SAM-GPT: Hilbert Curve Enhanced Mamba for Brain Lesion Segmentation and VLM-based Analysis

Jinfu Wang[1*], **Qiyuan Wang**[1*], **Yunfei Liang**[2], **Kaipeng Wang**[3] and **Jinhua Zhao**[1†]
[1]Central China Normal University Wollongong Joint Institute, Central China Normal University, China
[2]University of Chinese Academy of Sciences, China
[3]Northeastern University, China
jw717@uowmail.edu.au, 1740919441@qq.com, liangyunfei23@mails.ucas.ac.cn,
1578203728@qq.com, jzhao1101@ccnu.edu.cn

Abstract

Recent breakthroughs in Vision-Language Models (VLMs) have shown strong potential in medical analysis, but they remain limited in the brain lesion image domain, especially when pathological regions occupy a small portion of the image. This is because VLMs tend to put excessive attention on background regions that are visually similar to target regions. Based on the findings, we propose SAM-GPT, a novel framework that leverages segmentation-derived spatial priors to support VLM-based lesion classification. It first uses an enhanced segmentation model to localize pathological regions for diagnostic tasks, and then converts lesion attributes (e.g., size, pixel range, lesion location) into linguistic guidance for a vision–language model. To improve small lesion recognition, we incorporate a new Hilbert scanning method into Mamba that improves both local spatial continuity and global spatial modeling. Experiments on benchmark datasets show that our model achieves an average Dice coefficient of 72.80% on the brain lesion segmentation task and an accuracy of 80.56% on the brain disease classification task, demonstrating the effectiveness of the proposed framework. The code is available at https://github.com/1JF-Wang/sam-gpt.

1 Introduction

Brain lesions profoundly affect quality of life, with ischemia and glioma being among the most common types [Hu *et al.*, 2020]. In MRI analysis, although the lesions originate from different mechanisms, they share the same imaging characteristics such as small lesion size, sparse distribution, blurred boundaries, and low contrast against surrounding tissues, which significantly increase the difficulty of lesion localization for vision-language models. Radiological diagnosis often relies on correlating clinical information with multimodal imaging data. However, the process is labor-intensive and demands extensive expert knowledge. Consequently, radiology diagnosis accuracy is vulnerable to human error and limited physician experience [Awadalla *et al.*, 2023].

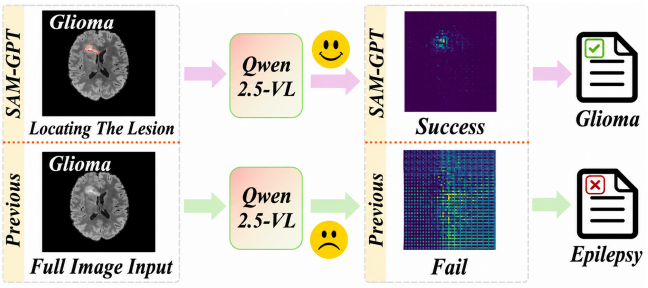

Figure 1: Comparison of glioma analysis pipelines, without prompt enhancement (bottom), VLMs fail to capture key features, leading to misclassification (epilepsy). In contrast, SAM-GPT (top) generates precise spatial prompts to capture the features successfully, enabling correct diagnosis (glioma).

Early downstream methods adopted encoder–decoder architectures, using CNNs for visual feature extraction and RNNs or Transformers for text generation [Chen *et al.*, 2020; Liu *et al.*, 2021]. Yet these models were barely trained on medical image-text pairs. Recently, the CLIP model [Radford *et al.*, 2021] has provided a new perspective for classification tasks. With the development of Vision-Language Models (VLMs), new models like Qwen2.5-VL [Bai *et al.*, 2025], GPT-4o [Hurst *et al.*, 2024] have recently emerged. These VLMs are pre-trained on vast amounts of image-text pairs. They then require only a limited number of medical pairs for fine-tuning to adapt to specific domains.

However, in complex medical scenarios, VLMs such as Qwen2.5-VL struggle to accurately localize lesions, often leading to misclassification, as shown in Figure 1. This stems from insufficient spatial grounding. Segmentation models trained on natural images perform poorly on medical lesion localization, while insufficient spatial continuity modeling further amplifies VLM misclassification for small and sparse lesions. Although State Space Models like Mamba efficiently capture long-range dependencies, standard row-by-row flattening separates vertically adjacent pixels by large intervals. This disruption limits the modeling of continuous lesion structures and hinders the capture of complex tumor boundaries [Liu *et al.*, 2024]. This motivates us to introduce a new scanning mechanism that better preserves both local continuity and global spatial relationships in medical images.

To mitigate these limitations, we propose a novel SAM-GPT framework, integrating SAM 2 [Ravi *et al.*, 2025] to guide Qwen2.5-VL in lesion localization. In order to leverage information in diagnostic images with varying contrasts (i.e., different modalities) to complement features, we design a novel modality interaction approach utilizing the Hilbert curve [Jagadish, 1997; Wu *et al.*, 2024]. Specifically, we embed a Hilbert Mamba Agent into the SAM 2 encoder for efficient multi-modal feature interaction and fusion. We further redesign the decoder as a 3D progressive upsampling architecture with a Memory Encoder and Memory Bank, replacing self-attention with Hilbert Mamba to better capture fine-grained lesion boundaries. Lastly, we utilized the lesion range, pixel area, and lesion location obtained from the segmentation results of SAM 2 to enhance Qwen2.5-VL's input prompts. Experiments on public brain lesion datasets (glioma, focal cortical dysplasia, and ischemic stroke lesions) show that SAM-GPT achieves an average Dice of 72.80% (up to 81.73%) and over 80% accuracy in lesion classification, establishing a new state-of-the-art.

Our main contributions are summarized as follows: (**i**) We propose a segmentation-guided VLM framework that converts lesion attributes into textual prompts, enabling more accurate diagnosis; (**ii**) we design a Hilbert Mamba based segmentation architecture to preserve local continuity and global dependencies; (**iii**) Extensive experiments on brain lesion datasets demonstrate that our model achieves state-of-the-art performance on multiple benchmarks and different tasks.

