ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

RefSAM3D: Adapting the Segment Anything Model with cross-modal references for three-dimensional medical image segmentation

Xiang Gao and Kai Lu*

Department of Anesthesiology, Nanjing Drum Tower Hospital, Nanjing University, Nanjing, Jiangsu, China

Abstract

The Segment Anything Model (SAM), originally built on a two-dimensional vision transformer, excels at capturing global patterns in two-dimensional natural images but faces challenges when applied to three-dimensional (3D) medical imaging modalities such as computed tomography and magnetic resonance imaging. These modalities require capturing spatial information in volumetric space for tasks such as organ segmentation and tumor quantification. To address this challenge, we introduce RefSAM3D, an adaptation of SAM for 3D medical imaging by incorporating a 3D image adapter and cross-modal reference prompt generation. Our approach modifies the visual encoder to handle 3D inputs and enhances the mask decoder for direct 3D mask generation. We also integrate textual prompts to improve segmentation accuracy and consistency in complex anatomical scenarios. By employing a hierarchical attention mechanism, our model effectively captures and integrates information across different scales. Extensive evaluations on multiple medical imaging datasets demonstrate that RefSAM3D outperforms state-of-the-art methods. Our work thus advances the application of SAM in accurately segmenting complex anatomical structures in medical imaging.

Keywords: Three-dimensional medical imaging; Cross-modal reference prompt; Volumetric segmentation; Vision transformer

***Corresponding author:**
Kai Lu
(961340955@qq.com)

Citation: Gao X, Lu K. RefSAM3D: Adapting the Segment Anything Model with cross-modal references for three-dimensional medical image segmentation. *Artif Intell Health*. 2025;2(4):114-128. doi: 10.36922/AIH025080010

Received: February 17, 2025

Revised: May 1, 2025

Accepted: June 23, 2025

Published online: August 14, 2025

1. Introduction

Medical image segmentation is a fundamental task in medical imaging, primarily aimed at identifying and extracting specific anatomical structures, such as organs, lesions, and tissues, from medical images. This process is crucial for numerous clinical applications, including computer-aided diagnosis, treatment planning, and disease progression monitoring. Accurate image segmentation provides precise volumetric and shape information about target structures, which is essential for further clinical applications such as disease diagnosis, quantitative analysis, and surgical planning.[1-3]

Currently, recent breakthroughs in foundational models for image segmentation[4,5] have yielded transformative results, leveraging extensive datasets to capture general representations that exhibit exceptional generalizability and performance. However, despite these strides, significant challenges arise when applying these models, particularly

the Segment Anything Model (SAM), to medical image segmentation. For example, Huang *et al.*[6] demonstrated that SAM performs suboptimally on medical data, especially with objects that have irregular shapes or low contrast. Three main factors limit SAM's effectiveness in this domain. First, medical images, which often differ significantly from natural images, tend to be smaller, irregular in shape, and low in contrast, complicating direct application of the model. Second, medical structures typically have blurred or indistinct boundaries, whereas SAM's pre-training data includes predominantly well-defined edges, reducing segmentation accuracy and stability. Finally, medical imaging data often exists in three-dimensional (3D) forms with rich volumetric details. Yet, SAM's hint engineering was developed for two-dimensional (2D) data, limiting its ability to leverage 3D spatial features essential in medical contexts.

To enhance SAM's performance in medical imaging tasks, it is crucial to adapt and fine-tune the model to address domain-specific challenges. Recent studies have shown that parameter-efficient transfer learning (PETL) techniques, such as Low-Rank Adaptation[7] and Adapters,[8] are effective in this context. For instance, Med-Tuning[9] reduces the domain gap between natural images and medical volumes by incorporating Med-Adapter modules into pretrained visual foundation models. SAMed[10] employs the Low-Rank Adaptation fine-tuning strategy to adjust the image encoder, prompt encoder, and mask decoder of the SAM, achieving a balance between performance and deployment cost. However, these approaches predominantly focus on pure 2D adaptation, not fully exploiting the 3D information inherent in volumetric medical data. Nowadays, research is gradually shifting focus to better utilize the extensive data available in the 3D domain. The related methodologies can be categorized into two main approaches: one relies on prompt design based on SAM,[11-13] and the other achieves fully automatic segmentation when the segmented objects exhibit relatively regular shapes and positions.[14,15] The automatic prompt generation fails to leverage specialized medical knowledge and struggles to capture critical features due to blurred boundaries and small targets in medical images. These limitations result in suboptimal performance of automated methods, indicating further optimization.

In this paper, we propose Ref-SAM3D, an innovative approach that integrates textual prompts to enhance segmentation accuracy and consistency in complex anatomical scenarios. By incorporating text-based cues, our method enables SAM to perform referring expression segmentation within a 3D context, allowing the model to process both visual inputs and semantic descriptions for more intelligent segmentation strategies. We introduce a hierarchical attention mechanism that significantly improves the model's ability to capture and integrate information across different scales. This mechanism focuses on critical feature layers while filtering out irrelevant data, thereby enhancing segmentation precision and robustness, particularly in complex 3D structures. By integrating information across multiple scales, the model achieves a nuanced understanding of volumetric data, leading to more precise medical image segmentation. In addition, we adapt the visual encoder to handle 3D inputs and enhance the mask decoder for direct 3D mask generation, bridging the gap between SAM's 2D architecture and the demands of 3D medical imaging. This adaptation is crucial for ensuring the model's applicability and effectiveness in this domain. We evaluate our approach on multiple medical imaging datasets, demonstrating its superior performance compared to state-of-the-art methods. Our experiments highlight the effectiveness of our model in accurately segmenting complex anatomical structures, thereby advancing the application of SAM in medical imaging. The contributions of our work are as follows:

(i) We introduce a cross-modal reference prompt generation mechanism that integrates text and image embeddings into a unified feature space, facilitating effective cross-modal interaction.
(ii) We develop a hierarchical attention mechanism that significantly improves the model's ability to capture and integrate information across different scales, leading to improved segmentation precision and robustness, particularly in complex 3D structures.
(iii) We achieve state-of-the-art results across multiple benchmarks, demonstrating superior performance in 3D medical image segmentation tasks.

2. Related work

2.1. Vision foundation models (VFMs)

With the rapid development of foundation models in computer vision, recent research has focused on leveraging large-scale pre-training to create adaptable models with zero-shot and few-shot generalization capabilities.[16-19] These VFMs draw inspiration from language foundation models like generative pre-trained transformers (GPT) series, showing remarkable adaptability across domains and tasks using pre-training and fine-tuning paradigms.[20] Notable examples include the Contrastive Language-Image Pre-training (CLIP) model[21] and the A Large-scale ImaGe and Noisy-text embedding (ALIGN) model,[22] which employ image-text pairs to achieve zero-shot generalization across tasks such as classification and video understanding. Building on these foundations, segmentation-specific models such as the segment-everything-everywhere model[23] and SegGPT[24] have emerged to address more

complex tasks. The segment-everything-everywhere model enhances VFM capabilities by introducing a universal prompting scheme that enables semantic-aware open-set segmentation, expanding their use in real-world scenarios. SegGPT, in turn, standardizes segmentation data and employs in-context learning for both images and videos, allowing it to handle diverse segmentation tasks without requiring additional task-specific training. Complementing these advances, DINOv2[25] scales up Vision Transformer (ViT) pre-training by increasing data and model size, producing more general and transferable visual features that simplify fine-tuning across a wide range of tasks, further broadening VFM applicability. The SAM[4] is one of the most notable VFMs for general-purpose image segmentation. Pre-trained on 11 million images and 1 billion masks, SAM enables interactive, prompt-driven zero-shot segmentation across a wide variety of visual tasks. Its impressive versatility has made it a key model for applications such as image segmentation, inpainting, and tracking. However, it still faces limitations in specific domains such as medical imaging, camouflage detection, and shadow segmentation.[26]

