

Article

VMedSAM-X: A State-Enhanced Dual-Branch Encoder for Efficient Promptable Medical Image Segmentation

Hengwei Zhang, Wei Li  and Yazhi Liu *

College of Artificial Intelligence, North China University of Science and Technology, Tangshan 063210, China; zhw@stu.ncst.edu.cn (H.Z.); lw@ncst.edu.cn (W.L.)

* Correspondence: liuyazhi@ncst.edu.cn

Abstract

Medical image segmentation plays a crucial role in clinical diagnosis and treatment planning. However, existing segmentation frameworks frequently exhibit high computational complexity and often fail to retain fine-grained structural details—especially along intricate anatomical boundaries such as blood vessels and tumor margins. To overcome these limitations, we propose VMedSAM-X, an efficient and computationally economical medical image segmentation framework that incorporates structured state space modeling into the Medical Segment Anything Model (MedSAM) architecture. The proposed method adopts a state-enhanced encoder that combines extended long short-term memory (xLSTM) with two-dimensional selective scanning (SS2D) and a dual-path cross-attention mechanism to enhance long-range dependency modeling while maintaining linear computational complexity. Experiments conducted on the 1024×1024 ACDC cardiac MRI dataset show that the proposed encoder reduces floating-point operations from 369.44 G to 17.36 G and achieves a $2.4\times$ improvement in inference speed compared with the Vision Transformer (ViT)-based encoder. Additional evaluations on the SegTHOR and MSD-Lung datasets demonstrate consistent improvements in Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) metrics over MedSAM and Vision Mamba U-Net (VM-UNet) baselines. These results indicate that the proposed framework provides an effective and computationally efficient solution for high-resolution medical image segmentation.

Keywords: medical image segmentation; MedSAM; VM-UNet; state space model; xLSTM; Mamba; computational efficiency

1. Introduction

Medical image segmentation plays a crucial role in computer-aided diagnosis, treatment planning, and disease monitoring, as it enables precise delineation of anatomical structures and pathological regions [1,2]. Thanks to the rapid development of deep learning, convolutional neural networks (CNNs)—particularly the U-shaped convolutional network (U-Net) and its variants—have achieved remarkable success in a wide range of medical imaging tasks by leveraging their strong local feature extraction capability and encoder–decoder design [3–5]. However, the inherently limited receptive field of convolutional operations restricted their ability to model long-range dependencies, which are essential for capturing global anatomical context in high-resolution medical images.

To address this limitation, vision transformers (ViT) were introduced to medical image segmentation, leveraging self-attention mechanisms to model global relationships among image tokens [6–9]. Recently, promptable segmentation frameworks such as the



Academic Editor: Thomas Lindner

Received: 5 April 2026

Revised: 22 April 2026

Accepted: 22 April 2026

Published: 24 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Segment Anything Model (SAM) and its medical adaptation Medical Segment Anything Model (MedSAM) further advanced this paradigm by enabling flexible segmentation through prompt guidance [10–14]. Despite their strong performance and generalization ability, transformer-based encoders suffered from quadratic computational complexity with respect to the number of tokens, leading to high memory consumption and computational cost, especially when processing high-resolution medical images.

In parallel, state space models (SSM) recently emerged as an efficient alternative for long-sequence modeling. Architectures such as Mamba introduced selective scanning mechanisms that achieved linear computational complexity while preserving the capability to capture long-range dependencies [15–18]. In addition, recurrent memory-based architectures, such as visual long short-term memory (xLSTM), demonstrated effectiveness in modeling sequential spatial context through gated state transitions [19]. These developments provided new opportunities to design lightweight vision encoders that retain global context modeling without relying on computationally expensive self-attention.

However, directly replacing the transformer encoder in promptable segmentation frameworks with a single alternative backbone may result in insufficient feature diversity [20]. Convolutional and state space models excelled at capturing structural and local spatial patterns, while recurrent memory mechanisms were more suitable for modeling long-range contextual dynamics. Effectively integrating these complementary representations into a unified architecture remains an open problem in efficient medical segmentation.

To address these challenges, we propose VMedSAM-X, a state-enhanced promptable segmentation framework that redesigns the MedSAM encoder using a dual-branch architecture. As illustrated in Figure 1, the proposed encoder consists of a Two-Dimensional Selective Scanning (SS2D) branch for efficient spatial structure modeling and an xLSTM branch for long-range contextual aggregation. To enable deep interaction between the two branches, we introduce a bidirectional cross-attention mechanism that allows each branch to dynamically refine its representation based on complementary features from the other branch. The fused features are then projected back to the original embedding space and combined with a residual connection to ensure compatibility with the prompt encoder and mask decoder of MedSAM.

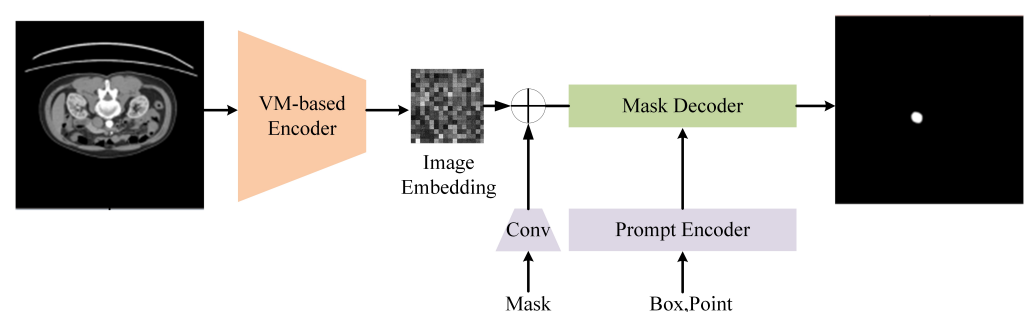


Figure 1. Overall architecture of the proposed VMedSAM-X framework. The framework consists of three main components: a multi-scale state-enhanced encoder, a prompt encoder, and a mask decoder. The input image is first processed by the proposed encoder, which integrates SS2D and xLSTM to capture long-range dependencies with linear computational complexity. The prompt encoder accepts user-provided prompts (e.g., boxes or points) to provide positional guidance. The mask decoder then fuses the encoded image features with the prompt information to generate the final segmentation mask.

The main contributions of this work are summarized as follows:

1. **A state-enhanced encoder for MedSAM:** We redesign the MedSAM image encoder using a dual-branch architecture that integrates selective state space modeling and visual long-short-term memory (LSTM), enabling efficient modeling of long-range dependencies while reducing the computational burden compared to standard ViT-based encoders.
2. **A bidirectional cross-attention fusion mechanism:** We propose a cross-attention module that facilitates mutual refinement of structural and contextual representations, thereby enabling dynamic alignment and synergistic fusion of complementary information across the two branches.
3. **Superior efficiency and competitive accuracy:** Experimental results on five public datasets show that the proposed framework cuts the computational cost by 95.3% (from 369.44 GB to 17.36 GB) compared to the ViT-based MedSAM encoder. This is achieved while maintaining or improving segmentation accuracy across different modalities and anatomical structures.

2. Related Work

2.1. CNN-Based Medical Image Segmentation

CNNs were the dominant approach for medical image segmentation due to their strong local feature extraction capability and computational efficiency [21]. U-Net established a widely adopted encoder–decoder architecture with skip connections, enabling precise localization by combining low-level spatial features with high-level semantic representations [3]. Subsequent variants such as U-Net++ and Attention U-Net further improved feature fusion and region focus through dense skip connections and attention mechanisms [22,23].

Despite these advances, CNN-based models inherently relied on local convolution operations, which limited their ability to capture long-range dependencies and global contextual information in high-resolution medical images [5,24,25]. This limitation motivated the exploration of architectures with stronger global modeling capabilities.

2.2. Transformer-Based Segmentation Models

Transformer architectures addressed the limitation of CNNs by leveraging self-attention mechanisms to model global dependencies. ViT demonstrated that image patches can be treated as tokens to enable long-range interaction modeling [6]. Hybrid architectures such as TransUNet combined CNN feature extraction with transformer-based global modeling, achieving improved segmentation performance [7].

Hierarchical transformers such as Swin Transformer further improved efficiency by restricting attention to local windows while enabling cross-window communication [26]. These models were successfully applied to medical segmentation tasks and extensively analyzed in recent surveys [8,9].

However, the computational and memory complexity of self-attention grew quadratically with input resolution, which posed significant challenges for high-resolution medical images [27,28]. This limitation became particularly critical in clinical scenarios requiring real-time processing or deployment on resource-constrained hardware.

2.3. Foundation Models for Medical Image Segmentation

The emergence of foundation models introduced a new paradigm for medical image segmentation. SAM enabled promptable segmentation across diverse visual domains through large-scale pretraining [10]. MedSAM adapted this framework to medical imaging by fine-tuning on a large multi-modal dataset, significantly improving performance on medical benchmarks [11].

Recent studies showed that while MedSAM demonstrated strong generalization capability, its performance varied significantly across tasks and anatomical structures [29,30]. In particular, its performance degraded on small or complex structures, and substantial improvements often required task-specific fine-tuning [31,32].

However, most existing efforts focused on prompt design and fine-tuning strategies, while the efficiency of the underlying transformer-based image encoder remained largely underexplored. The high computational cost of the ViT backbone limited the practical deployment of these models in real-world clinical environments.