2 Related Works

Vision-Language Models in Medicine. Early automated medical diagnosis relied on encoder-decoder architectures integrating CNNs and RNNs [Chen *et al.*, 2020]. Later, instruction-tuned Large Language Models (LLMs) such as Med-PaLM M [Tu *et al.*, 2024] and XrayGPT [Thawakar *et al.*, 2024] have demonstrated strong clinical capabilities. More recently, general vision-language models like Qwen2.5-VL [Bai *et al.*, 2025] have further advanced visual recognition tasks, including medical image diagnosis. Despite their success in medical diagnosis, most operate on coarse and noisy features, which limits their ability to recognize small pathological regions.

Medical Segmentation Models. 3D medical segmentation architectures have transitioned from CNNs to Transformers, UNETR [Hatamizadeh *et al.*, 2022b] and SwinUNETR [Hatamizadeh *et al.*, 2022a] set strong baselines. The Segment Anything Model (SAM) [Kirillov *et al.*, 2023] further introduced foundation models to the medical domain, such as MedSAM [Ma *et al.*, 2024]. Most recently, SAM 2 [Ravi *et al.*, 2025] has been adapted for 3D segmentation, improving the segmentation ability. However, when applied directly to small and sparse brain lesions, these models often lead to coarse localization.

Visual State Space Models. State Space Models (SSMs), particularly Mamba [Gu and Dao, 2024], enable efficient long-sequence modeling. In the vision domain, VMamba [Liu *et al.*, 2024] introduced 2D scanning mechanisms, which were extended by SegMamba [Xing *et al.*, 2024] to 3D medical segmentation tasks, demonstrating competitive performance compared with Vision Transformers. However, traditional scan methods disrupt 2D spatial adjacency during sequence flattening, causing models to struggle to capture fine-grained features.

To address these issues, our framework captures local spatial dependencies and precise small lesion features to generate a more accurate diagnosis, achieving state-of-the-art performance in both lesion segmentation and diagnostic classification.

3 Methods

Figure 2 illustrates SAM-GPT, a two-stage framework for brain lesion segmentation and VLM-based diagnostic classification. Specifically, in the first stage, HilbertSAM outlines the lesion boundaries from the input image sequence $X = \{x_i^m | m = 1, 2, ..., M; i = 1, 2, ..., N\}$, where M denotes the number of modalities (e.g., T1, T2, FLAIR) and N denotes the flattened spatial tokens of each image slice. The process generates size data and spatial location information for the lesions. Qwen2.5-VL utilizes these spatial priors to enhance the prompt, classifies the lesions, and combines the lesion information from the first stage with the classification information from the second stage to generate the final diagnostic result.

HilbertSAM primarily consists of a Hilbert-Enhanced Encoder and a Hilbert-Enhanced Decoder. Specifically, the Hilbert-Enhanced Encoder differs from the SAM 2 Encoder by supporting both single-modal and multi-modal inputs. To enhance the recognition capability of small lesions, we modified the SAM 2 decoder, the Hilbert-Enhanced Decoder effectively leverages features stored in the Memory Bank while incorporating Hilbert Mamba to scan different windows and perform progressive upsampling, ultimately achieving precise localization of lesion spatial positions.

3.1 Preliminaries

State Space Models. Structured State Space Models (S4) [Gu *et al.*, 2022] and Mamba [Gu and Dao, 2024] are sequence models that utilize a hidden state $h(t) \in \mathbb{R}^{N\times 1}$ to map a one-dimensional input signal $x(t) \in \mathbb{R}$ to an output $y(t) \in \mathbb{R}$, where N indicates the dimensionality of the hidden state. Mathematically, the state space model (SSM) [Chen, 1984] transforms input data through the ordinary differential equations (ODE) $h'(t) = Ah(t) + Bx(t)$ and $y(t) = Ch(t)$, where $A \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^{N\times 1}$, and $C \in \mathbb{R}^{1\times N}$ are the system's evolution and projection parameters. For deep learning applications, the continuous system is discretized via Zero-Order Hold (ZOH), resulting in a discrete model, $h_k = \bar{A}h_{k-1} + \bar{B}x_k$ and $y_k = Ch_k$, where the parameters are transformed as $\bar{A} = \exp(\Delta A)$ and $\bar{B} = (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B$. This framework is further leveraged by converting to a convolutional neural network (CNN) form for parallel computation, represented as $\mathbf{y} = \mathbf{x} \star \mathbf{M}$, with $\mathbf{M}$ being the convolution kernel derived from the state space model.