2.2. Adaptation of the SAM in medical imaging

The adaptation of SAM for medical imaging has evolved rapidly, driven by its impressive zero-shot performance in natural image segmentation. Initial evaluation studies[27-30] examined SAM's applicability to medical image segmentation, but its performance often fell short due to the domain gap between natural and medical images. For instance, He *et al.*[28] noted a performance gap of up to 70% in Dice scores compared to domain-specific models. This highlighted the need for task-specific fine-tuning. Following this, research attention shifted from evaluation to the adaptation of SAM for medical images.[12,13,15,17] Several studies have experimented with fine-tuning SAM by modifying its prompt design to handle the specific characteristics of medical data. SAM-Med2D,[31] for example, leveraged more comprehensive prompts, including points, bounding boxes, and masks, to optimize SAM for 2D medical image segmentation, whereas the medical SAM adapter[12] incorporated point prompts and adapters to inject medical domain knowledge into SAM's architecture. Although these approaches enhanced SAM's performance, the creation of prompts for each 2D slice of 3D medical data proved to be labor-intensive. Efforts to adapt SAM for 3D medical image segmentation have focused on overcoming this limitation. MedLSAM[32] and SAM3D[33] applied SAM to 3D datasets, with approaches such as SAMed[10] and Med-Tuning[9] employing techniques such as Low-Rank Adaptation to fine-tune SAM for 3D tasks. However, most of these methods have not fully addressed the critical need to account for 3D volumetric or temporal information, which is vital for medical image segmentation. Innovations such as 3DSAM-Adapter[13] and modality-agnostic SAM (MA-SAM)[34] have incorporated 3D convolutional adapters to transform SAM's 2D architecture into one capable of recognizing 3D structures. Similarly, SAMMed3D[11] introduced a framework to generate 3D prompts from 2D points, helping SAM process volumetric data more effectively. The success of these 3D adaptations highlights the importance of leveraging spatial information for more accurate segmentation. Recent trends indicate a shift toward prompt-free or semiautomatic systems, like AutoSAM Adapter,[15] which aim to maintain SAM's zero-shot capabilities while minimizing manual prompt generation.

2.3. PETL

With the widespread adoption of foundational models, PETL has garnered significant attention. PETL methods can be categorized into three main groups. One approach is addition-based methods, which involve integrating lightweight adapters or prompts into the original model. These adapters or prompts allow the fine-tuning of only a small number of additional parameters, enabling the model to adapt to specific tasks while preserving the majority of its pre-trained weights. This approach minimizes the computational overhead associated with training large models, as only the newly introduced components require optimization.[9,35] Another strategy focuses on specification-based methods, which prioritize the identification and tuning of a small proportion of influential parameters from the original model. This method often employs techniques such as sensitivity analysis to determine which parameters have the most significant impact on the model's performance for a given task. By selectively updating these parameters, specification-based methods aim to achieve efficient adaptation while reducing training burden and maintaining high performance levels.[10,13] In addition, reparameterization-based methods leverage low-rank representations to minimize the number of trainable parameters during the fine-tuning process. Techniques such as Low-Rank Adaptation and factorized tuning allow models to maintain their expressive power while significantly reducing the number of parameters that need to be adjusted. This approach not only enhances efficiency but also enables strong performance across various PETL tasks, as it effectively captures the essential features required for adaptation.[7] Recently, PETL techniques have been successfully utilized to adapt VFMs for a wide range of downstream tasks, including image classification, object detection, and, notably, medical image segmentation. Researchers have explored ways to fine-tune vision models

efficiently while addressing the unique challenges posed by these complex tasks.[35-37]

2.4. Image segmentation by referring expressions

Referring image segmentation is a task that involves segmenting a specific object in an image based on a natural language description. This task requires the model to understand both the visual content of the image and the semantic meaning of the text, making it a challenging problem at the intersection of computer vision and natural language processing. With the advent of large-scale vision-language models, the performance of referring image segmentation has significantly improved. Models such as CLIP[38] and ALIGN[39] leverage large datasets of image-text pairs to learn joint embeddings that can be used for various vision-language tasks, including referring image segmentation. These models have demonstrated strong zero-shot and few-shot capabilities, enabling them to generalize well to unseen tasks and datasets. Recent advances have seen the adoption of transformer architectures for referring expression-based image segmentation. Transformer-based models, such as the ViT,[40] have been adapted to this task by integrating textual information into the visual processing pipeline. Ding *et al.*[41] introduced a vision-language transformer approach that leverages transformer and multi-head attention mechanisms to establish deep interactions between vision and language features, significantly enhancing holistic understanding. Similarly, cross-modal attention mechanisms have become a key component in modern referring image segmentation models. These mechanisms enable the model to effectively combine visual and textual features by computing attention scores between the two modalities. Li *et al.*[42] introduced the hierarchical dense attention module to fuse hierarchical visual semantic information with sparse embeddings to obtain fine-grained dense embeddings, and an implicit tracking module to generate a tracking token and provide historical information for the mask decoder.

3. Method

3.1. Overview of Ref-SAM3D

The original SAM, built on a 2D ViT, is proficient in capturing global patterns within 2D natural images. However, its applicability is limited when it comes to medical imaging modalities such as computed tomography (CT) and magnetic resonance imaging (MRI), which involve 3D volumetric data. In these contexts, 3D information is essential for applications such as organ segmentation and tumor quantification, as the characteristics of these structures must be captured from a 3D perspective. Relying solely on 2D views can result in reduced accuracy due to potential boundary blurring and non-standard scanning postures. Moreover, medical images differ significantly from natural images in both content and structure, demanding higher anatomical precision and detail. Directly applying segmentation models trained on natural images to medical domains thus yields limited effectiveness. Figure 1 shows the proposed method, RefSAM3D.

3.2. 3D volumetric input processing

To enhance SAM's performance in medical imaging tasks, the model needs to be adapted and fine-tuned to accommodate the domain-specific challenges. We introduced a 3D image adapter to enable SAM's processing of volumetric data.

We first modified the visual encoder to handle 3D volumetric inputs. Given a 3D medical volume $V \in R^{C\times D\times H\times W}$, where *C, D, H*, and *W* denote the channel, depth, height, and width, respectively, we extracted the 3D features through the following steps.

3.2.1. Patch embedding

We approximated a $k \times k \times k$ convolution (with $k = 14$) by employing a combination of $1 \times k \times k$ and $k \times 1 \times 1$ 3D convolutions. The $1 \times k \times k$ convolution was initialized with pre-trained 2D convolution weights, which remain frozen during fine-tuning. To manage the complexity of the model, we applied depth-wise convolutions for the newly introduced $k \times 1 \times 1$ convolutions, reducing the number of parameters that require tuning.

3.2.2. Positional encoding

In the pre-trained ViT model, we introduced an additional learnable lookup table with dimensions ($C\times D$) to encode the positional information for 3D points (*d, h,* and *w*). By summing the positional embedding from the frozen *(h, w)* table with the learnable depth-axis embedding, we provided accurate positional encoding for the 3D data.

3.2.3. Attention block

The attention block was directly adjusted to accommodate 3D features. For 2D inputs, the query size was (*B, HW, C*), which is easily modified to (*B, DHW, C*) for 3D inputs while retaining all pretrained weights. We adopted a sliding window mechanism, similar to that in the Swin Transformer, to mitigate memory overhead resulting from the increased dimensionality, optimizing the model's performance and memory footprint.