2.4. State Space Models in Vision

SSMs recently emerged as an efficient alternative to attention-based architectures. The structured state space model (S4) demonstrated that long-range dependencies could be modeled with linear computational complexity [33]. Mamba further improved efficiency by introducing selective state updates, enabling dynamic information flow control [15].

Recent extensions such as VMamba and Visual Mamba adapted SSMs to two-dimensional visual data through multi-directional scanning mechanisms, achieving competitive performance in vision tasks [34–36]. These approaches enabled efficient global context aggregation while maintaining linear complexity.

Despite these advantages, existing SSM-based methods primarily focused on spatial dependency modeling and often lacked explicit mechanisms for capturing high-level semantic context, which is critical for accurate medical image segmentation.

2.5. Recurrent Memory Models in Medical Image Segmentation

Recurrent neural networks, particularly long short-term memory (LSTM), were widely used to model sequential dependencies in medical image analysis [37]. By introducing gating mechanisms, LSTM could selectively retain and update information over long sequences, making it suitable for capturing contextual continuity.

The recently proposed xLSTM extended traditional LSTM by introducing matrix-based memory and dual gating mechanisms, enabling richer representation and improved long-range dependency modeling [19]. In medical imaging, xLSTM-based architectures demonstrated promising results, particularly in scenarios with limited annotations and complex anatomical structures [38–40].

Compared with state space models that focus on spatial propagation, xLSTM maintained gated memory states, enabling selective retention of high-level semantic information. This made it particularly suitable for medical image segmentation tasks, where capturing anatomical consistency and contextual continuity was essential.

2.6. Summary

In summary, CNN-based methods provided strong local feature extraction but struggled with global context modeling, while transformer-based and foundation models offered improved global representation at the cost of high computational complexity. State space models provided efficient global dependency modeling but lacked explicit semantic memory mechanisms, whereas recurrent models captured contextual continuity but were limited in spatial modeling. These complementary characteristics motivate the proposed VMMedSAM-X framework, which integrates SS2D-based spatial modeling with xLSTM-based semantic memory through a unified dual-branch architecture [41,42].

3. Materials and Methods

To address the complementary limitations identified in Section 2—namely, the local receptive fields of convolutional neural networks, the quadratic complexity of Transform, and the lack of explicit semantic memory in existing SSM-based encoders—we introduce a

multi-scale state-enhanced image encoder into the MedSAM architecture and incorporate bidirectional cross-attention layers at the deep stages of feature extraction.

3.1. Overall Architecture

The overall architecture of VMedSAM-X is based on MedSAM's prompt-based segmentation paradigm, as shown in Figure 1. While retaining MedSAM's mask decoder and prompt encoder, it incorporates our newly designed multi-scale state-enhanced image encoder.

Unlike conventional ViT-based image encoders, our proposed image encoder models long-range dependencies in its core SS2D path with linear computational complexity and restricts computationally intensive fusion operations to deep-layer low-resolution features. Compared to transformer-based encoders, this approach achieves more efficient modeling of global dependencies.

The prompt encoder is responsible for converting user-provided spatial prompts (such as bounding boxes or points) into embedding representations aligned with image features. The mask decoder is responsible for fusing the visual features output by the image encoder with the prompt embeddings to ultimately generate a segmentation mask. In VMedSAM-X, both the prompt encoder and the mask decoder fully inherit the original architecture and pre-trained weights of MedSAM; they are kept frozen during training without any structural modifications.

3.2. Multi-Scale State-Enhanced Encoder

Medical images typically possess high spatial resolution and complex organizational structures. In traditional convolutional neural networks, feature extraction relies primarily on local convolution operations, whose receptive fields are limited by kernel size, making it difficult to effectively model long-range spatial dependencies. Although Transformer-based architectures with self-attention mechanisms can establish global correlations, their computational complexity grows quadratically with image size, resulting in prohibitive computational costs for high-resolution medical image segmentation tasks.

To address these issues, this paper constructs a multi-scale state-enhanced image encoder, as shown in Figure 2. The encoder adopts a pyramidal multi-scale architecture that first converts the input image into a serialized feature representation through patch embedding, followed by feature extraction and enhancement across multiple stages. For an input feature map $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, where H and W denote spatial resolution and C denotes the number of channels, each stage in the encoding path first reduces spatial resolution and increases channel count through patch merging operations, forming a hierarchical feature pyramid: resolution progressively decreases from $H/16$ to $H/128$, corresponding to channel counts increasing from C to $8C$.

Specifically, Stage 1 and Stage 2 each consist of two VSS blocks based on SS2D, followed by patch merging operations that halve the spatial resolution and double the channel dimensions. Stage 3 incorporates nine cross-attention fusion blocks, each containing parallel SS2D and xLSTM branches with a bidirectional cross-attention mechanism to integrate structural and semantic features. Stage 4 includes two such blocks. The output feature maps at the deepest stage have a resolution of $H/128 \times W/128$ with $8C$ channels.

At each resolution level, multiple Visual State Space (VSS) modules are stacked to capture long-range dependencies while maintaining linear computational complexity. Specifically, the VSS module employs the SS2D, which unfolds the feature map into sequences along both horizontal and vertical directions and performs state updates. The core idea of SS2D is to model spatial dependencies by scanning the feature map sequentially along both horizontal and vertical directions. At each step, the model updates a hidden state that

aggregates information from previously scanned positions, allowing long-range context to propagate efficiently across the entire image without the quadratic cost of self-attention. Formally, the discrete state update process can be expressed as:

$$\mathbf{h}_t = \bar{\mathbf{A}}\mathbf{h}_{t-1} + \bar{\mathbf{B}}\mathbf{x}_t \quad (1)$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{h}_t \quad (2)$$

where $\mathbf{x}_t \in \mathbb{R}^d$ is the input feature sequence, $\mathbf{h}_t \in \mathbb{R}^n$ is the hidden state, $\mathbf{y}_t \in \mathbb{R}^d$ is the output feature, and $\bar{\mathbf{A}}$, $\bar{\mathbf{B}}$, $\mathbf{C}$ are discretized state space parameters. Through this state propagation mechanism, the model can capture long-range dependencies while maintaining linear computational complexity.

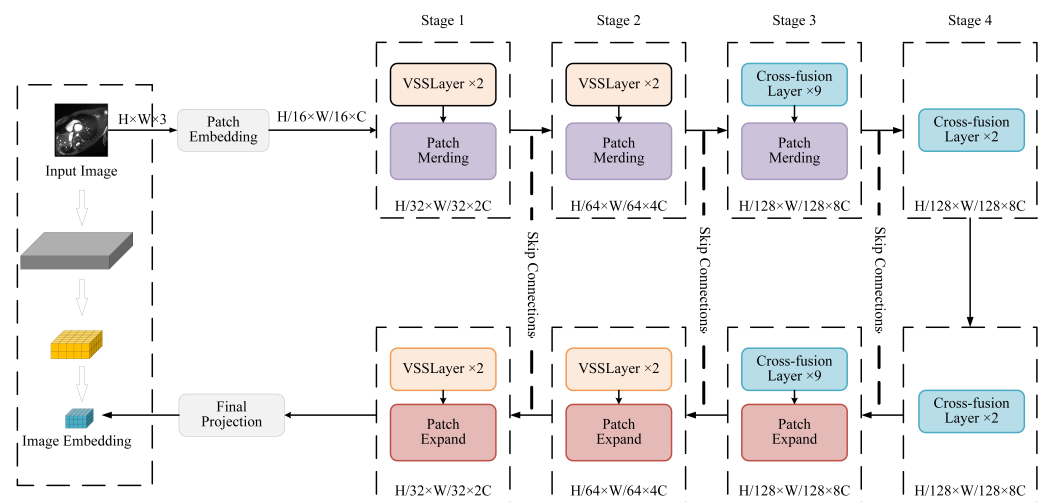


Figure 2. Architecture of the multi-scale state-enhanced encoder. The encoder adopts a four-stage pyramidal structure. Stage 1 and Stage 2 consist of SS2D blocks; Stage 3 and Stage 4 incorporate cross-attention fusion blocks with SS2D and xLSTM branches. The numbers of blocks in each stage are detailed in Section 3.2.

The encoder consists of multiple stages, with each stage progressively reducing spatial resolution through downsampling operations while increasing feature channel counts, thereby forming multi-level semantic representations. The feature representation at the l -th stage can be expressed as:

$$\mathbf{F}_l = \mathcal{E}_l(\mathbf{F}_{l-1}), \quad \mathcal{E}_l = \text{SS2D} \circ \text{PatchMerging} \quad (3)$$

Shallow features primarily retain fine-grained boundary and texture information, while deep features contain more abstract semantic information. By introducing the SS2D module at each encoding stage, the model enhances global information interaction across different scales.

Through the multi-scale state-enhanced encoder, the model can obtain rich local detail information and global semantic information while maintaining low computational complexity, providing more stable feature representations for subsequent feature fusion and decoder reconstruction.

3.3. Dual-Path Cross-Attention Mechanism

In the multi-scale state-enhanced encoder, the SS2D module can effectively model long-range dependencies in image features. However, this structure primarily focuses on the propagation of spatial structural information and has relatively limited expressive power for high-level semantic information. Target regions in medical images often exhibit

complex morphological characteristics, and relying solely on a single-path feature modeling can easily lead to insufficient semantic expression, thereby affecting segmentation accuracy.