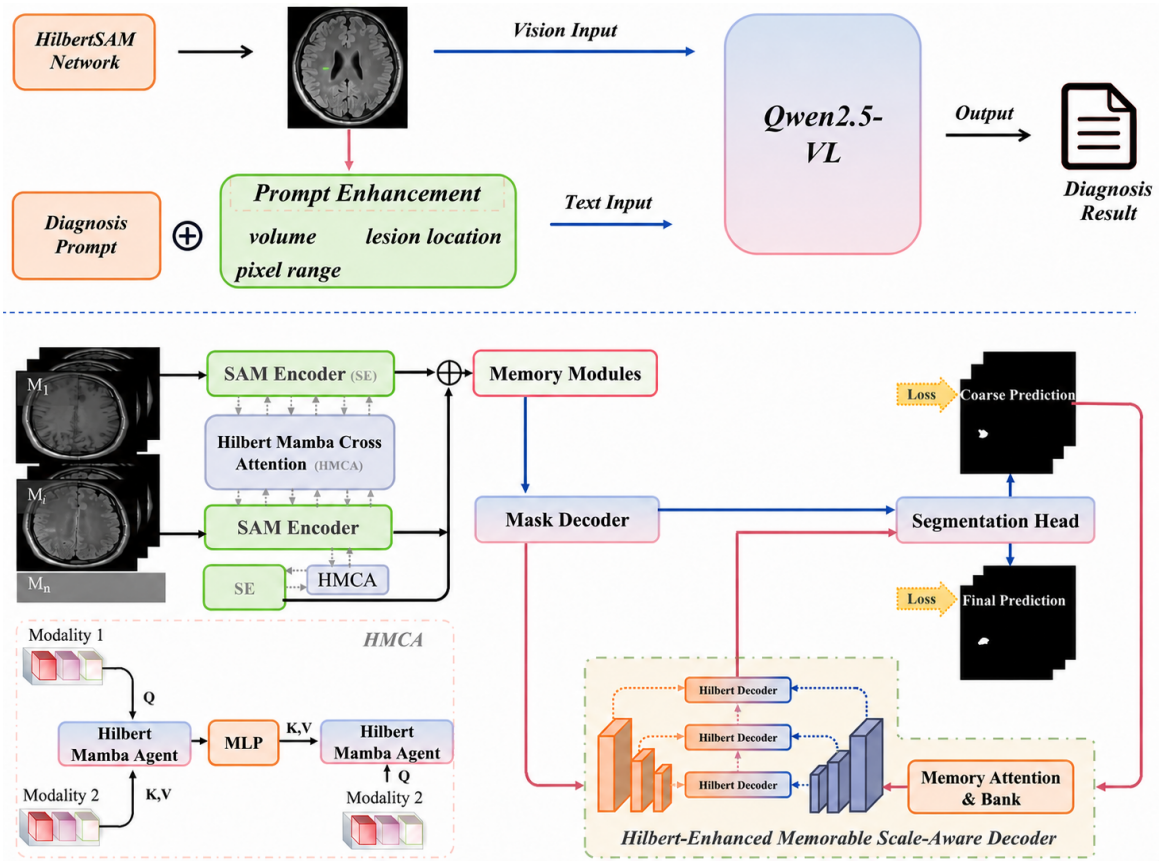

Figure 2: The framework of SAM-GPT (top) integrates HilbertSAM with Qwen2.5-VL for brain lesion analysis. HilbertSAM first delineates lesion boundaries to extract spatial priors (e.g., size, location), which are then converted into enhanced textual prompts to guide Qwen2.5-VL. The structure of HilbertSAM (bottom) features a Hilbert-Enhanced Encoder that fuses multi-modal inputs (M_1,...,M_n) via the Hilbert Mamba Cross Attention (HMCA) module, utilizing 3D Hilbert scanning for efficient feature interaction. The Hilbert-Enhanced Decoder incorporates a memory bank and Hilbert Mamba to progressively upsample features, ensuring precise localization of small lesions.

Hilbert Curve. The Hilbert curve is a type of space-filling curve (SFC) [Moon *et al.*, 2001] that can be represented as $\sigma : [0,1] \rightarrow [0,1] \times [0,1]$, mapping any point in the one-dimensional interval $[0,1]$ to coordinates in a two-dimensional unit square. We denote the order of the Hilbert curve as n, which approximately corresponds to the height and width of each frame in discrete cases. For any two points u and v in the interval $[0,1]$, their spatial-to-linear ratio (SLR) is defined as:

$$SLR = \frac{\|\sigma(u) - \sigma(v)\|^2}{|u - v|}. \quad (1)$$

The expansion factor (DF) of a space-filling curve is defined as the upper bound of the SLR. A lower DF value indicates a tighter mapping in the unit square, aligning with the requirements for preserving locality. Research indicates that the DF of the Hilbert curve is 6, while the DF of the conventional Zigzag curve is given by $4^n - 2^{n+1} + 2$, which increases towards infinity as the order n of the curve increases. Therefore, as image resolution improves, the Hilbert curve can more effectively maintain locality when mapping any two points in a one-dimensional sequence to multi-dimensional space compared to the Zigzag curve. To establish global and local correlations, we attempt to incorporate the Hilbert curve into state-space models (SSMs).

3.2 Hilbert Mamba

As shown in Figure 3, we first construct a 3D Hilbert curve that adheres to the principles of Hilbert curves, scanning through every point in the 3D Hilbert space. The 3D Hilbert algorithm recursively subdivides the 3D space into cubic subspaces and aggregates the generated Hilbert segments into a complete traversal curve. Subsequently, we convert each point on the 3D curve into a unique one-dimensional index. This indexing process flattens 3D space following the scanning trajectory of the Hilbert curve, mathematically represented as:

$$\text{Index} = \hat{x} \cdot Q \cdot T + \hat{y} \cdot T + \hat{z}, \quad (2)$$

where $(\hat{x}, \hat{y}, \hat{z})$ denotes the coordinate index of each point, and Q represents the width length of the input features. Based on the Hilbert index, the features $\hat{F} \in \mathbb{R}^{C \times T \times H/4 \times W/4}$ (C, H, W respectively denote channel, height, and width dimensions, T denotes the temporal dimension, i.e., slice depth) are flattened into a one-dimensional sequence $\hat{V} \in$