3.2.4. Bottleneck

As in other studies, we enhanced the bottleneck layer to better adapt to 3D tasks. Specifically, we replaced

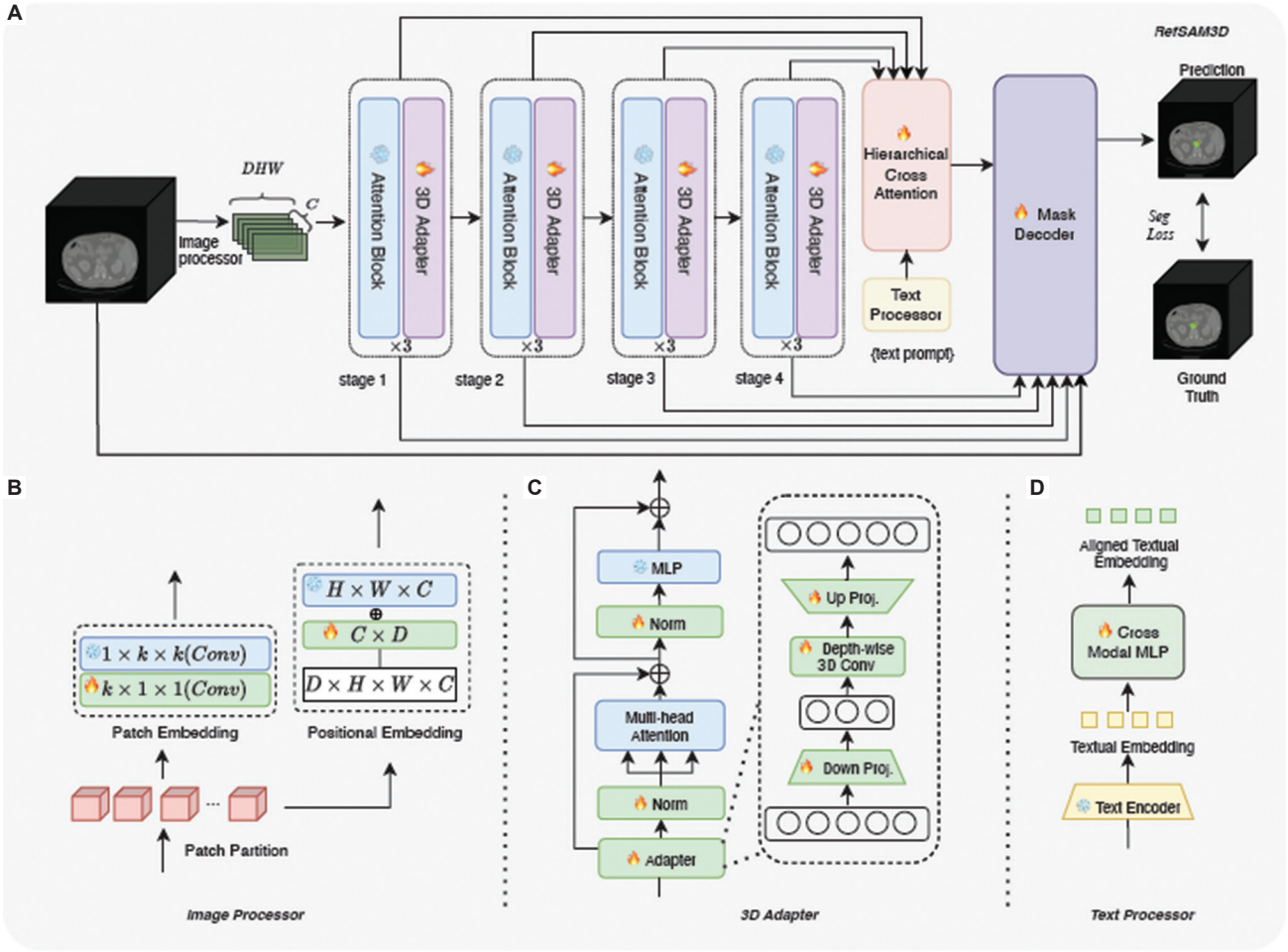

Figure 1. The proposed RefSAM3D method. (A) The overview of our proposed RefSAM3D for three-dimensional (3D) medical image segmentation, which integrates hierarchical cross-attention between image and text modalities to generate accurate segmentation predictions. (B) The design of the image processor, which includes patch partitioning, convolution-based patch embedding, and positional embedding, is used to process volumetric 3D medical data. (C) The framework of the 3D adapter incorporates multi-head attention, depth-wise 3D convolution, and up/down projection for efficient feature extraction and adaptation. (D) The pipeline of the text processor encodes textual prompts and aligns them with visual embeddings using a cross-modal multilayer perceptron for enhanced segmentation guidance.

2D convolutions with 3D ones and trained these layers from scratch to improve performance. To avoid the computational expense of fully fine-tuning a 3D ViT, we employed a lightweight adapter for efficient fine-tuning. The adapter comprised a down-projection and an up-projection linear layer, formulated as shown in Equation I:

$$\text{Adapter}\left(X\right) = \text{X} + \text{Act}(XW_{Down})W_{Up} \quad \text{(I)}$$

where $X \in \mathbb{R}^{N \times C}$ represents the input feature, $W_{Down} \in \mathbb{R}^{C \times N'}$ and $W_{Up}\mathbb{R}^{N' \times C}$ are the down-projection and up-projection layers, and Act (·) is the activation function. In addition, we incorporated depth-wise convolutions after the down-projection layer to enhance 3D spatial awareness.

3.3. Cross-modal reference prompt generation

3.3.1. Text encoder

Within the SAM framework, we carefully designed a text encoder to process textual prompts related to image segmentation tasks. Specifically, we employed the text encoder from the CLIP model, which can convert input textual prompts, such as "perform liver segmentation," into corresponding text embedding vectors.

The textual prompt was first tokenized into a sequence of tokens $\text{T} = t_{ll}^{L} = 1$. These tokens were then input into the CLIP text encoder to obtain the final embedding representation. The output of the text encoder is expressed as the formula shown in Equation II:

$$\mathcal{F}_e = \varepsilon_t(\text{T}) \in \mathbb{R}^{L \times C_e} \quad \text{(II)}$$

Here, $\mathcal{F}_e$ is the sequence of L word embeddings, each with C_e dimensions, i.e., $\mathcal{F}_e = f_{ii=1}^{L}$, where each word is represented by a C_e -dimensional embedding. By applying a pooling operation over these word embeddings, we obtained a sentence-level embedding $\mathcal{F}_e^s \in \mathbb{R}^{C_e}$.

3.3.2. Cross-modal projector

While text embeddings derived from pre-trained language models capture rich semantic representations, a significant gap exists between these representations and those obtained from visual encoders. This semantic disparity poses challenges in cross-modal fusion, as the two modalities do not naturally reside in the same embedding space. To address this, we adopted a strategy inspired by vision-and-language bidirectional encoder representations from transformers, wherein we employed a multilayer perceptron to align the text and image embeddings. This allows both modalities to be projected into a unified feature space, enabling more effective interaction. Specifically, for each word embedding f_i in $\mathcal{F}_e$, the sparse embedding can be obtained by adopting the cross-modal multilayer perceptron (Equation III):

$$\mathcal{F}_i^s = MLP(f_i) \in \mathbb{R}_{C_v} \quad \text{(III)}$$

3.3.3. Image feature extraction

As previously mentioned, we integrated lightweight adapters into our 3D SAM to efficiently adapt the model for processing volumetric medical images. In this step, we extracted the features produced by each attention block as cross-attention visual hierarchical features.

Let $\mathrm{V}_i \in \mathbb{R}^{B\times D_i H_i W_i \times C}$ denote the output of the i^{th} attention block, where B is the batch size, and H_i, W_i and D_i represent the height, width, and depth of the feature maps, respectively. This extraction allowed us to leverage the unique focus of each attention block on different aspects of the input data, capturing a rich representation of 3D spatial patterns. The adapted features are computed as Equation IV:

$$V_i^{'} = \mathrm{Adapter}_i(\mathrm{V}_i), \quad \forall i \in \{1,2,\ldots..,N\} \quad \text{(IV)}$$

where N = 4. We can obtain a collection of image features, as depicted in Equation V:

$$V^{'} = V_1^{'}, V_2^{'}, \ldots., V_N^{'} \quad \text{(V)}$$

3.3.4. Hierarchical cross-attention

The hierarchical cross-attention architecture is designed to integrate multi-level visual features with textual inputs, enabling a deeper understanding of cross-modal data in 3D tasks such as medical image analysis. By extracting hierarchical features from each attention block in a 3D SAM, the architecture leverages the fact that each layer focuses on different aspects of the input data, from low-level details to high-level semantics. This structure enhances the model's ability to relate complex 3D spatial patterns with corresponding textual prompts, improving cross-modal understanding. Figure 2 shows the hierarchical cross-attention architecture.

