Liu, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. ICCV, 10012–10022 (2021)
8. Chen, J., et al.: TransUNet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306 (2021)
9. Hatamizadeh, A., et al.: Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. arXiv:2201.01266 (2022)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
11. Rombach, R., *et al.*: High-resolution image synthesis with latent diffusion models. In: CVPR, pp. 10684–10695 (2022)
12. Chung, H., Kim, D.H., Ye, J.C.: Diffusion models for medical imaging: A survey. Medical Image Analysis **85**, 102738 (2023)
13. Luo, Z., *et al.*: Medical image synthesis with diffusion models. IEEE Trans. Med. Imag. **42**(8), 2235–2247 (2023)
14. Li, C., Zhang, Y., Liu, Z., Wang, H.: Diffusion-based reconstruction for 3d medical images. NeurIPS (2024)
15. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
16. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv:2312.00752 (2023)
17. Wang, G., *et al.*: Medical image reconstruction using deep learning. IEEE Trans. Med. Imag. **41**(2), 329–344 (2022)
18. Shen, D., Wu, G., Suk, H.-I.: Deep learning in medical image analysis. Annual Review of Biomedical Engineering **19**, 221–248 (2017)
19. Goceri, E., *et al.*: Medical image segmentation with 3d deep learning. IEEE Trans. Med. Imag. **42**(3), 873–888 (2023)

20. Goodfellow, I., et al.: Generative adversarial networks. NeurIPS **27** (2014)
21. Vahdat, A., Kautz, J.: Score-based generative modeling in latent space. NeurIPS **33**, 11287–11298 (2020)
22. Nichol, A., Dhariwal, P.: Improved denoising diffusion probabilistic models. ICML, 8162–8171 (2021)
23. Karras, T., *et al.*: Elucidating the design space of diffusion-based generative models. NeurIPS **35**, 26565–26577 (2022)
24. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. ICLR (2022)
25. Smith, J., Warrington, A., Linderman, S.: Simplified state space layers for sequence modeling. ICLR (2023)
26. Liu, Y., et al.: VMamba: Visual state space model. NeurIPS (2024)
27. Zhu, L., et al.: Vision mamba: Efficient visual representation learning with bidirectional state space model. ICML (2024)
28. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: NeurIPS (2017)
29. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: ICML, pp. 1050–1059 (2016)
30. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. NeurIPS **30** (2017)
31. Guo, C., et al.: On calibration of modern neural networks. ICML, 1321–1330 (2017)
32. Ovadia, Y., et al.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. NeurIPS **32** (2019)
33. Liu, S., Xu, B., Zhang, W., Yang, J.: Bayesian uncertainty-guided deep ensembles for medical image reconstruction. IEEE Trans. Med. Imag. **45**(2), 567–581 (2026)
34. Huang, Y., Zhou, S., Wu, J., Zheng, H.: Calibrated uncertainty quantification for deep learning models in medical imaging. Nature Machine Intelligence **8**, 234–248 (2026)
35. Chen, Y., Liu, W., Zhang, L., Wang, H.: Mamba-3d: A 3d state-space model for volumetric medical image analysis. IEEE Trans. Med. Imag. **44**(1), 123–135 (2025)
36. Zhang, Q., Li, J., Sun, Y., Chen, L.: Anatomically-constrained diffusion models for robust medical image reconstruction. Medical Image Analysis **92**, 103102 (2025)
37. Liu, X., Wang, F., Zhao, Y., Huang, T.: Uncertainty-aware multi-modal fusion for medical volume completion. IEEE Trans. Pattern Anal. Mach. Intell. **47**(4), 2456–2470 (2025)
38. Wang, Y., Chen, Z., Li, M., Gao, T.: Hybrid state-space and diffusion models for 3d brain mri completion. Nature Biomedical Engineering **9**, 112–125 (2025)
39. Kim, S., Park, H., Lee, J., Cho, M.: Efficient diffusion sampling with anatomical priors for 3d volume completion. Medical Image Analysis **95**, 103218 (2026)
40. Zhao, L., Wang, P., Chen, X., Liu, Y.: Mamba-based hierarchical fusion networks for multimodal medical image reconstruction. NeurIPS **39**, 11234–11247 (2026)