To further enhance the model's ability to model global semantic information, this paper introduces a semantic memory modeling branch in the encoding stage and designs a dual-path cross-attention mechanism to achieve information fusion between structural features and semantic features. As shown in Figure 3, the cross-fusion module consists of two parallel branches: the SS2D-based spatial structure encoding branch and the xLSTM-based semantic memory modeling branch, with information interaction between the two paths achieved through bidirectional cross-attention.

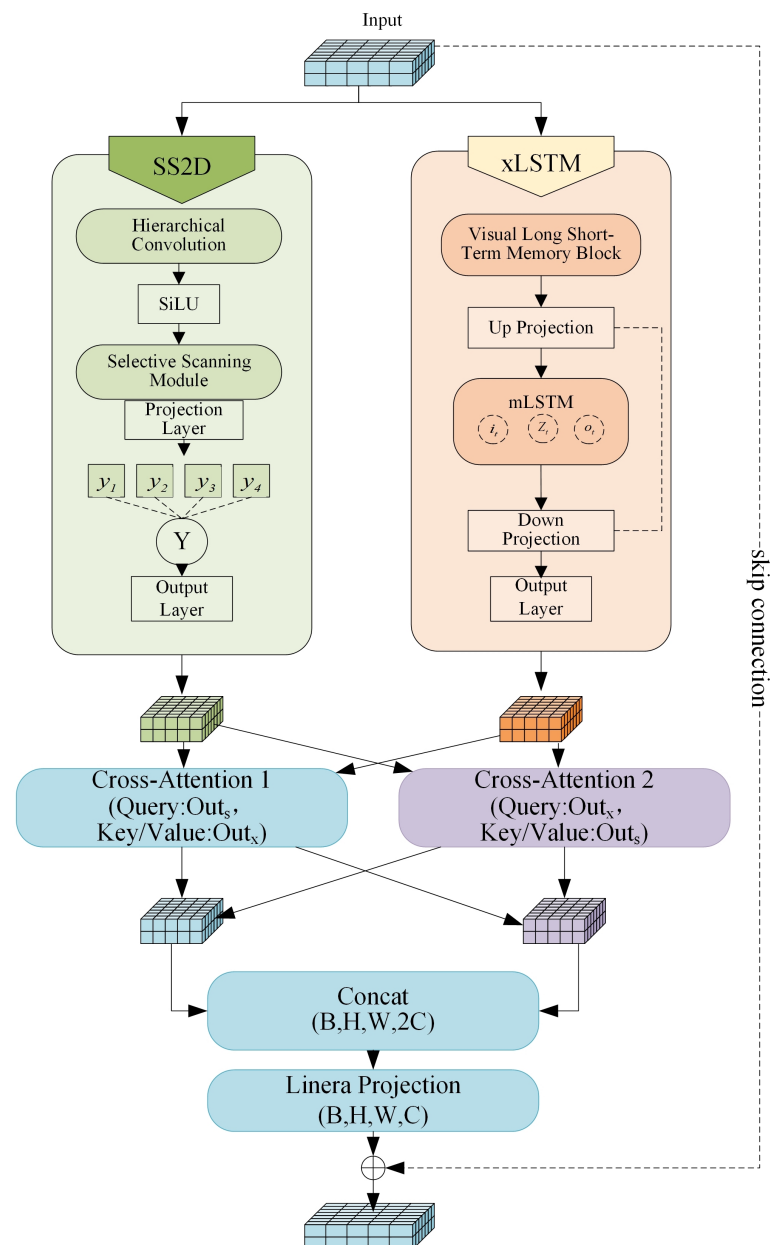


Figure 3. Illustration of the proposed dual-path cross-attention fusion module. The module consists of two parallel branches: an SS2D-based branch for spatial structural feature encoding and an xLSTM-based branch for semantic memory modeling. High-level features are first projected into sequence representations and processed by xLSTM to capture long-range semantic dependencies. Bidirectional cross-attention is then employed to enable information interaction between the structural and semantic feature streams, facilitating adaptive feature fusion and enhancing global context modeling capability.

Specifically, in the deeper stages of the encoder, the high-level features output by the multi-scale state-enhanced encoder are reshaped into sequence form and fed into xLSTM for semantic modeling. Unlike SS2D, which focuses on spatial propagation, xLSTM maintains a gated memory cell that selectively retains or forgets information over long sequences. This mechanism is particularly effective for preserving high-level semantic context across distant regions, such as the continuity of elongated organs or the global shape of anatomical structures. For an input feature sequence $\mathbf{X} \in \mathbb{R}^{L \times d}$, the xLSTM state update process can be expressed as:

$$\mathbf{i}_t = \sigma(\mathbf{W}_i \mathbf{x}_t + \mathbf{R}_i \mathbf{h}_{t-1} + \mathbf{b}_i) \quad (4)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_f \mathbf{x}_t + \mathbf{R}_f \mathbf{h}_{t-1} + \mathbf{b}_f) \quad (5)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o \mathbf{x}_t + \mathbf{R}_o \mathbf{h}_{t-1} + \mathbf{b}_o) \quad (6)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}_c \mathbf{x}_t + \mathbf{R}_c \mathbf{h}_{t-1} + \mathbf{b}_c) \quad (7)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad (8)$$

where $\mathbf{i}_t$, $\mathbf{f}_t$, and $\mathbf{o}_t$ denote the input gate, forget gate, and output gate, respectively, and $\mathbf{c}_t$ is the memory cell state. Through this gating mechanism, the network can effectively retain key semantic information in long sequences. In this paper, the high-level features output by the encoder are first reshaped into sequence form and input to the xLSTM module for semantic modeling. This process can capture cross-region semantic associations, forming more stable global feature representations.

Since the SS2D encoder branch primarily extracts spatial structural information, while the xLSTM branch focuses on semantic memory modeling, the two types of features exhibit differences in representation. To achieve effective fusion, we design a bidirectional cross-attention mechanism that enables mutual feature refinement between the two branches. The intuition behind this design is that structural features can benefit from semantic guidance to resolve ambiguities in low-contrast regions, while semantic features can be grounded by structural details to improve boundary precision. The cross-attention operation allows each branch to query the other and adaptively incorporate complementary information. Formally, let $\mathbf{F}_{\text{ss2d}} \in \mathbb{R}^{L \times d}$ denote the output features from the SS2D branch and $\mathbf{F}_{\text{xlstm}} \in \mathbb{R}^{L \times d}$ denote the output features from the xLSTM branch, where $L = H \times W$ is the sequence length and d is the feature dimension. Two separate cross-attention operations are performed:

$$\mathbf{F}_{\text{ss2d} \rightarrow \text{xlstm}} = \text{CrossAttn}(\mathbf{F}_{\text{ss2d}}, \mathbf{F}_{\text{xlstm}}) \quad (9)$$

$$\mathbf{F}_{\text{xlstm} \rightarrow \text{ss2d}} = \text{CrossAttn}(\mathbf{F}_{\text{xlstm}}, \mathbf{F}_{\text{ss2d}}) \quad (10)$$

Each cross-attention operation follows the standard formulation. Taking $\text{CrossAttn}(\mathbf{F}_{\text{ss2d}}, \mathbf{F}_{\text{xlstm}})$ as an example, the query $\mathbf{Q}$ is derived from $\mathbf{F}_{\text{ss2d}}$, while the key $\mathbf{K}$ and value $\mathbf{V}$ are derived from $\mathbf{F}_{\text{xlstm}}$:

$$\mathbf{Q} = \mathbf{F}_{\text{ss2d}} \mathbf{W}_Q, \quad \mathbf{K} = \mathbf{F}_{\text{xlstm}} \mathbf{W}_K, \quad \mathbf{V} = \mathbf{F}_{\text{xlstm}} \mathbf{W}_V \quad (11)$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d}$ are learnable projection matrices. The attention output is then:

$$\mathbf{F}_{\text{ss2d} \rightarrow \text{xlstm}} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right) \mathbf{V} \quad (12)$$

The second operation $\text{CrossAttn}(\mathbf{F}_{\text{xlstm}}, \mathbf{F}_{\text{ss2d}})$ is computed symmetrically, with $\mathbf{F}_{\text{xlstm}}$ serving as the query and $\mathbf{F}_{\text{ss2d}}$ as the key and value.

The two cross-attention outputs are then concatenated along the channel dimension and projected back to the original feature dimension:

$$\mathbf{F}_{\text{fused}} = \mathbf{W}_{\text{proj}}[\mathbf{F}_{\text{ss2d} \rightarrow \text{xlstm}} \parallel \mathbf{F}_{\text{xlstm} \rightarrow \text{ss2d}}] \quad (13)$$

where $\parallel$ denotes concatenation and $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{2d \times d}$ is a learnable linear projection. Finally, a residual connection is applied to preserve the original structural information:

$$\mathbf{F}_{\text{out}} = \mathbf{F}_{\text{ss2d}} + \mathbf{F}_{\text{fused}} \quad (14)$$

This dual-path fusion strategy enables the model to simultaneously utilize the spatial structural information provided by the SS2D encoder and the long-range semantic memory provided by the xLSTM branch, resulting in more stable and accurate feature representations for complex medical images. By introducing the bidirectional cross-attention mechanism, the model achieves effective fusion of multi-source features while maintaining low computational complexity, providing richer semantic information for fine-grained segmentation in the subsequent decoding stage.