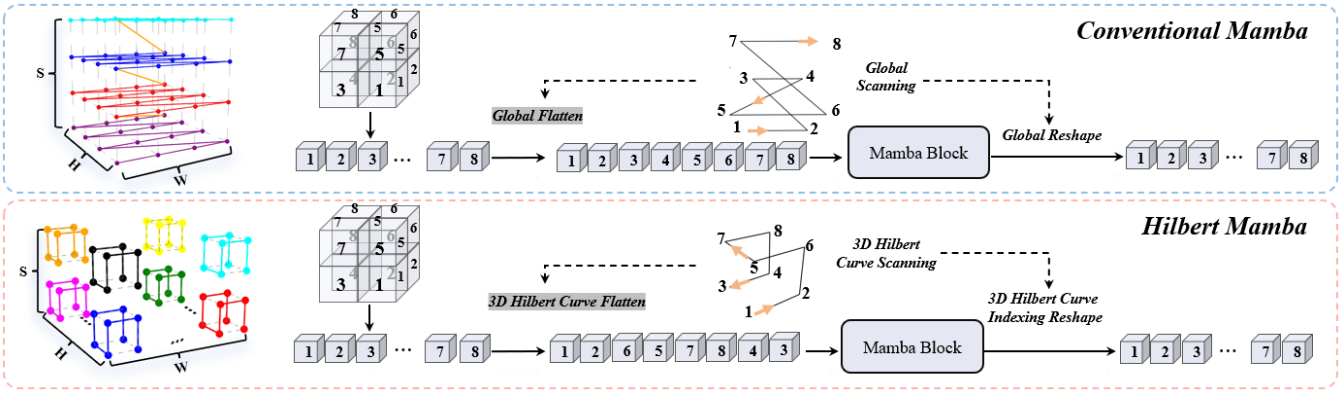

Figure 3: Compared with conventional scanning (top), Hilbert scanning (bottom) preserves local spatial continuity (as shown in the figure, spatially adjacent elements are assigned to nearby positions in one-dimensional sequence). Then the sequence is processed by the Mamba block and reshaped using Hilbert indexing, enabling denser aggregation of local spatiotemporal features for small and sparse lesion segmentation.

$\mathbb{R}^{C\times(H/4\times W/4\times T)}$, which enhances local dependencies compared to the global sequence V. After the local feature enhancement via Hilbert scanning, the sequence is input into the Mamba block, which leverages spatiotemporal corrections in a local context. Ultimately, we reshape the output sequence back to its original dimensions, defined as $\hat{F}' \in \mathbb{R}^{C\times T\times H/4\times W/4}$. Unlike other scanning methods, the Mamba enhanced with Hilbert scanning places greater emphasis on capturing local dependencies across both spatial and temporal dimensions.

3.3 Hilbert Mamba Applications

Hilbert-Enhanced Encoder. To enable multimodal fusion in the SAM 2 encoder and improve sensitivity to small lesions, we designed a novel Hilbert Mamba Cross Attention specifically for the Encoder of SAM 2 as illustrated in Figure 2. The Hilbert Mamba Cross Attention primarily consists of our newly designed Hilbert Mamba Agent. The Hilbert Mamba Agent dynamically aggregates cross-modal features through Hilbert-based localized scanning and QKV interaction provided by different modalities.

After the local features are scanned by the Hilbert Mamba, to facilitate interaction between different modalities, the Hilbert Mamba Agent incorporates a novel approach to Agent queries. Specifically, given two input modalities m_1 and m_2, the Hilbert Mamba Agent accepts the Query from m_1 and the Key and Value from m_2. We introduce a new weight B that replaces the Q values of m_1. To prevent a significant increase in computational cost due to a large N, the dimension of B differs from that of Q, where B has a dimension of n, and the value of n is much smaller than N. Mathematically, the interaction of the modality can be expressed as:

$$\begin{aligned} O_1 =&\sigma(Q_{m1}B_1^T + C_2)\sigma(B_1K_{m2}^T + C_1)V_{m2} \\ &+DWC(V_{m2}). \end{aligned} \tag{3}$$

where $C_1 \in \mathbb{R}^{n\times N}$, $C_2 \in \mathbb{R}^{N\times n}$, $Q_{m1}, K_{m2}, V_{m2} \in \mathbb{R}^{N\times C}$, and $B_1 \in \mathbb{R}^{n\times C}$. C_1 and C_2 are set as learnable parameters for biases. $\sigma(\cdot)$ denotes the Softmax function. Depthwise convolution (DWC) is introduced to mitigate the issue of insufficient feature diversity in agent attention.

To enhance the deep interaction between different imaging modalities and facilitate continuous low-level feature extraction, after the first Hilbert Mamba, we utilize a Multi-Layer Perceptron (MLP) to extract the first fusion's K and V. We then interact the Query values from m_2 with the extracted K and V values. During the Q, K, V interaction, the Hilbert Mamba scans the newly synthesized sequence again. By continuously scanning local regions, we aim to improve the segmentation of small lesions at the pixel level (small lesions may occupy only 1-10 pixels in the overall pixel space).

Hilbert-Enhanced Decoder. Hilbert-Enhanced Decoder is composed of a memory encoder, a memory bank and a Hilbert-Enhanced Scale-Aware decoder. Memory encoder and memory bank create efficient and rich memory representations and store these memory representations. Hilbert-Enhanced Scale-Aware decoder utilizes these representations to optimize the continuity of the three-dimensional feature output. Specifically, to help the decoder decode subtle features from the feature sequence, we attempt to combine a progressive decoding approach with a Hilbert Mamba enhanced 3D decoder. Specifically, we replace the self-attention mechanism in the original decoder with Hilbert Mamba, enabling the decoder to re-scan output tokens and reconstruct adjacent-token features instead of relying on global scanning. Structurally, the Hilbert-Enhanced Decoder employs three 3D decoders to fuse the pyramid feature maps of the current frame f^t from the mask decoder and the previous frame f^{t-1} from the memory bank at the same resolution through concatenation and convolution. In addition, transposed-convolution is used in the 3D decoder to perform 2x upsampling on the feature maps, which are then fused with upper feature maps. The process is formulated as follows:

$$f_j = \begin{cases} \text{Conv}(\text{Cat}(f_j^t, f_j^{t-1})), & j = \frac{1}{16}, \\ \text{Conv}(\text{Cat}(f_j^t, f_j^{t-1}, \text{UP}(f_{j/2}))), & j \in \left\{\frac{1}{8}, \frac{1}{4}\right\}. \end{cases} \tag{4}$$

where $\text{UP}(\cdot)$ represents the transposed-convolution. The

output feature maps of the Hilbert-Enhanced Scale-Aware Decoder are combined with the feature maps of the mask decoder to enhance the perception capacity of small targets.

3.4 Prompt Enhancement

Considering the Qwen2.5-VL model's limitation in capturing critical lesion information, we propose a Prompt Enhancement method. We use the HilbertSAM to accurately delineate lesion regions and generate lesion-aware spatial priors. Specifically, the segmentation results from HilbertSAM are converted into enhancement prompts, including lesion pixel regions, boundaries, and locations, to assist Qwen2.5-VL in accurate image recognition. We further analyze theoretically why enhancement prompts can improve the classification capability of Qwen2.5-VL.