In this architecture, the inputs include both the hierarchical image features, $V^{'} = V_1^{'}, V_2^{'}, \ldots., V_N^{'}$, derived from each attention block, and a textual prompt T, which encodes the semantic information. These inputs are fused through a cross-attention mechanism where each layer of visual features interacts with the textual input, allowing mutual enrichment of modalities. The output is a cross-modal prompt that combines visual and textual information, which can be fed into SAM's prompt encoder to guide tasks such as segmentation or object detection in 3D medical images.

In the hierarchical cross-attention architecture, the cross-attention mechanism is designed to facilitate interaction between the hierarchical image features and

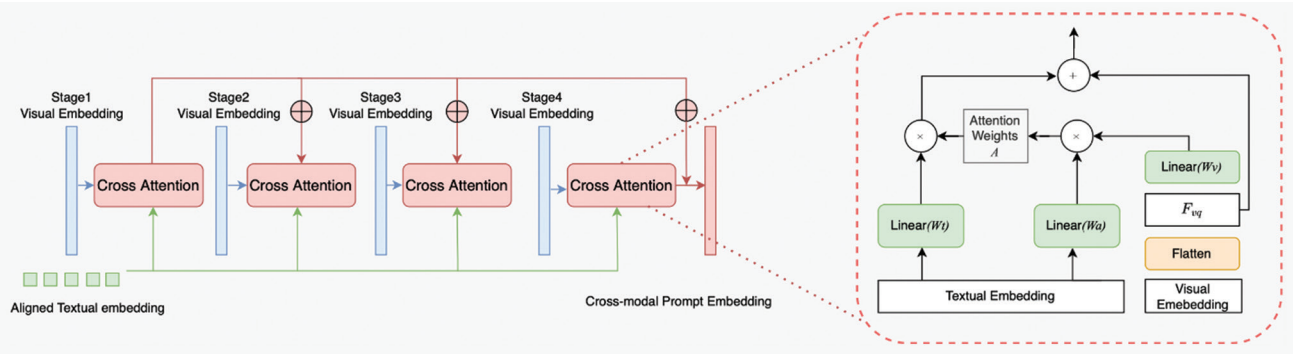

Figure 2. The structure of the cross-modal prompt embedding module. The left part illustrates the overall architecture, where hierarchical visual embeddings from four stages interact with aligned textual embeddings using cross-attention mechanisms to generate cross-modal prompt embeddings. The right part details the cross-attention mechanism, showing how attention weights are computed to align textual and visual embeddings through linear transformations and fusion, enabling effective multi-modal integration for downstream tasks.

the textual prompt. As mentioned above, $V_i^{'}$ represents the adapted feature maps extracted from the ith attention block, and the textual prompt is $T \in \mathbb{R}^{B\times L\times C}$.

The cross-attention process can be formally expressed as follows (Equation VI). For each hierarchical feature $F_i^{'}$, we computed the attention scores A_i with respect to the text T:

$$A_i = \text{softmax}\left(\frac{Q_i K^T}{\sqrt{d_k}}\right) \quad \text{(VI)}$$

where $\mathrm{Q}_i \in \mathbb{R}^{B\times D_i H_i W_i \times C}$ are the queries derived from $F_i^{'}$, and $K \in \mathbb{R}^{B\times L\times C}$ are the keys derived from the textual prompt T. The dimensionality d_k represents the size of the keys, which is a scaling factor to ensure stable gradients during training. The attention output O_i for each feature block can then be computed as in Equation VII:

$$O_i = A_i V_i \quad \text{(VII)}$$

Where V_i denotes the values corresponding to $F_i^{'}$ and is similarly dimensioned as $F_i^{'}$. The final output from the cross-attention mechanism can be represented as Equation VIII:

$$\mathrm{O} = [\mathrm{O}_1, \mathrm{O}_2, \ldots, \mathrm{O}_N] \in \mathbb{R}^{B\times DHW\times C} \quad \text{(VIII)}$$

resulting in a combined output that integrates both visual and textual information across multiple layers. This enriched representation was then utilized as a cross-modal prompt in the subsequent stages of SAM's prompt encoder, effectively bridging the gap between visual features and semantic understanding derived from text.

3.4. Lightweight mask decoder

The original SAM mask decoder comprises merely two transformer layers, two transposed convolution layers, and a multilayer perception layer. In the context of 3D medical image processing tasks, we replaced the 2D convolutions with 3D convolutions to enable direct 3D mask generation. Given that many anatomical structures or lesions in medical images are relatively small, it is often necessary to achieve higher resolution images to ensure better distinction of the segmented elements.

In the image encoder of the SAM, the patch embedding process of the transformer backbone embeds each 16 × 16 patch into a feature vector, resulting in a 16 × 16 down-sampling of the input. The SAM mask decoder employs two consecutive transposed convolution layers to up-sample the feature map by a factor of four. However, the final prediction generated by SAM still has a resolution that is four times lower than the original input shape. To address this problem, we employed progressive up-sampling, making moderate adjustments to the SAM decoder by integrating two additional transposed convolution operations. With each layer up-sampling the feature maps by a factor of two, the four transposed convolutional layers progressively restored feature maps to their original input resolution. In addition, we introduced a multilayer aggregation mechanism, designing a network akin to a "U-shaped" architecture. We combined intermediate feature maps from stages 1–4 during the image encoder phase with prompts generated during the cross-modal reference prompt generation phase to enrich the mask features. After up-sampling the mask feature map to the original resolution, we concatenated it with the original image and used another 3D convolution to fuse the information and generate the final mask to better leverage information from the original resolution.

4. Experiments

4.1. Experimental setup

We conducted a comprehensive evaluation of our segmentation method across four medical image segmentation tasks, encompassing three distinct imaging modalities: CT-based tumor segmentation, MRI-based cardiac segmentation, and multi-organ segmentation from multi-modal datasets. Our approach was rigorously compared against state-of-the-art methods on CT imaging tasks. In addition, we assessed our method's performance on MRI cardiac and multi-organ segmentation tasks, providing a thorough analysis of its generalization capabilities and conducting an in-depth ablation study to elucidate the contributions of its constituent components.

4.1.1. Datasets

The kidney tumor segmentation (KiTS21) dataset[43] is a comprehensive collection designed for the segmentation of kidneys, tumors, and cysts in CT imaging. It comprises 300 publicly available training cases and 100 withheld testing cases. The dataset is formatted in 3D CT with files stored in the.nii.gz format. The image dimensions exhibit significant variability, with voxel spacing ranging from (0.5, 0.44, 0.44) mm to (5.0, 1.04, 1.04) mm and sizes ranging from (29, 512, 512) to (1,059, 512, 796). The dataset includes annotations for three anatomical structures: kidneys, tumors, and cysts. These structures are consistently present across all training cases, with cysts appearing in 49.33% of the cases. This dataset serves as a critical resource for advancing automated segmentation techniques in medical imaging analysis.

The Medical Segmentation Decathlon (MSD) pancreas tumor dataset[12] consists of 281 contrast-enhanced

abdominal CT scans with annotations for both the pancreas and pancreatic tumors. This dataset is part of the MSD pancreas segmentation challenge. Each CT volume has a resolution of 512 × 512 pixels, with the number of slices per scan ranging from 37 to 751. The authors filtered the dataset to retain only the axial view images containing more than 5% pancreatic content. Consistent with previous studies, we merged the pancreas and pancreatic tumor masks into a single entity for segmentation.