3.4. Experimental Setup and Data Preprocessing

To comprehensively evaluate the performance of the proposed method in medical image segmentation tasks, experiments are conducted on multiple public medical imaging datasets, with comparative analysis against various mainstream segmentation models. All experiments are performed under unified data preprocessing and experimental conditions to ensure fairness and reproducibility.

3.4.1. Datasets

Five representative datasets are selected, covering different modalities and anatomical regions. These datasets encompass complex scenarios ranging from regular to irregular boundaries and from single organs to multiple organs, facilitating validation of the model's generalization capability across different clinical tasks.

- segTHOR [43]: 120 thoracic CT scans (80 training, 40 testing) for organ segmentation of five structures including trachea and esophagus.
- AMOS22 [44]: 800 abdominal CT/MRI cases with 15 annotated abdominal organs.
- ACDC [45]: 200 cardiac MRI images for cardiovascular structure segmentation.
- MSD-Lung [46]: 96 thin-slice CT scans of non-small cell lung cancer patients for lung tumor segmentation.
- KiTS23 [47]: 599 CT cases of kidney and tumors for renal and tumor segmentation.

3.4.2. Data Preprocessing

To strictly prevent data leakage, all data preprocessing is performed after the data has been split. Specifically, all cases are first randomly divided into training and testing sets in an 8:1:1 ratio; subsequently, all preprocessing operations—including intensity normalization, extraction of valid slices, and resolution standardization—are performed independently for each case within their respective subsets. Operations that rely on annotation information for valid slice extraction are also performed after splitting, thereby ensuring that all 2D slices generated from the same patient belong to the same data subset.

Significant differences exist across datasets in imaging quality, voxel intensity ranges, and annotation granularity. To ensure training stability and improve data utilization, a unified preprocessing pipeline is designed:

- Label cleaning and small structure removal: 3D connected component analysis is used to remove artifact structures smaller than 1000 pixels, and small regions smaller than

100 pixels are removed at the 2D slice level. Multi-tumor cases are independently instance-labeled to ensure semantic consistency.

- Intensity normalization: For CT data, clinical window width and level (WL = 40, WW = 400) are applied for intensity clipping, followed by unified normalization to [0, 255]. For MRI data, non-background region windowing using the 0.5–99.5% percentile range is applied to suppress extreme noise.
- Effective slice extraction: Slices containing organs/lesions are automatically cropped by detecting non-zero voxel ranges in annotations, removing irrelevant empty slices to improve training efficiency.
- Spatial resolution unification: All slices are interpolated to a unified resolution of 1024×1024 using nearest-neighbor interpolation to avoid label boundary smoothing.
- Sample expansion: Each 3D case is transformed into multiple 2D slices with clear anatomical semantics after preprocessing, expanding the training sample scale from case-level to slice-level, effectively alleviating the sample scarcity problem faced by deep learning models in medical imaging scenarios.

All quantitative metrics reported in this paper are calculated at the level of two-dimensional slices. For each slice in the test set, the Dice similarity coefficient, intersection over union, 95% Hausdorff distance, and average surface distance are calculated independently; the final results are obtained as the arithmetic mean over all test slices. It should be noted that since the dataset is split at the case level, all slices from a single patient are fully assigned to either the training set or the test set. Therefore, averaging at the slice level does not introduce cross-case information leakage, and the evaluation results accurately reflect the model's generalization performance on independent cases.

3.4.3. Prompting Protocol

All experiments use bounding box prompts as spatial guidance to simulate user-provided input in interactive segmentation scenarios. For each foreground object, the prompt is defined as the tightest axis-aligned bounding box enclosing the corresponding ground-truth mask.

During training, random perturbations are applied to the bounding box coordinates to simulate the variability of user annotations. Specifically, each bounding box is independently shifted by 0–10 pixels with a 50% probability, enhancing robustness against imperfect prompts.

During inference, bounding boxes are generated using the same strategy without perturbations, approximating ideal user-provided inputs. This protocol follows common practices in promptable segmentation and is consistently applied to all comparison methods.

3.4.4. Experimental Environment

During model training, the officially released pre-trained weights of MedSAM V1(March 2024) are loaded for initialization. The prompt encoder and mask decoder modules are retained and frozen, while the image encoder is redesigned as the proposed state-enhanced encoder and trained on the target datasets. This setup constitutes a fine-tuning procedure, where only the redesigned encoder parameters are updated, while the prompt learning components remain fixed. This design choice serves two purposes: (1) it isolates the contribution of the encoder redesign from the prompt learning capability of MedSAM, ensuring that any observed performance improvement can be attributed to the proposed encoder architecture; (2) it preserves the generalization ability inherited from MedSAM, which is essential for fair comparison with baseline methods that also rely on pre-trained weights.

All experiments are conducted under the Linux Ubuntu 20.04 LTS operating system with the following hardware configuration: NVIDIA GeForce RTX 3090 GPU (NVIDIA Corporation, Santa Clara, CA, USA), Intel® Xeon® Gold 6226R CPU @ 2.90GHz (Intel Corporation, Santa Clara, CA, USA), and 256 GB RAM. CUDA version 11.8 is used.

The model is implemented in PyTorch (version 2.1.0) and optimized using AdamW with an initial learning rate of 1×10^{-4} and a weight decay of 0.01. A cosine annealing scheduler is employed to gradually decrease the learning rate during training. The model is trained for 300 epochs with a batch size of 4. Early stopping is applied based on the validation Dice Similarity Coefficient (DSC) with a patience of 10 epochs, and the checkpoint with the best validation performance is selected for final evaluation. The loss function is defined as a combination of Dice loss and binary cross-entropy loss:

$$\mathcal{L} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{BCE}}.$$

Automatic mixed precision (AMP) is used to accelerate training and reduce memory consumption.

To ensure reproducibility, all experiments are conducted with a fixed random seed (2023). Each experiment is repeated three times with different random initializations, and the reported results correspond to the mean performance across runs.

To ensure fair comparison, all experiments are conducted under the same hardware and software environment. For baseline methods, results are either directly taken from the original publications or reproduced under the same experimental settings when implementations are available. During inference, all test images are uniformly resized to 1024×1024 (or corresponding 3D patch sizes), with a batch size of 1 to avoid memory overflow and ensure consistent evaluation across different models.

4. Results

This section presents the experimental evaluation of the proposed VMedSAM-X framework. The model was assessed on five publicly available medical image segmentation datasets, including segTHOR (thoracic CT), MSD-Lung (lung tumor CT), KiTS23 (kidney tumor CT), AMOS22 (abdominal CT/MRI), and ACDC (cardiac MRI). Details regarding preprocessing, training protocols, and evaluation metrics are provided in Section 3.

4.1. Quantitative Evaluation on Multiple Datasets

To comprehensively assess segmentation performance, the proposed method was compared with several representative convolutional, transformer-based, and foundation-model-based approaches, including U-Net, VM-UNet, TransUNet, SwinUNet, MedSAM, nnU-Net, UNETR, Swin-UNETR, and UX-Net. Performance was evaluated using Dice similarity coefficient (DSC), intersection over union (IoU), 95% Hausdorff distance (HD95), and average surface distance (ASD).

To assess the stability and reliability of the proposed method, all experiments for VMedSAM-X are repeated three times with different random initializations. The reported results for our method in Tables 1–5 are therefore presented as means $\pm$ standard deviations. For baseline methods, results are either cited from the original publications or reproduced as single-run best values under the same experimental conditions. Detailed per-class segmentation results for the multi-organ datasets (segTHOR, AMOS22, and ACDC) are provided in the Appendix A.

Table 1. Quantitative results on the segTHOR dataset.

Method	DSC (%)	IoU (%)	HD95 (mm)	ASD (mm)
U-Net	79.2	66.2	6.12	2.44
VM-UNet	86.1	75.9	4.85	1.91
TransUNet	80.7	67.6	5.78	2.18
SwinUNet	82.8	70.6	5.41	2.03
MedSAM	87.3	78.5	4.27	1.58
nnU-Net	88.9	80.1	4.02	1.47
UNETR	84.2	72.6	5.33	2.07
Swin-UNETR	89.6	81.2	3.94	1.41
UX-Net	90.1	82.1	3.88	1.39
Ours	90.8 ± 0.3	84.0 ± 0.4	3.87 ± 0.12	1.34 ± 0.06

Table 2. Quantitative results on the MSD-Lung dataset.

Method	DSC (%)	IoU (%)	HD95 (mm)	ASD (mm)
U-Net	74.2	61.2	7.92	3.71
VM-UNet	88.1	78.9	5.13	2.36
TransUNet	90.1	82.1	4.67	1.95
SwinUNet	89.2	80.6	4.98	2.01
MedSAM	92.5	87.2	4.52	2.82
nnU-Net	92.8	87.6	3.98	1.74
UNETR	87.4	77.7	5.91	2.65
Swin-UNETR	93.1	88.1	3.62	1.66
UX-Net	93.3	88.4	3.44	1.59
Ours	93.4 ± 0.4	88.5 ± 0.5	3.27 ± 0.15	1.64 ± 0.08

Table 3. Quantitative results on the KiTS23 dataset.