We define the output of the model without enhancement prompts as $R_{\text{base}} = f(I)$, and the output of the model with enhancement prompts as $R_{\text{enhanced}} = f(I, P)$, we introduce a gain function $G(P)$ to measure the knowledge gain of Qwen2.5-VL introduced by enhancement prompts:

$$G(P) = \alpha \cdot \Delta f(I, P) + \beta \cdot \mathcal{H}(I, P), \tag{5}$$

where $\Delta f(I, P)$ is the improvement of the model's capability of prediction under the enhancement prompts, defined as:

$$\Delta f(I, P) = f(I, P) - f(I). \tag{6}$$

The term $\mathcal{H}(I, P)$ represents the information richness of the image–prompt pairs, α and β are weighting factors.

To evaluate whether the enhanced classification result is close to the ground-truth diagnosis, we define loss of knowledge as:

$$L_{\text{enhanced}} = \mathbb{E}_{(Y,I,P)}[L(Y, R_{\text{enhanced}})], \tag{7}$$

where Y denotes the ground-truth diagnosis label. The function can be further expressed as:

$$\begin{aligned} L(Y, R_{\text{enhanced}}) =& \gamma \cdot \|R_{\text{enhanced}} - Y\|^2 \\ &+ \delta \cdot \mathcal{C}(R_{\text{enhanced}}, Y), \end{aligned} \tag{8}$$

where $\|R_{\text{enhanced}} - Y\|^2$ is the Mean Squared Error (MSE) measuring the deviation between the enhanced output of Qwen2.5-VL and the ground-truth knowledge. Additionally, $\mathcal{C}(R_{\text{enhanced}}, Y)$ represents the loss of classification task between the enhanced prediction and the true label. γ and δ control the weights of the different loss components.

Consequently, the classification result can also be conceptually expressed as:

$$R_{\text{enhanced}} = R_{\text{base}} + G(P) \cdot \mathcal{P}(Y|I, P), \tag{9}$$

where $G(P) \cdot \mathcal{P}(Y|I, P)$ represents the classification gain brought by Prompt Enhancement. $\mathcal{P}(Y|I, P)$ denotes the prompt conditioned confidence of Qwen2.5-VL for category Y, namely the confidence that the prediction is correct based on I and P. The overall metric measuring loss of knowledge is defined as:

$$J_{\text{sa}} = L_{\text{enhanced}} + J_{\text{reg}}, \tag{10}$$

where J_{reg} is a regularization term intended to prevent overfitting.

Overall, the Prompt Enhancement method is proposed to provide lesion-aware spatial knowledge for Qwen2.5-VL. A lower value of J_{sa} indicates that the gain effect introduced by Prompt Enhancement is more effective and that the enhanced model output is closer to the ground truth.

4 Experiments

4.1 Analysis of Misclassification

To validate the interference of image backgrounds on Qwen2.5-VL's ability to discern images, we designed the following experiment: First, we selected 200 medical images (including 50 T1 images of brain gliomas, 50 FLAIR images of brain gliomas, 50 T1 images of epilepsy, and 50 FLAIR images of epilepsy) and directly fed them to Qwen2.5-VL for discrimination. Next, we utilized the unfinetuned SAM 2 to segment these images. Finally, we employed the MedSAM 2 [Zhu *et al.*, 2024b], to segment the images and then provided them to Qwen2.5-VL for assessment. The experimental results presented in Table 1 demonstrate that the image background significantly interferes with Qwen2.5-VL's correct identification of lesions and that precise removal of the image background shows a positive correlation with Qwen2.5-VL's accuracy in recognizing lesions.

4.2 Datasets

In the task of precise image background removal, we conducted a comparative evaluation of the Hilbert-enhanced SAM 2 network against state-of-the-art methods on three publicly available MRI brain lesion benchmark datasets (e.g., BraTS2021 [Baid *et al.*, 2021], FCD2023 [Schuch *et al.*, 2023], ISLES2022 [Petzsche *et al.*, 2022]). We utilized the data format (3D and 2D) of each dataset to ensure fair comparison. We employed metrics such as the Dice coefficient, Intersection over Union (IoU), Precision, and Recall to quantitatively compare the performance of different methods.

For the lesion classification task of Qwen2.5-VL in recognizing lesion types, considering that the raw medical image format cannot be recognized by VLMs, we randomly selected 4,800 labeled 2D slices from the constructed brain lesion dataset (the dataset includes 1,600 slices of gliomas, 1,600 slices of epilepsy, and 1,600 slices of cerebral infarction). We conducted a comparative evaluation of our SAM-GPT network on the 4,800 slices against the latest methods. We utilized ACC, PREC, RECALL, and F1 scores for the quantitative comparison of different methods.

4.3 Implementation Details

The proposed SAM-GPT was implemented using Python 3.10 and PyTorch 2.1.0 on a Ubuntu 22.04 server equipped with ten V100 GPUs. In our SAM-GPT model, we configured the initial learning rate to 1×10^{-5}. Throughout training,

Method	Acc	Recall	Prec	F1
Qwen2.5-VL	67%	66%	73%	0.69
SAM 2 & Qwen2.5-VL	61%	64%	72%	0.68
MedSAM 2 & Qwen2.5-VL	75%	79%	74%	0.76

Table 1: Impact of background interference on Qwen2.5-VL performance. Removing background using medical segmentation models significantly improves classification accuracy.