The liver tumor segmentation benchmark (LiTS)[44] dataset is a publicly available benchmark dataset focused on liver and liver tumor segmentation. It was created to evaluate and compare the performance of automated liver and liver tumor segmentation algorithms. The LiTS dataset comprises 201 abdominal CT scans, of which 194 contain liver lesions. The dataset is divided into 131 training cases and 70 testing cases. The resolution and quality of the CT images vary, with axial resolutions ranging from 0.56 mm to 1.0 mm and z-direction resolutions ranging from 0.45 mm to 6.0 mm.

The MSD colon dataset[45] is a publicly available benchmark dataset focused on primary colon cancer segmentation from CT images. The dataset consists of 190 abdominal CT scans in total, which are divided into 126 training cases and 64 testing cases. Each case is annotated with segmentation masks identifying the primary colon cancer regions.

For cardiac segmentation, we utilized the multi-modality whole heart segmentation (MM-WHS) Challenge 2017 dataset,[46] which contains 20 CT and 20 MRI scans with pixel-level ground-truth annotations. These scans were collected in a real clinical setting and include five anatomical labels: left ventricle blood cavity, right ventricle blood cavity, left atrium blood cavity, right atrium blood cavity, and ascending aorta. In our experiments, only the CT scans were used, which contain between 177 and 363 slices, each with a resolution of 512 × 512 pixels and voxel spacing ranging from 0.3 to 0.6 mm.

The Beyond the Cranial Vault (BTCV) challenge dataset[47] comprises 30 CT volumes, each manually labeled with 13 different abdominal organs. The number of slices per scan ranges between 85 and 198, with a slice thickness varying between 2.5 mm and 5.0 mm. All scans have an axial resolution of 512 × 512, whereas the in-plane resolution varies from 0.54×0.54 mm^2 to 0.98×0.98 mm^2. We followed the data split proposed by Tang *et al.*,[48] utilizing 24 cases for training and 6 cases for testing.

For evaluating the model's generalization ability, we also used the multi-modality abdominal multi-organ segmentation challenge (AMOS22) dataset.[49] This dataset includes abdominal CT and MRI scans from different patients, with each scan annotated for 15 organs. In line with the approach in MA-SAM, we limited our evaluation to the 12 organs common to both the AMOS22 and BTCV datasets. For generalization testing, we utilize 300 CT scans and 60 MRI scans from the AMOS22 training and validation sets.

4.1.2. Implementation details

We implemented our method and benchmarked it against baseline models using PyTorch (version 2.7.1) and the medical open network for AI framework, specifically utilizing SAM-B for all experiments, which employs ViT-B as the image encoder backbone. The training was conducted on an NVIDIA A40 GPU (United States) with a batch size of 1, using the AdamW optimizer with a linear learning rate scheduler for a total of 200 epochs. The initial learning rate was set to 1e-4, with a momentum of 0.9 and a weight decay of 1e-5. Data preprocessing involved adjusting the isotropic spacing to 1 mm. For data augmentation, we applied various transformations, including random rotation, flipping, erasing, shearing, scaling, translation, posterization, contrast adjustments, brightness modifications, and sharpness enhancements. During training, we also sampled foreground and background patches at a 1:1 ratio. For single-organ cancer segmentation, we assessed our method's performance through comparisons with state-of-the-art volumetric segmentation and fine-tuning techniques, using the dice coefficient and normalized surface dice (NSD) as evaluation metrics, similar to Sam-med3d [11]. For multi-organ segmentation, we employed the dice coefficient and Hausdorff distance (HD) as evaluation metrics. For each dataset, we designed specific text prompts to guide the segmentation process, as shown in Table 1. These prompts were carefully crafted to provide clear anatomical context while maintaining consistency across different organs and pathologies.

4.2. Comparison with state-of-the-art methods

Our method was extensively evaluated against a wide range of state-of-the-art 3D medical image segmentation techniques on both CT and MRI datasets. These techniques include the convoluted neural network-based no new (nn)U-Net[50]—an automated configuration framework evolved from the U-Net architecture[51]—and the Swin U-Net transformers (Swin-UNETR),[52] which employs a hierarchical encoder structure for 3D segmentation tasks. Furthermore, we also considered nnFormer,[53] a model that integrates both local and global volumetric self-attention mechanisms, and UNETR++,[54] which enhances segmentation accuracy and efficiency through

the introduction of an efficient pairing attention module. In addition, we compared our approach with 3D UNet-eXpanded Network (UX-Net),[55] a method designed to create a simple, efficient, and lightweight network that combines the capabilities of hierarchical transformers with the advantages of ConvNet modules. We also evaluated SAM-B, which is the base model of SAM trained on natural images and directly applied to medical images without adaptation. Finally, our method was benchmarked against the latest SAM adaptation techniques, including 3DSAM-adapter[13]—a promptable 3D medical image segmentation model—and MA-SAM,[34] a framework that utilizes parameter-efficient fine-tuning strategies and 3D adapters.

The results presented in Table 2 demonstrate that our proposed Ref-SAM3D method consistently outperformed other approaches across a wide range of tasks, achieving the highest scores in nearly all scenarios, particularly excelling in challenging tumor types. In kidney tumor segmentation, despite challenges such as low contrast with surrounding tissues, blurred boundaries, and high morphological heterogeneity, Ref-SAM3D achieved a dice score of 95.53% and an NSD of 99.45%, surpassing other methods. For pancreatic tumors, which constitute less than 0.5% of CT images and exhibit diverse shapes, Ref-SAM3D achieved a dice score of 82.42%, representing a 2.12% improvement over existing state-of-the-art techniques. In liver tumor segmentation, Ref-SAM3D attained a dice score of 80.10%, effectively handling variations in grayscale and irregular shapes. Despite the extensive distribution and complex anatomical structure of colorectal cancer lesions,

Table 1. Datasets used in our experiments and their corresponding prompt content descriptions

Task	Dataset name	Prompt content
Kidney tumor segmentation	KiTS21 Challenge	CT images, kidneys, tumors, and cysts segmentation, spacing (0.5, 0.44, 0.44) mm to (5.0, 1.04, 1.04) mm, dimensions (29, 512, 512) to (1,059, 512, 796)
Pancreas tumor segmentation	MSD pancreas	CT images, pancreas tumor segmentation, resolution 512×512, slices 37–751
Liver tumor segmentation	LiTS dataset	CT images, liver tumor segmentation, axial resolution 0.56–1.0 mm, z-direction resolution 0.45–6.0 mm
Colon cancer segmentation	MSD colon dataset	CT images, colon cancer segmentation, and abdominal scans
MRI cardiac segmentation	MM-WHS Challenge	MRI images, cardiac structure segmentation (LVC, RVC, LAC, RAC, AA), resolution 512×512, voxel spacing 0.3–0.6 mm
Abdominal multi-organ segmentation	BTCV Challenge	CT images, abdominal organ segmentation (13 organs), slice thickness 2.5–5.0 mm, in-plane resolution 0.54×0.54 mm^2 to 0.98×0.98 mm^2
Multi-modality abdominal multi-organ segmentation	AMOS22 dataset	CT and MRI images, abdominal organ segmentation (15 organs), varying modalities and resolutions

Abbreviations: AA: Ascending aorta; AMOS: Abdominal Multi-Organ Segmentation; BTCV: Beyond the Cranial Vault; CT: Computed tomography; MM-WHS: Multi-Modality Whole Heart Segmentation; MRI: Magnetic resonance imaging; MSD: Medical Segmentation Decathlon; LAC: Left atrium blood cavity; LiTS: Liver Tumor Segmentation Benchmark; LVC: Left ventricle blood cavity; RAC: Right atrium blood cavity; RVC: Right ventricle blood cavity.