Method	DSC (%)	IoU (%)	HD95 (mm)	ASD (mm)
U-Net	88.6	79.7	8.02	3.46
VM-UNet	90.3	82.5	7.44	3.11
TransUNet	90.9	83.4	7.11	3.02
SwinUNet	91.5	84.3	6.83	2.92
MedSAM	92.4	86.7	8.37	3.52
nnU-Net	93.1	87.1	7.44	3.03
UNETR	90.4	82.6	8.33	3.61
Swin-UNETR	93.6	88.0	6.91	2.84
UX-Net	94.0	88.7	6.52	2.76
Ours	94.2 ± 0.3	89.5 ± 0.4	5.68 ± 0.20	2.83 ± 0.10

Table 4. Quantitative results on the AMOS22 dataset.

Method	DSC (%)	IoU (%)	HD95 (mm)	ASD (mm)
U-Net	84.8	73.5	6.02	2.41
VM-UNet	88.2	79.0	4.93	1.97
TransUNet	89.3	80.7	4.67	1.88
SwinUNet	90.9	83.5	4.12	1.69
MedSAM	86.9	77.8	5.50	1.86
nnU-Net	91.2	84.0	4.21	1.71
UNETR	88.4	79.4	5.83	2.03
Swin-UNETR	92.4	86.5	3.96	1.63
UX-Net	92.8	87.2	3.88	1.57
Ours	93.1 ± 0.4	87.5 ± 0.5	3.27 ± 0.15	1.38 ± 0.07

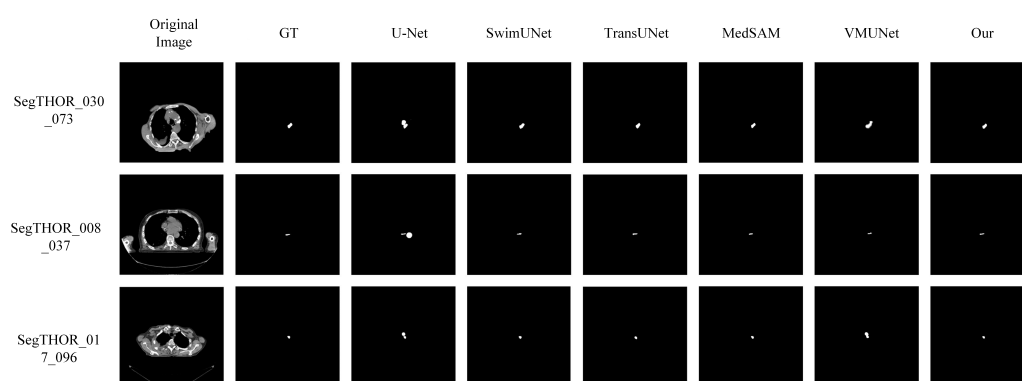
Table 5. Quantitative results on the ACDC dataset.

Method	DSC (%)	IoU (%)	HD95 (mm)	ASD (mm)
U-Net	89.6	81.2	2.93	1.22
VM-UNet	88.5	79.5	4.33	1.49
TransUNet	91.2	83.9	2.21	0.91
SwinUNet	93.0	87.5	1.39	0.66
MedSAM	90.4	82.5	3.92	1.28
nnU-Net	92.8	86.9	1.52	0.71
UNETR	90.1	82.8	2.31	0.96
Swin-UNETR	93.9	88.4	1.36	0.63
UX-Net	93.3	87.7	1.41	0.69
Ours	95.7 ± 0.2	92.0 ± 0.3	1.18 ± 0.05	0.41 ± 0.03

4.1.1. segTHOR Thoracic Organ Segmentation

The segTHOR dataset contains small thoracic organs with ambiguous boundaries and high anatomical variability, posing challenges for accurate segmentation. As shown in Table 1, the proposed method achieved the highest DSC of 90.8% and IoU of 84.0%. In addition, the method produced the lowest HD95 (3.87 mm) and ASD (1.34 mm), indicating improved boundary localization compared with the competing approaches.

As shown in Figure 4, U-Net and TransUNet exhibit weak responses in small target regions, with incomplete detection and noticeable boundary discontinuities. SwinUNet improves overall localization but still shows boundary blurring and shrinkage in low-contrast areas. Vision Mamba U-Net (VM-UNet) and MedSAM achieve relatively complete coverage of target structures, yet under-segmentation remains observable in elongated or low-contrast regions. In contrast, the proposed method consistently maintains complete and continuous target contours across all test samples, with superior shape consistency and boundary adherence to the ground truth. These qualitative observations align with the quantitative results in Table 1, where the proposed method achieves the highest Dice (90.8%) and lowest HD95 (3.87 mm).

**Figure 4.** Qualitative segmentation results on the segTHOR dataset. From left to right: input image, ground truth, U-Net, TransUNet, SwinUNet, MedSAM, VM-UNet, and the proposed VMMedSAM-X.

4.1.2. MSD-Lung Lung Tumor Segmentation

The challenge of lung tumor segmentation in the MSD-Lung dataset stems from irregular lesion shapes and heterogeneous intensities. The quantitative results in Table 2 demonstrate that the proposed framework achieved competitive performance, obtaining the highest DSC (93.4%) and IoU (88.5%), as well as a reduced HD95 value of 3.27 mm. These results suggest improved robustness in capturing fine tumor boundaries.

Figure 5 presents representative segmentation results. U-Net detects only partial tumor regions with obvious shrinkage. SwinUNet and TransUNet produce relatively complete

tumor coverage but exhibit over-smoothing or slight boundary shifts. MedSAM and VM-UNet capture the main tumor structure but show subtle boundary offsets or incomplete filling in some cases. The proposed method accurately delineates tumor boundaries with consistent shape completeness, even in low-contrast or irregularly shaped regions. This is consistent with the quantitative results in Table 2, where the proposed method achieves the highest Dice (93.4%) and the lowest HD95 (3.27 mm).

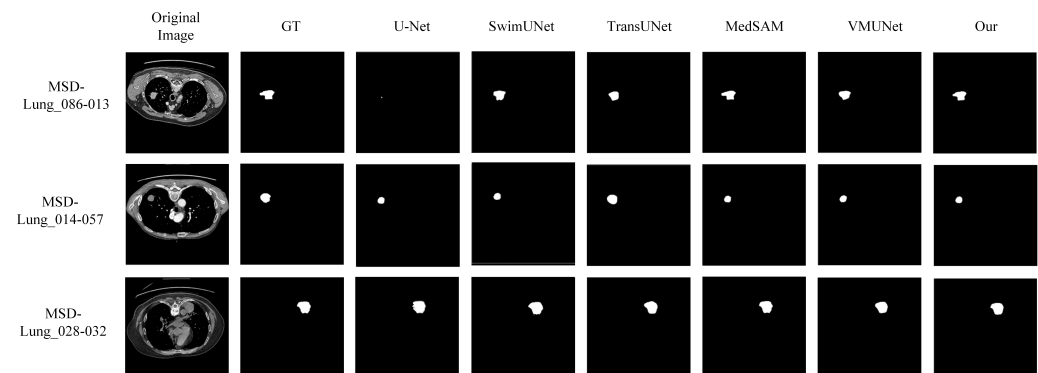


Figure 5. Qualitative segmentation results on the MSD-Lung dataset. From left to right: input image, ground truth, U-Net, TransUNet, SwinUNet, MedSAM, VM-UNet, and the proposed VMMedSAM-X.

4.1.3. KiTS23 Kidney Tumor Segmentation

The KiTS23 dataset includes cases with large inter-patient variability and complex tumor morphology. As presented in Table 3, the proposed method achieved a DSC of 94.2% and IoU of 89.5%, outperforming all compared methods. Notably, the HD95 was reduced to 5.68 mm, indicating more stable boundary predictions on challenging structures.

Figure 6 visualizes the segmentation results. U-Net produces approximate localization but suffers from shape distortion and rough boundaries. SwinUNet and TransUNet recover the overall shape better but still exhibit over-smoothing or boundary expansion in complex areas. MedSAM and VM-UNet maintain relatively consistent shape fidelity, though minor boundary offsets or over-filling occur in some slices. The proposed method generates segmentation masks that closely align with the ground truth, preserving fine structural details while maintaining shape integrity. This is supported by the quantitative results in Table 3, where the proposed method achieves the highest Dice (94.2%) and the lowest HD95 (5.68 mm).

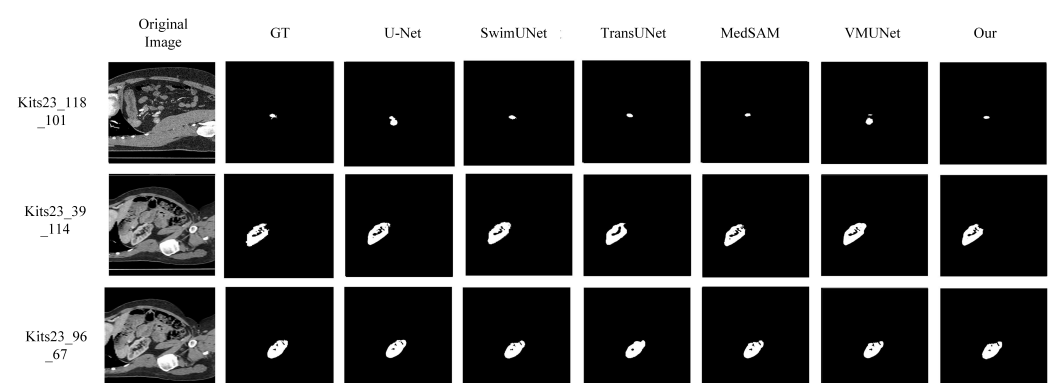


Figure 6. Qualitative segmentation results on the KiTS23 dataset. From left to right: input image, ground truth, U-Net, TransUNet, SwinUNet, MedSAM, VM-UNet, and the proposed VMMedSAM-X.