Method	Year	BraTS2021				FCD2023				ISLES2022			
		Dice(%)	IoU(%)	Prec(%)	Recall(%)	Dice(%)	IoU(%)	Prec(%)	Recall(%)	Dice(%)	IoU(%)	Prec(%)	Recall(%)
UNETR	2022	52.26	48.09	56.75	49.32	26.68	16.09	33.45	25.56	37.27	23.76	39.46	34.32
SwinUNETR	2021	53.27	47.14	54.43	53.33	28.64	16.30	29.71	27.86	37.62	24.06	33.98	41.91
nnFormer	2021	62.36	54.80	63.50	61.45	43.88	30.35	46.91	45.61	50.12	35.21	52.83	46.97
MixUNETR	2025	61.98	52.14	59.65	62.31	38.61	24.52	35.56	41.72	47.74	32.95	49.21	46.39
STU-NET	2023	59.80	51.06	60.32	57.11	27.42	16.75	28.81	26.18	22.54	15.82	25.13	28.80
MedSAM	2024	68.44	55.08	69.60	68.21	56.43	40.77	51.20	58.61	66.65	54.27	64.51	**71.46**
SAM Med 2D	2024	71.38	61.46	73.93	67.35	57.51	42.96	57.30	58.91	67.02	52.16	65.19	69.34
SAM 2 (zero-shot)	2025	00.03	00.01	00.02	00.03	00.03	00.01	00.01	00.02	00.01	00.01	00.02	00.04
SAM Med 3D	2024	73.56	64.77	75.89	71.33	54.01	37.66	51.09	55.22	69.34	56.40	70.77	67.54
MedSAM 2	2024	76.17	68.43	**81.33**	75.22	64.22	50.13	**62.78**	67.01	68.57	57.31	69.65	66.45
FS-MedSAM2	2024	78.84	70.71	79.90	72.11	65.70	**50.98**	61.01	69.53	65.15	55.07	65.98	64.75
HilbertSAM (Ours)	2025	**81.73**	**72.80**	80.29	**79.56**	**66.31**	49.83	61.39	**73.21**	**70.38**	**57.43**	**71.04**	69.59

Table 2: Comparison of segmentation methods on brain lesion datasets. The bolded parts indicate the method with the best performance for that metric. SAM 2 (zero-shot) indicates the original SAM 2 model evaluated without medical fine-tuning or prompts.

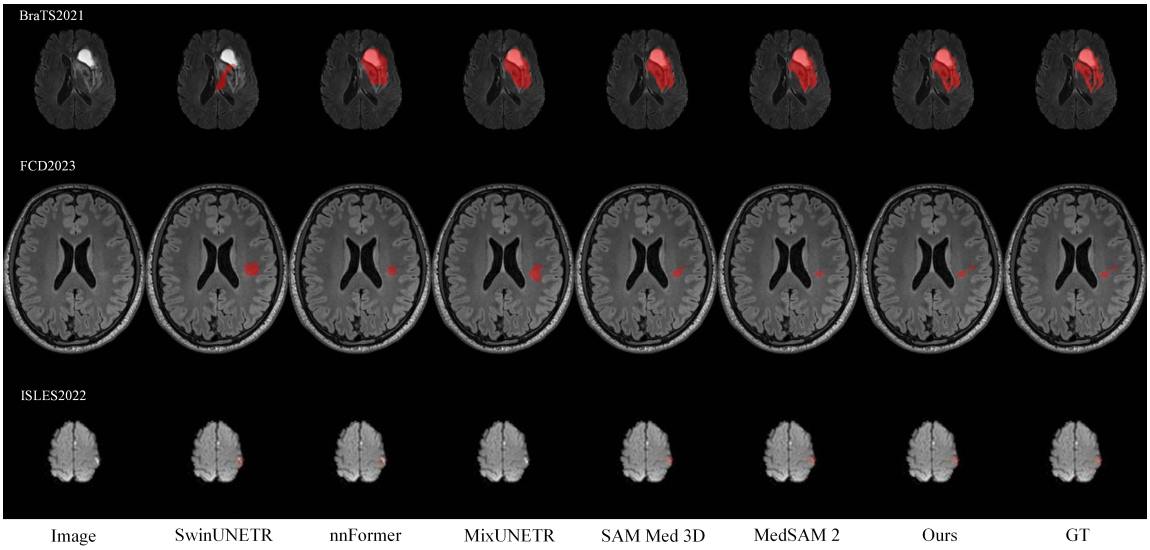

Figure 4: Visual comparison of segmentation results. Compared with existing methods, our segmentation is more consistent with the GT (ground truth) and achieves more accurate localization of small lesion regions.

a cosine annealing scheduler is implemented, which adjusts the learning rate via cosine annealing. A warm-up phase is employed for the first five epochs, during which the learning rate is gradually increased to stabilize the training process.

4.4 Comparisons with State-of-the-art Methods

To evaluate segmentation performance, we compare HilbertSAM with eleven state-of-the-art methods, including five single-modality 3D segmentation networks (e.g., UNETR [Hatamizadeh *et al.*, 2022b], SwinUNETR [Hatamizadeh *et al.*, 2022a], nnFormer [Zhou *et al.*, 2023], MixUNETR [Shen *et al.*, 2025], STU-NET [Huang *et al.*, 2023]), six single-modality segmentation networks based on SAM [Kirillov *et al.*, 2023] and SAM 2 (e.g., SAM Med 2D [Cheng *et al.*, 2023], SAM 2 [Ravi *et al.*, 2025], SAM Med 3D [Wang *et al.*, 2025], MedSAM 2 [Zhu *et al.*, 2024b], FSMedSAM 2 [Bai *et al.*, 2024]). Notably, the performance of zero-shot SAM 2 baseline is limited, illustrating the gap between generic segmentation models and medical-adapted segmentation models for small and sparse lesion targets. As shown in Table 2, SAM-GPT achieved the highest Dice coefficient (81.73%), IOU (72.80%), and Recall (79.56%) in the benchmark BraTS2021 dataset tests. It also shows consistent performance on the other two datasets, indicating its superior capability in brain lesion perception.

Method	Dice	IoU	Params	GFLOPs	Time
SAM 2 (zero-shot)	00.03%	00.01%	44M	128	156ms
VMamba+SAM 2	71.56%	67.91%	102M	201	294ms
EVMamba+SAM 2	75.70%	70.22%	113M	267	322ms
JamMa+SAM 2	73.56%	68.84%	**64M**	239	229ms
HilbertSAM (Ours)	**81.73%**	**72.80%**	97M	**195**	**178ms**

Table 3: Comparison of different Mamba scanning strategies integrated into SAM 2, where HM denotes Hilbert Mamba, bold values are marked among effective Mamba variants.