Table 2. Comparison with classical medical image segmentation methods on four tumor segmentation datasets

Methods	Kidney tumor		Pancreas tumor		Liver tumor		Colon cancer	
	Dice	NSD	Dice	NSD	Dice	NSD	Dice	NSD
nnU-Net	73.07	77.47	41.65	62.54	60.10	75.41	43.91	52.52
Swin-UNETR	65.54	72.04	40.57	60.05	50.26	64.32	35.21	42.94
UNETR++	56.49	60.04	37.25	53.59	37.13	51.99	25.36	30.68
nnFormer	45.14	42.28	36.53	53.97	45.54	60.67	24.28	32.19
3D UX-Net	57.59	58.55	34.83	52.56	45.54	60.67	28.50	32.73
SAM-B (10 pts/slice)	40.07	34.96	30.55	32.91	8.56	5.97	39.14	42.70
3DSAM-adapter (10 points/volume)	74.91	84.35	57.47	79.62	56.61	69.52	49.99	65.67
MA-SAM (1 relaxed 3D bounding box/slice)	93.38	98.91	80.30	97.19	75.23	92.31	65.45	81.40
Ref-SAM3D	95.53	99.45	82.42	98.41	80.10	93.23	70.14	88.90

Note: All data presented as percentages (%).
Abbreviations: 3D: Three-dimensional; nn: No new; NSD: Normalized surface Dice; SAM: Segment Anything Model; UNETR: U-Net Transformers; UX-Net: UNet-eXpanded Network.

Ref-SAM3D achieved a dice score of 70.14%, marking a 10.11% increase over current technologies. It is noteworthy that traditional methods like nnU-Net perform well on certain tasks, yet overall, they fall short compared to newer methods such as Ref-SAM3D. Particularly, when dealing with tumors that have blurred boundaries and diverse morphologies, Ref-SAM3D demonstrated significant advantages. These findings underscore the exceptional performance of Ref-SAM3D in addressing a variety of complex medical image segmentation challenges. Figure 3 shows the qualitative visualizations of these tasks.

In the domain of multi-organ segmentation, we conducted experiments on the BTCV dataset. The Ref-SAM3D approach demonstrated exceptional capability, achieving a dice score of 97.1% for spleen segmentation, as shown in Table 3, which surpasses all comparative methods. The left and right kidneys attained dice scores of 96.1% and 94.9%, respectively. The esophagus achieved a dice score of 85.2%, surpassing other methods, whereas the liver and stomach achieved scores of 97.3% and 94.1%, respectively. Furthermore, Ref-SAM3D demonstrated efficiency in handling complex anatomical structures, such as the pancreas and aorta, achieving dice scores of 87.5% and 92.3%, respectively. Ref-SAM3D achieved an average HD value of 2.34, underscoring its superior boundary precision. Figure 4 shows qualitative visualizations of BTCV tasks. From the qualitative visualization results, Ref-SAM3D demonstrated superior performance in multi-organ segmentation tasks. The method accurately identified and segmented boundaries between different organs, maintaining high segmentation precision even in cases with blurred organ boundaries or complex anatomical structures. Notably, Ref-SAM3D maintained stable segmentation performance for both small organs such as the pancreas, and elongated structures, such as the aorta, further validating the reliability of the quantitative evaluation metrics.

In addition, in the context of cardiac tumor segmentation using MRI, as shown in Figure 5, a qualitative assessment of predicted masks from various segmentation models indicates that our AutoSAM Adapter produced visually superior results, especially in terms of boundary precision, when compared to existing state-of-the-art methods.

4.3. Generalization evaluation

To assess the generalization capabilities of Ref-SAM3D, we conducted comprehensive experiments across heterogeneous datasets and imaging modalities. Our evaluation framework encompassed two distinct scenarios: cross-modality generalization on the AMOS22 dataset (comprising both CT and MRI modalities) and cross-dataset adaptation using the MM-WHS cardiac imaging dataset.

In the zero-shot generalization experiments, we evaluated the model's transferability by applying our Ref-SAM3D, trained exclusively on the BTCV CT dataset, to the AMOS22 dataset without any additional fine-tuning. The quantitative results demonstrated remarkable

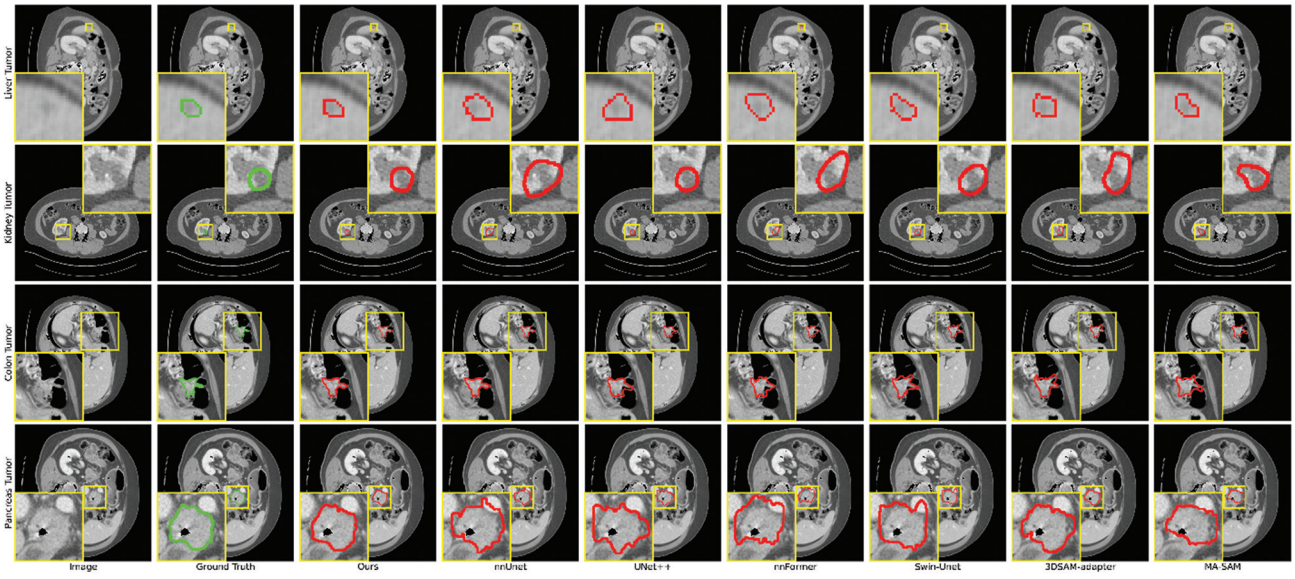

Figure 3. Qualitative visualizations of the proposed method and baseline approaches on liver tumor, kidney tumor, pancreas tumor, and colon cancer segmentation tasks

Abbreviations: 3D: Three-dimensional; nn: No new; NSD: Normalized surface Dice; SAM: Segment Anything Model; UNETR: U-Net Transformers; UX-Net: UNet-eXpanded Network