4.1.4. AMOS22 Abdominal Multi-Organ Segmentation

The AMOS22 dataset involves multi-organ segmentation in abdominal CT and MRI, where organs are often closely adjacent. Table 4 shows that the proposed model achieved

a DSC of 93.1% and IoU of 87.5%, with lower HD95 and ASD values than the baseline methods. This indicates that the proposed architecture is capable of handling complex inter-organ boundaries.

Figure 7 shows representative segmentation results. U-Net detects target regions but exhibits noticeable shrinkage and boundary deviation. SwinUNet and TransUNet improve overall localization but still show boundary ambiguity in adjacent organ regions. MedSAM and VM-UNet achieve relatively complete segmentation with good shape consistency, though minor boundary expansion or detail loss can be observed in some slices. The proposed method produces segmentation results that closely match the ground truth, with accurate boundary delineation even in regions with close organ proximity. This is consistent with the quantitative results in Table 4, where the proposed method achieves the highest Dice (93.1%) and the lowest HD95 (3.27 mm).

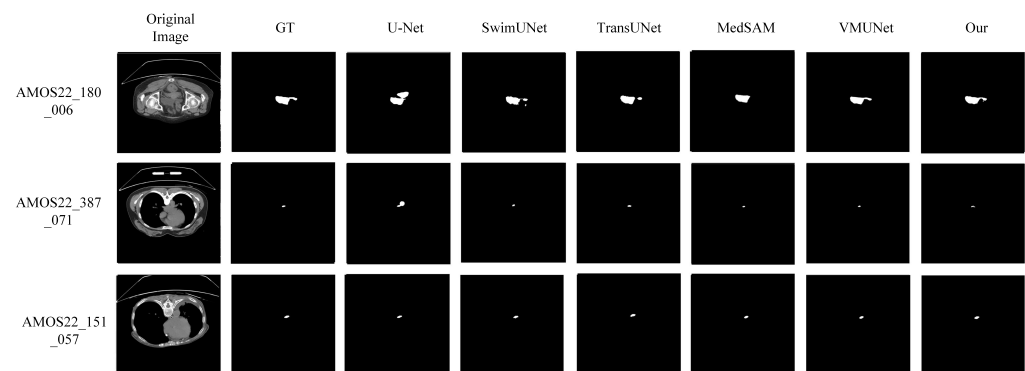


Figure 7. Qualitative segmentation results on the AMOS22 dataset. From left to right: input image, ground truth, U-Net, TransUNet, SwinUNet, MedSAM, VM-UNet, and the proposed VMMedSAM-X.

4.1.5. ACDC Cardiac MRI Segmentation

Cardiac MRI segmentation requires accurate delineation of highly deformable structures across cardiac phases. As shown in Table 5, the proposed approach achieved the highest DSC (95.7%) and IoU (92.0%), while also obtaining the lowest HD95 (1.18 mm) and ASD (0.41 mm). These results suggest that the model generalizes well to dynamic anatomical structures.

Figure 8 presents segmentation results on the ACDC dataset. U-Net detects the approximate cardiac region but exhibits boundary blurring, especially near the myocardium–blood pool interface. SwinUNet and TransUNet achieve better overall shape recovery but still show local boundary discontinuities. MedSAM and VM-UNet capture the main cardiac structures with relatively stable boundaries, though minor boundary offset remains in some slices. The proposed method generates segmentation masks with superior boundary precision and shape consistency, closely matching the ground truth even in challenging regions such as the right ventricular wall. This aligns with the quantitative results in Table 5, where the proposed method achieves the highest Dice (95.7%) and the lowest HD95 (1.18 mm).

4.2. Computational Complexity Analysis

In addition to segmentation accuracy, computational efficiency is an important consideration for clinical deployment. Table 6 compares the proposed method with representative models in terms of floating-point operations (FLOPs), number of parameters, and inference time. All complexity statistics are computed using the fvcore library with an input resolution of 1024×1024 and a batch size of 1, which corresponds to the actual inference setting used in our experiments. For encoder-based comparisons, FLOPs are measured on

the encoder component only, excluding the prompt encoder and mask decoder, to isolate the efficiency gain from the redesigned architecture.

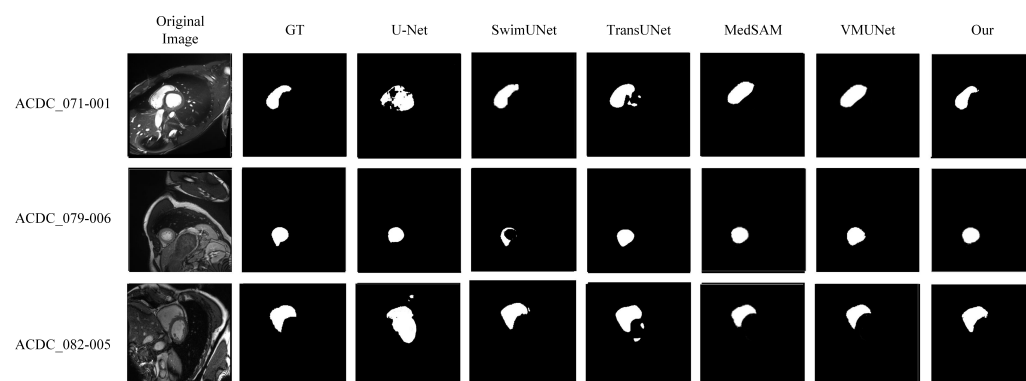


Figure 8. Qualitative segmentation results on the ACDC dataset. From left to right: input image, ground truth, U-Net, TransUNet, SwinUNet, MedSAM, VM-UNet, and the proposed VMMedSAM-X.

The proposed encoder requires 17.36 G FLOPs, which is substantially lower than the ViT-based MedSAM encoder (369.44 G). Compared with VM-UNet (22.41 G), the proposed model reduces FLOPs by approximately 22% while maintaining higher segmentation accuracy. The total number of parameters of the proposed framework is 110.87 M, slightly higher than VM-UNet but still lower than Swin-UNETR. Furthermore, the average inference time on a single 1024×1024 image is 28.6 ms, which is 79.1% faster than MedSAM and 16.4% faster than VM-UNet, demonstrating improved computational efficiency suitable for real-time clinical applications.

Table 6. Computational complexity comparison under 1024×1024 input resolution with batch size 1. FLOPs are measured on the encoder component only.

Method	FLOPs (G)	Params (M)	Inference Time (ms)
U-Net	28.54	31.04	18.6
TransUNet	124.67	105.32	52.3
SwinUNet	96.82	89.45	47.1
MedSAM (ViT-B)	369.44	89.67	136.8
VM-UNet	22.41	98.63	34.2
Swin-UNETR	218.53	112.36	78.5
Ours	17.36	110.87	28.6

4.3. Ablation Study

To investigate the contribution of each component in the proposed framework, ablation experiments were conducted on the segTHOR dataset, as summarized in Table 7.

The baseline model achieved a DSC of 79.2%. By introducing the SS2D module, the performance improved significantly to 86.8%, demonstrating the effectiveness of state space modeling in capturing long-range spatial dependencies. Further incorporating the xLSTM module led to additional improvement (88.9%), indicating that semantic memory enhances feature representation. Finally, integrating the dual-path cross-attention module resulted in the best performance of 90.8%.

Overall, each component contributes progressively to the final performance, and their combination leads to the most effective representation.

Table 7. Ablation study results on the segTHOR dataset.

Model	SS2D	xLSTM	Cross-Attention	DSC (%)
Baseline				79.2
#1	✓			86.8
#2	✓	✓		88.9
VMMedSAM-X	✓	✓	✓	90.8

4.4. Summary of Experimental Findings

Experimental evaluations on five datasets support three key conclusions regarding the proposed VMMedSAM-X framework.

First, the dual-branch encoder design offers a favorable trade-off between segmentation accuracy and computational cost. Across all evaluated datasets—covering single-lesion tasks and multi-organ scenarios, as well as CT and MRI modalities—VMMedSAM-X consistently matches or exceeds the performance of competing baselines, while reducing encoder FLOPs by over 95% compared to ViT-based MedSAM encoders. The ablation studies in Table 7 confirm that this performance stems from the incremental contributions of the SS2D pathway, the xLSTM branch, and the cross-attention mechanism.

Second, the complementary nature of the two encoder branches contributes to robust performance across structures of varying size and morphology. The SS2D pathway efficiently propagates spatial context, a property that benefits the delineation of elongated or irregular structures, while the xLSTM branch preserves semantic continuity over extended spatial ranges. The per-class results reported in the Appendix A support this observation: the method performs reliably on large, well-defined organs and on smaller, more variable structures where spatial ambiguity is often more pronounced.

Third, the consistency observed across datasets with different imaging characteristics indicates that the state-enhanced encoder generalizes effectively across the evaluation scope. The framework maintains stable performance on CT and MRI data as well as on thoracic, abdominal, and cardiac anatomical structures, without requiring modality-specific architectural adjustments.