To validate the effectiveness of the Hilbert Mamba scanning approach, We further compare different Mamba scanning strategies integrated with SAM 2—VMamba [Liu *et al.*,

Method	HMCA	DHM	PE	ACC(%)	Prec(%)	Recall(%)	F1 Score
M1				61.34	60.74	64.81	0.63
M2	✓			72.73	68.34	74.34	0.71
M3	✓	✓		75.43	71.41	83.47	0.77
SAM-GPT	✓	✓	✓	**80.56**	**83.23**	**87.30**	**0.85**

Table 4: Quantitative results of our network and the baseline networks constructed from the ablation study ("M1" to "M3") on 4,800 brain lesion slices.

Method	ACC(%)	Recall(%)	Prec(%)	F1 Score
R2Gen	54.12	53.56	57.36	0.55
R2Gen*	61.48	70.77	60.24	0.65
XrayGPT	60.22	59.45	62.75	0.61
XrayGPT*	64.81	69.60	65.31	0.67
XrayPULSE	61.81	67.42	59.86	0.63
XrayPULSE*	68.73	73.42	67.41	0.70
Med-PaLM M	66.81	65.13	70.27	0.68
Med-PaLM M*	72.63	74.32	69.48	0.72
MiniGPT4	69.33	83.90	64.90	0.73
MiniGPT4*	74.32	80.32	74.73	0.77
Qwen2.5-VL	67.31	69.79	80.35	0.75
SAM-GPT (Ours)	**80.56**	**83.23**	**87.30**	**0.85**

Table 5: Comparison of VLM methods. The asterisk (*) indicates that the SAM-enhanced prompt is re-inputted into the VLM model.

2024], EVMamba [Pei *et al.*, 2025], and JamMa [Lu and Du, 2025]—on the BraTS2021 dataset. As shown in Table 3, Hilbert Mamba combined with SAM 2 achieves a Dice coefficient of 81.73% and an IoU of 72.80%, outperforming the other scanning methods and maintaining a good balance between model complexity and computational efficiency.

4.5 Discrimination Capability Comparison

For the task of lesion discrimination accuracy, we compare SAM-GPT with representative VLM-based methods (e.g., R2Gen [Wang *et al.*, 2023], XrayGPT [Thawakar *et al.*, 2024], XrayPULSE [OpenMedLab, 2024], Med-PaLM M [Tu *et al.*, 2024], MiniGPT4 [Zhu *et al.*, 2024a]). As shown in Table 5, SAM-GPT achieved the highest ACC(80.56%), Recall(83.23%), Prec(87.30%), and F1 scores(0.85) in testing 4,800 lesion images. To validate the effectiveness of our proposed solution for misclassification, we re-input the outputs of SAM 2 into these five models, accordingly enhancing their prompts. As shown in Table 5, all five models demonstrated a notable performance increase when augmented with inputs from SAM 2.

4.6 Visual Comparison

Figure 4 visually compares the lesion localization results predicted by our network and state-of-the-art methods on the brain lesion dataset we constructed. Compared to other methods, our network is more effective in localizing lesions.

4.7 Ablation Study

Baseline Design

We perform ablation experiments to verify the effectiveness of three critical components of our network, namely the Hilbert Mamba Cross Attention (HMCA), the Hilbert Mamba in the Decoder of SAM 2 (DHM), and the Prompt Enhancement for Qwen2.5-VL (PE). First, we define a baseline model referred to as "M1", which excludes these three primary components. Next, by incorporating HMCA, we enhance the baseline model "M1", resulting in model "M2". Building upon "M2", we further integrate the Hilbert Mamba from the SAM 2 Decoder to construct "M3". Finally, we introduce the prompt enhancement for Qwen2.5-VL. This series of experiments aims to validate the contribution of each component to the overall performance of the network.

Quantitative Comparison

The ablation study results presented in Table 4 provide quantitative results of different components of the SAM-GPT method on our constructed dataset of 4,800 brain lesion images. Specifically, compared to "M1", "M2" improves performance by introducing HMCA, achieving an 11.39% increase in ACC. The addition of DHM in "M3" leads to further improvements, particularly with Recall rising from 74.34% to 83.47%. Finally, incorporating Prompt Enhancement (PE) into the complete SAM-GPT model results in optimal performance, achieving an ACC of 80.56%, a Prec of 83.23%, a Recall of 87.30%, and an F1 Score of 0.85.

5 Conclusion

In this work, we propose SAM-GPT to address VLM misclassification caused by background interference, effectively mitigating background distractions and enhancing the VLM model's ability to distinguish lesions across different diseases. By integrating the Hilbert curve with Mamba and SAM 2, SAM-GPT captures local feature continuity and leverages multimodal inputs to improve lesion localization between small lesions and normal tissues. Additionally, we introduce a Prompt Enhancement method that leverages lesion size, pixel regions, and lesion coordinates to assist the VLM model in accumulating prior knowledge about lesions. Comprehensive experiments demonstrate that our method surpasses the current state-of-the-art in both lesion localization and discrimination capabilities.

Acknowledgments

This research was funded by the Fundamental Research Funds for the Central Universities of China (Grant No. XJ2026000301) and the Central China Normal University Wollongong Joint Institute.

Contribution Statement

Equal contribution (∗): Jinfu Wang and Qiyuan Wang. Corresponding author (†): Jinhua Zhao.

References

[Awadalla *et al.*, 2023] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. OpenFlamingo: An open-source framework for training large autoregressive vision-language models. https://arxiv.org/abs/2308.01390, 2023.

[Bai *et al.*, 2024] Yunhao Bai, Qinji Yu, Boxiang Yun, Dakai Jin, Yingda Xia, and Yan Wang. FS-MedSAM2: Exploring the potential of SAM 2 for few-shot medical image segmentation without fine-tuning. https://arxiv.org/abs/2409.04298, 2024.

[Bai *et al.*, 2025] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, et al. Qwen2.5-VL technical report. https://arxiv.org/abs/2502.13923, 2025.