Table 3. Comparison of abdominal multi-organ segmentation results

Metric	Method	Spleen	R.Kd	L.Kd	GB	Eso.	Liver	Stomach	Aorta	IVC	Veins	Pancreas	AG	Average
Dice (%)	nnU-Net	97.0	95.3	95.3	63.5	77.5	**97.4**	89.1	90.1	88.5	79.0	87.1	**75.2**	86.3
	Swin-UNETR	95.6	94.2	94.3	63.6	75.5	96.6	79.2	89.9	83.7	75.0	82.2	67.3	83.1
	UNETR++	94.2	92.1	95.4	65.0	75.9	96.9	88.3	85.5	84.9	76.1	81.8	71.3	83.95
	nnFormer	93.5	94.9	95.0	64.1	79.5	96.8	90.1	89.7	85.9	77.8	85.6	73.9	85.6
	3D UX-Net	94.6	94.2	94.3	59.3	72.2	96.4	73.4	87.2	84.9	72.2	80.9	67.1	81.4
	3DSAM-adapter	94.3	96.1	94.1	62.9	79.9	96.1	83.8	88.4	85.3	75.6	83.1	69.4	84.1
	MA-SAM	96.7	95.1	95.4	68.2	82.1	96.9	92.8	91.1	87.5	79.8	86.6	73.9	87.2
	Ref-SAM3D	**97.1**	**94.9**	**96.1**	**70.3**	**85.2**	97.3	**94.1**	**92.3**	**88.8**	**80.4**	**87.5**	75.1	**88.3**
HD (%)	nnU-Net	1.07	1.19	1.19	7.49	8.56	**1.14**	4.84	14.11	2.87	5.67	**2.31**	**2.23**	4.39
	Swin-UNETR	1.21	1.41	1.37	2.25	5.82	1.70	13.75	5.92	4.46	7.58	3.53	3.40	4.37
	UNETR++	5.99	1.23	1.33	5.99	10.37	33.12	5.23	8.23	2.14	10.34	3.12	2.13	7.44
	nnFormer	78.03	1.41	1.43	3.00	4.92	1.38	4.24	7.53	4.02	6.53	2.96	2.76	9.95
	3D UX-Net	3.17	1.59	1.26	4.53	13.92	1.75	19.72	12.53	3.47	9.99	3.70	4.11	6.68
	3DSAM-adapter	3.38	1.23	1.21	2.23	5.43	1.15	4.00	6.47	7.88	5.18	4.71	3.94	3.90
	MA-SAM	**1.00**	**1.19**	1.07	1.59	3.77	1.36	3.87	5.29	3.12	3.25	3.93	2.57	2.67
	Ref-SAM3D	1.30	1.32	**1.00**	**1.21**	**3.18**	1.23	**3.77**	**4.12**	**2.30**	**3.12**	3.08	2.44	**2.34**

Abbreviations: 3D: Three-dimensional; AG: Average; Eso.: Esophagus; GB: Gall bladder; HD: Hausdorff distance; IVC: Inferior vena cava; L.Kd: Left kidney; nn: No new; R.kd: Right kidney; SAM: Segment Anything Model; UNETR: U-Net Transformers; UX-Net: UNet-eXpanded Network.

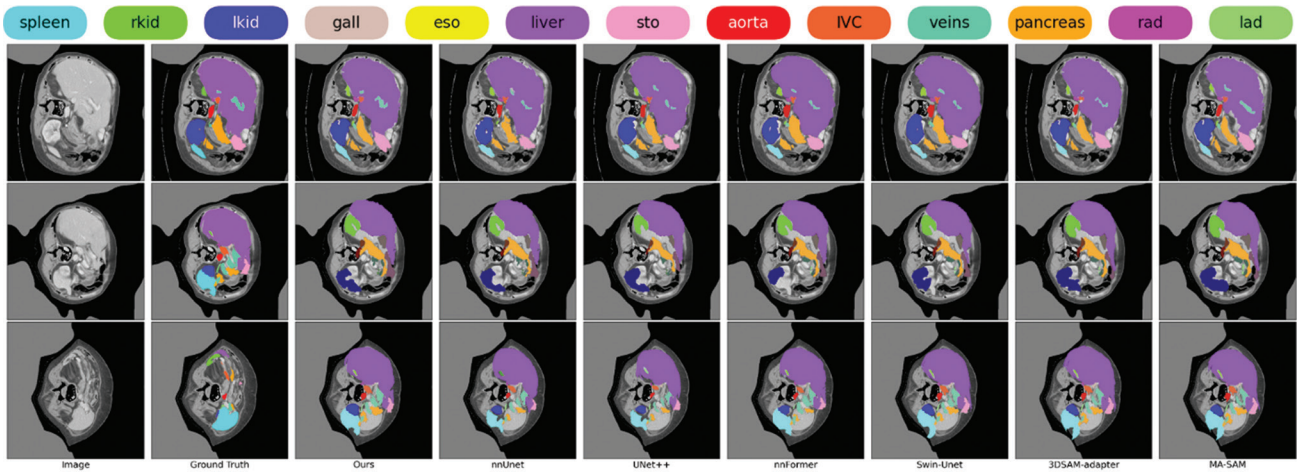

Figure 4. Qualitative visualization of segmentation results generated from our Ref-SAM3D method and other state-of-the-art methods on the Beyond the Cranial Vault dataset. Rkid and Lkid refer to the right and left kidneys, respectively. Sto, rad, and lad stand for stomach, respectively.
Abbreviations: 3D: Three-dimensional; IVC: Inferior vena cava; nn: No new; NSD: Normalized surface Dice; SAM: Segment Anything Model; UNETR: U-Net Transformers; UX-Net: UNet-eXpanded Network.

performance, achieving a mean dice coefficient of 85.7% on CT images, indicating robust generalization across different CT acquisition protocols and patient cohorts. Notably, in the challenging cross-modality scenario of MRI segmentation, our model maintained substantial performance with a dice score of 63.2% (±3.1%), significantly surpassing baseline methods, including nnU-Net (12.1%) and Swin-UNETR (15.3%).

Furthermore, when employing a five-shot fine-tuning strategy on the AMOS22 MRI data, Ref-SAM3D exhibited even more impressive results, achieving a dice score of 84.1% (Figure 6). This represents a substantial improvement over the fine-tuned versions of nnU-Net (72.4%) and Swin-UNETR (75.3%), demonstrating the model's superior adaptability and learning efficiency with minimal additional training data. These results underscore

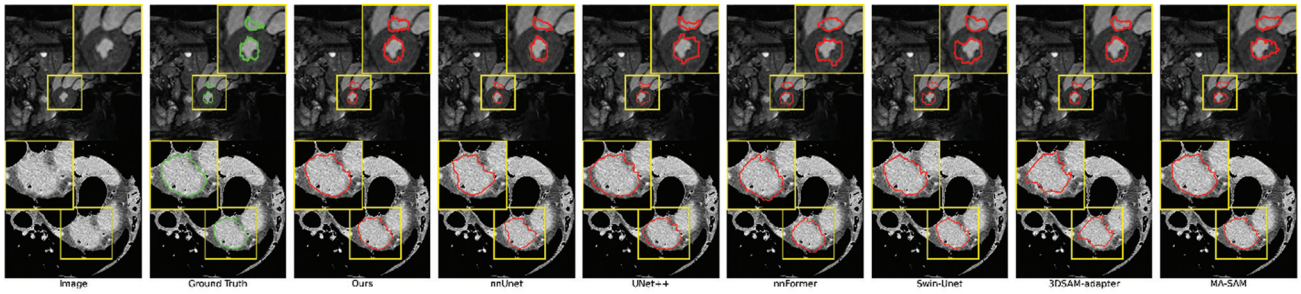

Figure 5. Qualitative visualization of segmentation results generated from different methods for magnetic resonance imaging cardiac tumor segmentation
Abbreviations: 3D: Three-dimensional; nn: No new; NSD: Normalized surface Dice; SAM: Segment Anything Model; UNETR: U-Net Transformers; UX-Net: UNet-eXpanded Network.

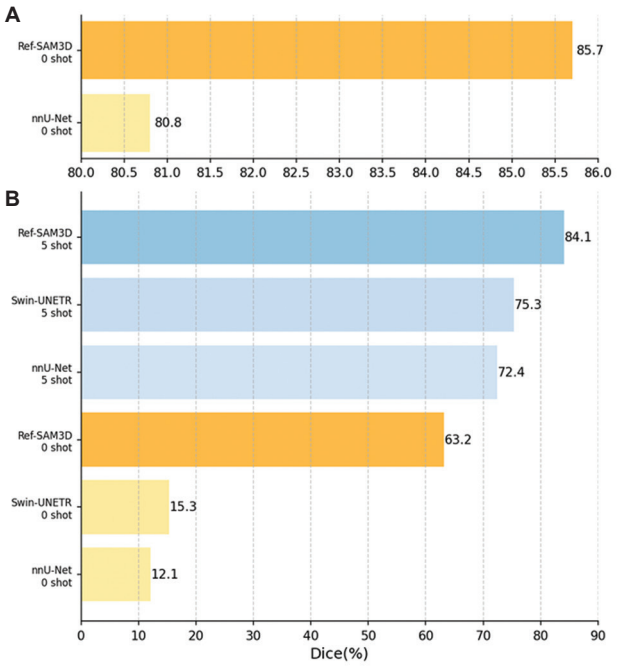

Figure 6. Comparison of zero-shot and five-shot generalization performance of Ref-SAM3D, nnU-Net, and Swin-UNETR on AMOS22 data. (A) Computed tomography (CT) and (B) magnetic resonance imaging (MRI) data.
Abbreviations: 3D: Three-dimensional; nn: No new; NSD: Normalized surface Dice; SAM: Segment Anything Model; UNETR: U-Net Transformers; UX-Net: UNet-eXpanded Network.