Several limitations should be noted. The parameter count of VMMedSAM-X (110.87 M) is slightly higher than that of some baselines; this suggests that efficiency gains are achieved through reduced FLOPs and faster inference rather than through model compression. Furthermore, the current framework operates on 2D slices and does not explicitly model volumetric context. Extending the method to 3D data and further optimizing the parameter budget remain important directions for future work.

5. Conclusions

In this study, we present VMMedSAM-X: an efficient medical image segmentation framework incorporating structured state-space modelling into the MedSAM architecture. The proposed design combines SS2D-based spatial propagation and xLSTM-based semantic memory, integrating them via a dual-path cross-attention mechanism to enable effective modelling of long-range dependencies and contextual consistency. Experimental results on multiple public datasets demonstrate that the proposed method achieves competitive segmentation performance while significantly reducing computational complexity. This demonstrates its effectiveness for high-resolution medical image analysis with favorable computational efficiency. Nevertheless, several limitations remain to be addressed. The current study focuses on 2D slice-based segmentation and is primarily evaluated on CT and MRI datasets. Generalization to volumetric data and other imaging modalities is left for future exploration. Additionally, reliance on fully annotated datasets may limit applicability in scenarios with scarce labels. Future work will investigate extensions to 3D architectures,

semi-supervised learning strategies and further optimisation for efficient deployment, with the aim of enhancing both scalability and practical applicability in clinical settings.

Author Contributions: Conceptualization, H.Z. and Y.L.; methodology, H.Z. and W.L.; software, W.L. and H.Z.; validation, W.L. and H.Z.; formal analysis, W.L.; investigation, W.L.; resources, Y.L.; data curation, W.L.; writing—original draft preparation, H.Z.; writing—review and editing, Y.L. and H.Z.; visualization, H.Z.; supervision, Y.L.; project administration, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available and can be accessed as follows: (1) ACDC (Automated Cardiac Diagnosis Challenge): The cardiac MRI dataset is available at <https://www.creatis.insa-lyon.fr/Challenge/acdc/> (accessed on 15 March 2025). (2) segTHOR (Segmentation of Thoracic Organs at Risk): The CT dataset is available at <https://competitions.codalab.org/competitions/21145> (accessed on 15 March 2025). (3) MSD-Lung (Medical Segmentation Decathlon—Lung Tumor): The CT dataset is part of the Medical Segmentation Decathlon, available at <http://medicaldecathlon.com/> (accessed on 15 March 2025). (4) KiTS23 (Kidney and Kidney Tumor Segmentation Challenge 2023): The CT dataset is available at <https://kits-challenge.org/kits23/> (accessed on 15 March 2025). (5) AMOS22 (Abdominal Multi-Organ Segmentation Challenge): The CT and MRI dataset is available at <https://amos22.grand-challenge.org/> (accessed on 15 March 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASD	Average Surface Distance
CNN	Convolutional Neural Network
CT	Computed Tomography
DSC	Dice Similarity Coefficient
HD95	95% Hausdorff Distance
IoU	Intersection over Union
LSTM	Long Short-Term Memory
MedSAM	Medical Segment Anything Model
MRI	Magnetic Resonance Imaging
MSD	Medical Segmentation Decathlon
SAM	Segment Anything Model
SS2D	Two-Dimensional Selective Scanning
SSM	State Space Model
U-Net	U-shaped Convolutional Network
ViT	Vision Transformer
VM-UNet	Vision Mamba U-Net
xLSTM	Extended Long Short-Term Memory

Appendix A. Per-Class Segmentation Results

To provide a more detailed evaluation of the proposed method, we report per-class segmentation performance on the multi-organ datasets, including segTHOR, ACDC, and AMOS22.

All values are reported as means $\pm$ standard deviations over three independent runs.

All metrics are computed at the slice level. The per-class results are obtained by grouping slices according to anatomical structures and averaging the corresponding metric values.

Due to the different numbers of slices associated with each organ, the arithmetic mean of the per-class results is not expected to be identical to the overall average computed across all slices.

Table A1. Per-class segmentation results (mean $\pm$ standard deviation) of VMedSAM-X on multi-organ datasets.

Dataset	Organ/Structure	DSC (%)	IoU (%)	HD95 (mm)	ASD (mm)
segTHOR	Heart	83.87 $\pm$ 7.59	72.89 $\pm$ 10.30	4.21 $\pm$ 1.52	1.78 $\pm$ 0.61
	Aorta	95.32 $\pm$ 3.15	91.22 $\pm$ 5.51	2.48 $\pm$ 0.91	0.91 $\pm$ 0.33
	Trachea	91.48 $\pm$ 4.30	84.57 $\pm$ 6.71	4.79 $\pm$ 1.68	1.63 $\pm$ 0.57
	Esophagus	90.73 $\pm$ 4.52	84.25 $\pm$ 6.25	3.92 $\pm$ 1.41	1.28 $\pm$ 0.46
	Average (slice-level, reported in Table 1)	90.8 $\pm$ 0.3	84.0 $\pm$ 0.4	3.87 $\pm$ 0.12	1.34 $\pm$ 0.06
ACDC	Left Ventricle (LV)	96.8 $\pm$ 1.5	93.8 $\pm$ 2.2	0.97 $\pm$ 0.35	0.32 $\pm$ 0.11
	Right Ventricle (RV)	93.5 $\pm$ 2.0	87.8 $\pm$ 3.1	1.42 $\pm$ 0.51	0.51 $\pm$ 0.18
	Myocardium (MYO)	96.2 $\pm$ 1.8	92.7 $\pm$ 2.8	1.15 $\pm$ 0.41	0.40 $\pm$ 0.14
	Average (slice-level, reported in Table 4)	95.7 $\pm$ 0.2	92.0 $\pm$ 0.3	1.18 $\pm$ 0.05	0.41 $\pm$ 0.03
AMOS22	Spleen	93.91 $\pm$ 6.21	89.04 $\pm$ 9.03	2.45 $\pm$ 0.88	0.82 $\pm$ 0.29
	Right Kidney	94.12 $\pm$ 4.11	89.14 $\pm$ 6.55	2.78 $\pm$ 1.02	0.95 $\pm$ 0.34
	Left Kidney	94.25 $\pm$ 3.25	89.30 $\pm$ 5.53	2.82 $\pm$ 0.99	0.98 $\pm$ 0.33
	Gallbladder	87.49 $\pm$ 8.11	77.57 $\pm$ 11.43	6.12 $\pm$ 2.21	2.15 $\pm$ 0.78
	Esophagus	82.16 $\pm$ 10.06	70.81 $\pm$ 12.88	7.45 $\pm$ 2.63	2.72 $\pm$ 0.98
	Liver	94.09 $\pm$ 5.08	89.21 $\pm$ 7.86	2.53 $\pm$ 0.91	0.88 $\pm$ 0.31
	Stomach	89.22 $\pm$ 8.42	81.43 $\pm$ 11.94	4.72 $\pm$ 1.68	1.62 $\pm$ 0.58
	Aorta	93.37 $\pm$ 4.64	87.88 $\pm$ 7.22	2.61 $\pm$ 0.94	0.91 $\pm$ 0.32
	Inferior Vena Cava	87.30 $\pm$ 8.59	76.38 $\pm$ 12.21	3.48 $\pm$ 1.25	1.22 $\pm$ 0.44
	Pancreas	80.60 $\pm$ 11.22	68.79 $\pm$ 13.84	7.18 $\pm$ 2.58	2.56 $\pm$ 0.92
	Right Adrenal Gland	71.38 $\pm$ 14.68	57.35 $\pm$ 16.38	9.45 $\pm$ 3.40	3.38 $\pm$ 1.22
	Left Adrenal Gland	77.20 $\pm$ 8.90	62.90 $\pm$ 11.20	8.92 $\pm$ 3.21	3.18 $\pm$ 1.15
	Duodenum	79.03 $\pm$ 12.03	66.78 $\pm$ 14.84	8.78 $\pm$ 3.16	3.10 $\pm$ 1.12
	Bladder	90.32 $\pm$ 8.88	83.33 $\pm$ 12.36	3.25 $\pm$ 1.17	1.14 $\pm$ 0.41
	Prostate/Uterus	90.88 $\pm$ 5.98	83.79 $\pm$ 9.24	3.08 $\pm$ 1.11	1.08 $\pm$ 0.39
	Average (slice-level, reported in Table 4)	93.1 $\pm$ 0.4	87.5 $\pm$ 0.5	3.27 $\pm$ 0.15	1.38 $\pm$ 0.07

References

1. Xu, G.; Udupa, J.K.; Luo, J.; Zhao, S.; Yu, Y.; Raymond, S.B.; Peng, H.; Ning, L.; Rath, Y.; Liu, W.; et al. Is the medical image segmentation problem solved? A survey of current developments and future directions. *arXiv* **2025**, arXiv:2508.20139. [[CrossRef](#)]
2. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.W.M.; Van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [[CrossRef](#)]
3. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*; Springer: Cham, Switzerland, 2015; pp. 234–241.
4. Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2016; pp. 424–432.
5. Kabil, A.; Khoriba, G.; Yousef, M.; Rashed, E.A. Advances in medical image segmentation: A comprehensive survey with a focus on lumbar spine applications. *Comput. Biol. Med.* **2025**, *198*, 111171. [[CrossRef](#)]
6. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Virtual, 5 October 2021.

7. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306. [[CrossRef](#)]
8. Azad, R.; Rauf, Z.; Khan, A.R.; Rathore, S.; Khan, S.H.; Shah, N.S.; Farooq, U.; Asif, H.; Asif, A.; Zahoor, U.; et al. A recent survey of vision transformers for medical image segmentation. *arXiv* **2024**, arXiv:2402.04899.
9. Kumar, S.S. Advancements in medical image segmentation: A review of transformer models. *Comput. Electr. Eng.* **2025**, *123*, 110099. [[CrossRef](#)]
10. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2023; pp. 4015–4026.
11. Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; Wang, B. Segment anything in medical images. *Nat. Commun.* **2024**, *15*, 654. [[CrossRef](#)] [[PubMed](#)]
12. Fan, W.; Li, A.; Xu, M.; Sun, W.; Man, F. Practical application of SAM for breast nodules segmentation. *Front. Oncol.* **2026**, *16*, 1756011. [[CrossRef](#)] [[PubMed](#)]
13. Nguyen, E.; Liu, H.; Ruan, D. Necessity and impact of specialization of large foundation model for medical segmentation tasks. *Med. Phys.* **2024**, *52*, 321–328. [[CrossRef](#)]
14. Awad, M.A.; Mabrouk, M.S.; Elnokrashy, A.F. Medical image segmentation using transformer encoders and prompt-based learning: A systematic review of adaptation strategies, performance, and challenges. In *Proceedings of the 2025 Twelfth International Conference on Intelligent Computing and Information Systems (ICICIS)*; IEEE: New York, NY, USA, 2025; pp. 281–288.
15. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752. [[CrossRef](#)]
16. Zhang, R.; Guo, H.; Tian, K.; Zhou, J.; Yan, M.; Zhang, Z.; Zhao, S. Unified medical image segmentation with state space modeling snake. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*; Association for Computing Machinery: New York, NY, USA, 2025; pp. 7825–7834.
17. Dong, X.; Zhou, B.; Yin, C.; Liao, I.Y.; Jin, Z.; Xu, Z. ÆMMamba: An efficient medical segmentation model with edge enhancement. *IEEE J. Biomed. Health Inform.* **2026**, *30*, 1889–1901. [[CrossRef](#)]
18. Tang, F.; Nian, B.; Li, Y.; Jiang, Z.; Yang, J.; Liu, W.; Zhou, K. MambaMIM: Pre-training Mamba with state space token interpolation and its application to medical image segmentation. *Med. Image Anal.* **2025**, *103*, 103606. [[CrossRef](#)]
19. Beck, M.; Pöppel, K.; Spanring, M.; Auer, A.; Prudnikova, O.; Kopp, M.; Klambauer, G.; Brandstetter, J.; Hochreiter, S. xLSTM: Extended long short-term memory. In *Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 107547–107603.
20. Dutta, P.; Bose, S.; Roy, S.K.; Mitra, S. Are Vision-xLSTM-embedded U-Nets better at segmenting medical images? *Neural Netw.* **2025**, *192*, 107925. [[CrossRef](#)]
21. Shen, D.; Wu, G.; Suk, H.I. Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **2017**, *19*, 221–248. [[CrossRef](#)]
22. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **2018**, *39*, 1856–1867. [[CrossRef](#)] [[PubMed](#)]
23. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999. [[CrossRef](#)]
24. Abdellaoui, C.; Belkacem, S.; Messaoudi, N. Deep learning architectures for medical image segmentation: An organized analysis of CNN-based models and uses. *Bull. Electr. Eng. Inform.* **2026**, *15*, 424–437. [[CrossRef](#)]
25. Oliveira, J.V.S.; Vieira, D.F.; Silva, M.P.; Fernandes, D.L.; Ribeiro, M.H.F.; Oliveira, H.N. Strategies for deep learning in volumetric medical imaging: A survey. In *Proceedings of SIBGRAPI*; IEEE: New York, NY, USA, 2025.
26. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2021.
27. Tay, Y.; Dehghani, M.; Bahri, D.; Metzler, D. Efficient transformers: A survey. *ACM Comput. Surv.* **2022**, *55*, 109. [[CrossRef](#)]
28. Kim, J.W.; Khan, A.U.; Banerjee, I. Systematic review of hybrid vision transformer architectures for radiological image analysis. *J. Imaging Inform. Med.* **2025**, *38*, 3248–3262. [[CrossRef](#)] [[PubMed](#)]
29. Moglia, A.; Leccardi, M.; Cavicchioli, M.; Maccarini, A.; Marcon, M.; Mainardi, L.; Cerveri, P. Generalist models in medical image segmentation: A survey and performance comparison with task-specific approaches. *Inf. Fusion* **2026**, *127*, 103709. [[CrossRef](#)]
30. Noh, S.; Lee, B.-D. A narrative review of foundation models for medical image segmentation: Zero-shot performance evaluation on diverse modalities. *Quant. Imaging Med. Surg.* **2025**, *15*, 5825–5858. [[CrossRef](#)]
31. Ayllon, E.M.; Mantegna, M.; Shen, L.; Soda, P.; Guarrasi, V.; Tortora, M. Can foundation models really segment tumors? A benchmarking odyssey in lung CT imaging. In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*; IEEE: New York, NY, USA, 2025.
32. Lee, H.H.; Gu, Y.; Zhao, T.; Xu, Y.; Yang, J.; Usuyama, N.; Wong, C.; Wei, M.; Landman, B.A.; Huo, Y.; et al. Foundation Models for Biomedical Image Segmentation: A Survey. *arXiv* **2024**, arXiv:2401.07654. [[CrossRef](#)]

33. Gu, A.; Goel, K.; Ré, C. Efficiently modeling long sequences with structured state spaces. Presented at the International Conference on Learning Representations (ICLR), Virtual Event, 25–29 April 2022.
34. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. VMamba: Visual state space model. *arXiv* **2024**, arXiv:2401.10166.
35. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision Mamba: Efficient visual representation learning with bidirectional state space model. *arXiv* **2024**, arXiv:2401.09417. [[CrossRef](#)]
36. Sun, Y.; Wang, J.; Yin, R. From channel-spatial attention to state space models: A review of evolving mechanisms in tumour segmentation. *Clin. Transl. Discov.* **2026**, *6*, e70127. [[CrossRef](#)]
37. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)]
38. Qayyum, A.; Mazher, M.; Niederer, S.A. Assessing self-supervised xLSTM-UNet architectures for head and neck tumor segmentation. In *Challenge on Head and Neck Tumor Segmentation for MRI-Guided Applications*; Springer: Cham, Switzerland, 2024; pp. 166–178.
39. Novikov, A.A.; Major, D.; Wimmer, M.; Lenis, D.; Bühler, K. Deep sequential segmentation of organs in volumetric medical scans. *IEEE Trans. Med. Imaging* **2018**, *38*, 1207–1215. [[CrossRef](#)] [[PubMed](#)]
40. Ter-Sarkisov, A. One shot model for COVID-19 classification and lesions segmentation in chest CT scans using long short-term memory network with attention mechanism. *IEEE Intell. Syst.* **2022**, *37*, 54–64. [[CrossRef](#)]
41. Guo, M.-H.; Xu, T.-X.; Liu, J.-J.; Liu, Z.-N.; Jiang, P.-T.; Mu, T.-J.; Zhang, S.-H.; Martin, R.R.; Cheng, M.-M.; Hu, S.-M. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [[CrossRef](#)]
42. Liu, X.; Zhang, C.; Zhang, L. Vision Mamba: A comprehensive survey and taxonomy. *arXiv* **2024**, arXiv:2405.04404. [[CrossRef](#)]
43. Lambert, Z.; Petitjean, C.; Dubray, B.; Kuan, S. SegTHOR: Segmentation of thoracic organs at risk in CT images. In *Proceedings of the 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*; IEEE: New York, NY, USA, 2020; pp. 1–6.
44. Ji, Y.; Bai, H.; Ge, C.; Yang, J.; Zhu, Y.; Zhang, R.; Li, Z.; Zhanng, L.; Ma, W.; Wan, X.; et al. AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*; IEEE: New York, NY, USA, 2022; Volume 35, pp. 36722–36732.
45. Bernard, O.; Lalande, A.; Zotti, C.; Cervenansky, F.; Yang, X.; Heng, P.A.; Cetin, I.; Lekadir, K.; Camara, O.; Ballester, M.A.G.; et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Trans. Med. Imaging* **2018**, *37*, 2514–2525. [[CrossRef](#)] [[PubMed](#)]
46. Antonelli, M.; Reinke, A.; Bakas, S.; Farahani, K.; Kopp-Schneider, A.; Landman, B.A.; Litjens, G.; Menze, B.; Ronneberger, O.; Summers, R.M.; et al. The medical segmentation decathlon. *Nat. Commun.* **2022**, *13*, 4128. [[CrossRef](#)]
47. Heller, N.; Isensee, F.; Tejpaul, R.; Wood, A.; Papanikolopoulos, N.; Weight, C. The 2023 Kidney and Kidney Tumor Segmentation Challenge (KiTS23). In *Kidney and Kidney Tumor Segmentation: MICCAI 2023 Challenge, KiTS 2023*; Springer: Cham, Switzerland, 2024; pp. 1–10.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.