[Baid *et al.*, 2021] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. https://arxiv.org/abs/2107.02314, 2021.

[Chen *et al.*, 2020] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1439–1449, Online, November 2020. Association for Computational Linguistics.

[Chen, 1984] Chi-Tsong Chen. *Linear system theory and design*. Holt, Rinehart and Winston, New York, USA, 1984.

[Cheng *et al.*, 2023] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, et al. SAM-Med2D. https://arxiv.org/abs/2308.16184, 2023.

[Gu and Dao, 2024] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. https://arxiv.org/abs/2312.00752, 2024.

[Gu *et al.*, 2022] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, pages 1–25, Online, April 2022.

[Hatamizadeh *et al.*, 2022a] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 272–284, Cham, Switzerland, September 2022. Springer International Publishing.

[Hatamizadeh *et al.*, 2022b] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. UNETR: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, Waikoloa, Hawaii, USA, January 2022. IEEE.

[Hu *et al.*, 2020] Xiaojun Hu, Weijian Luo, Jiliang Hu, Sheng Guo, Weilin Huang, Matthew R. Scott, Roland Wiest, Michael Dahlweid, and Mauricio Reyes. Brain SegNet: 3D local refinement network for brain lesion segmentation. *BMC Medical Imaging*, 20(1):17, Feb 2020.

[Huang *et al.*, 2023] Ziyan Huang, Haoyu Wang, Zhongying Deng, Jin Ye, et al. STU-Net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. https://arxiv.org/abs/2304.06716, 2023.

[Hurst *et al.*, 2024] Aaron Hurst, Adam Lerer, Adam P. Goucher, et al. GPT-4o system card. https://arxiv.org/abs/2410.21276, 2024.

[Jagadish, 1997] H. V. Jagadish. Analysis of the hilbert curve for representing two-dimensional space. *Information Processing Letters*, 62(1):17–22, April 1997.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, Paris, France, October 2023. IEEE.

[Liu *et al.*, 2021] Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, et al. Auto-encoding knowledge graph for unsupervised medical report generation. In *Advances in Neural Information Processing Systems*, pages 16266–16279, Online, December 2021. Curran Associates, Inc.

[Liu *et al.*, 2024] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual state space model. In *Advances in Neural Information Processing Systems*, pages 103031–103063. Curran Associates, Inc., December 2024.

[Lu and Du, 2025] Xiaoyong Lu and Songlin Du. JamMa: Ultra-lightweight local feature matching with joint mamba. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14934–14943, Nashville, Tennessee, USA, June 2025. IEEE.

[Ma *et al.*, 2024] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, January 2024.

[Moon *et al.*, 2001] Bongki Moon, H. V. Jagadish, Christos Faloutsos, and Joel H. Saltz. Analysis of the clustering properties of the Hilbert space-filling curve. *IEEE Transactions on Knowledge and Data Engineering*, 13(1):124–141, January–February 2001.

[OpenMedLab, 2024] OpenMedLab. XrayPULSE. https://github.com/openmedlab/XrayPULSE, 2024.

[Pei *et al.*, 2025] Xiaohuan Pei, Tao Huang, and Chang Xu. EfficientVMamba: Atrous selective scan for light weight visual mamba. In *Proceedings of the AAAI Conference*

on Artificial Intelligence, pages 6443–6451, Philadelphia, Pennsylvania, USA, February 2025. AAA Press.

[Petzsche *et al.*, 2022] Moritz R. Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, et al. ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific Data*, 9(1):762, December 2022.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[Ravi *et al.*, 2025] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *Proceedings of the International Conference on Learning Representations*, pages 28085–28128, April 2025.

[Schuch *et al.*, 2023] Fabiane Schuch, Lennart Walger, Matthias Schmitz, Bastian David, et al. An open presurgery MRI dataset of people with epilepsy and focal cortical dysplasia type II. *Scientific Data*, 10(1):475, July 2023.

[Shen *et al.*, 2025] Quanyou Shen, Bowen Zheng, Wenhao Li, Xiaoran Shi, et al. MixUNETR: A U-shaped network based on W-MSA and depth-wise convolution with channel and spatial interactions for zonal prostate segmentation in MRI. *Neural Networks*, 181:106782, January 2025.

[Thawakar *et al.*, 2024] Omkar Chakradhar Thawakar, Abdelrahman M. Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, et al. XrayGPT: Chest radiographs summarization using large medical vision-language models. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 440–448, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[Tu *et al.*, 2024] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, et al. Towards generalist biomedical AI. *NEJM AI*, 1:AIoa2300138, March 2024.

[Wang *et al.*, 2023] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2GenGPT: Radiology report generation with frozen LLMs. *Meta-Radiology*, 1:100033, September 2023.

[Wang *et al.*, 2025] Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, et al. SAM-Med3D: A vision foundation model for general-purpose segmentation on volumetric medical images. *IEEE Transactions on Neural Networks and Learning Systems*, 36:17599–17612, October 2025.

[Wu *et al.*, 2024] Hongtao Wu, Yijun Yang, Huihui Xu, Weiming Wang, Jinni Zhou, and Lei Zhu. RainMamba: Enhanced locality learning with state space models for video deraining. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7881–7890, Melbourne, VIC, Australia, October 2024. Association for Computing Machinery.

[Xing *et al.*, 2024] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. SegMamba: Long-range sequential modeling mamba for 3D medical image segmentation. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention*, pages 578–588, Marrakesh, Morocco, October 2024. Springer Nature Switzerland.

[Zhou *et al.*, 2023] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnFormer: Volumetric medical image segmentation via a 3D transformer. *IEEE Transactions on Image Processing*, 32:4036–4045, July 2023.

[Zhu *et al.*, 2024a] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, pages 18378–18394, Vienna, Austria, May 2024.

[Zhu *et al.*, 2024b] Jiayuan Zhu, Abdullah Hamdi, Yunli Qi, Yueming Jin, and Junde Wu. Medical SAM 2: Segment medical images as video via Segment Anything Model 2. https://arxiv.org/abs/2408.00874, 2024.