Ref-SAM3D's robust generalization capabilities and its potential as a versatile solution for medical image segmentation across different imaging modalities.

These experimental findings clearly demonstrate Ref-SAM3D's robust performance across different datasets and imaging modalities. The model's strong zero-shot generalization capabilities and impressive few-shot learning results suggest its practical value in real-world medical applications, where adapting to diverse imaging conditions with minimal additional training is essential. These characteristics position Ref-SAM3D as a promising solution for clinical deployment, particularly in scenarios requiring flexible and efficient medical image analysis tools.

Table 4. Ablation on each key component in our method

Parameters	Dice (%)	Hausdorff distance (%)
Ref-SAM3D	88.3	2.34
Without a text prompt	72.3	7.31
Without a cross-modal projector	80.1	4.22
Without hierarchical fusion	74.1	6.33

4.4. Ablation study

4.4.1. Effects of text prompt

The text prompt in our Ref-SAM3D model provided essential semantic guidance by bridging textual descriptions and visual features, enabling better interpretation of anatomical structures. The results, as shown in Table 4, without this component, the model's performance dropped significantly, with dice score decreasing from 88.3% to 72.3% (−16.0%) and HD increasing from 2.34% to 7.31% (+4.97%). This substantial degradation demonstrates that the text prompt is crucial for leveraging linguistic context to achieve precise medical image segmentation.

4.4.2. Effects of cross-modal projector

The cross-modal projector in Ref-SAM3D plays a vital role in aligning textual and visual inputs, facilitating effective integration of multi-modal information for improved segmentation. By harmonizing these inputs, the projector enhanced the model's ability to utilize semantic context from text alongside visual data. As shown in Table 4, removing this component resulted in an 8.2% decrease in dice score (from 88.3% to 80.1%) and an HD increase from 2.34% to 4.22%. These results confirm that when the cross-modal projector is removed, the model relies on unaligned embeddings, which can lead to less effective feature integration.

Table 5. The ablation experiments of each stage under the hierarchical cross-attention

Stages	Dice (%)	Hausdorff distance (%)
All stages	88.3	2.34
Stages 1 and 4	78.5	2.76
Stages 2 and 4	82.1	2.62
Stages 3 and 4	85.4	2.48
Stage 4 only	73.78	2.89

4.4.3. Effects of hierarchical cross-attention mechanism

The hierarchical fusion mechanism in Ref-SAM3D is pivotal for integrating information across various encoder layers, enabling the model to capture detailed, multi-level semantic features essential for precise segmentation. Ablation studies, summarized in Table 4, demonstrate the significance of this mechanism. Removing the hierarchical fusion led to a sharp decline in segmentation accuracy, with the dice coefficient dropping from 88.3% to 74.1%, and the HD increasing from 2.34% to 6.33%. This underscores the mechanism's role in effectively combining features across layers for better performance.

Moreover, Table 5 provides a systematic evaluation of each block level's contribution to the model. The results reveal that utilizing all layers (Stage 1–4) achieved the best performance, with a dice score of 88.3% and an HD of 2.34%. In contrast, excluding specific layers led to varied performance declines, with the shallow layers contributing significantly to contextual information and deeper layers enhancing fine-grained details. For example, when only deeper layers (Stages 3 and 4) were used, the dice score dropped to 78.5%, and the HD increased to 2.76%. In contrast, including only the shallow layers (Stages 1 and 2) yielded a dice score of 73.78% and an HD of 2.89%.

These findings underscore the necessity of a comprehensive fusion approach. Each layer's unique contributions—from the broad contextual cues in shallow layers to the detailed semantic information in deeper layers—work synergistically to enhance the model's ability to capture complex anatomical structures, ultimately improving overall segmentation accuracy and robustness.

5. Conclusion

We present Ref-SAM3D, a 3D-adapted SAM framework that synergizes cross-modal prompting and hierarchical attention to address medical segmentation challenges in volumetric imaging. Our model establishes a bidirectional interaction between visual data and semantic text descriptions, enabling intelligent segmentation through joint reasoning over volumetric imaging and clinical context. Three key innovations drive our methodology: (i) A cross-modal reference prompt generator that fuses text and image embeddings into a unified feature space through adaptive alignment, significantly enhancing spatial-semantic correlation, (ii) a multi-scale hierarchical attention mechanism that dynamically prioritizes critical anatomical features across dimensional scales while suppressing irrelevant noise, significantly improving segmentation robustness in intricate 3D topologies, and (iii) a volumetric architecture adaptation that transforms SAM's native 2D processing into true 3D computation through depth-aware convolutions and recursive mask refinement, effectively bridging the dimensional gap in medical imaging analysis. Extensive validation demonstrates state-of-the-art performance on complex segmentation tasks. While our approach is highly effective, future work is needed to focus on improving computational efficiency to enable real-time clinical applications, exploring semi-supervised learning techniques to address the challenge of limited labeled data. Overall, our method holds significant promise as a generalizable and robust segmentation framework, offering both fully automatic and promptable segmentation capabilities for a wide range of 3D medical imaging applications.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: Xiang Gao
Data curation: Xiang Gao
Investigation: Xiang Gao
Methodology: Xiang Gao
Visualization: Xiang Gao
Writing–original draft: Xiang Gao
Writing–review & editing: Kai Lu

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Data will be made available upon request to the corresponding author.

References

1. Obuchowicz R, Strzelecki M, Piorkowski A. Clinical applications of artificial intelligence in medical imaging and image processing-A review. *Cancers* (*Basel*). 2024;16(10):1870.

   doi: 10.3390/cancers16101870

2. Addimulam S, Mohammed MA, Karanam RK, *et al.* Deep learning-enhanced image segmentation for medical diagnostics. *Malays J Med Biol Res*. 2020;7(2):145-152.

3. Khalifa M, Albadawy M. AI in diagnostic imaging: Revolutionising accuracy and efficiency. In: *Computer Methods and Programs in Biomedicine Update*. Vol. 5; 2024.

4. Kirillov A, Mintun E, Ravi N, *et al.* Segment anything. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*; 2023. p. 4015-4026.

5. Zou X, Yang J, Zhang H, *et al. Segment Everything Everywhere all at Once.* arXiv *Preprint arXiv*: 2304.06718; 2023.

6. Huang Y, Yang X, Liu L, *et al.* Segment anything model for medical images? *Med Image Anal*. 2024;92:103061.

   doi: 10.1016/j.media.2023.103061

7. Hu EJ, Shen Y, Wallis P, *et al.* Lora: Low-rank adaptation of large language models. *arXiv preprint*:2106.09685, 2021.

8. Poth C, Sterz H, Paul I, *et al.* Adapters: A unified library for parameter-efficient and modular transfer learning. In: Feng Y, Lefever E, editors. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore*; 2023. p. 149-160.

9. Shen J, Wang W, Chen C, *et al. Medtuning: A New Parameter-efficient Tuning Framework for Medical Volumetric Segmentation. arXiv Preprint arXiv: 2304.10880*; 2024.

10. Zhang K, Liu D. *Customized Segment Anything Model for Medical Image Segmentation. arXiv preprint arXiv: 2304.13785*; 2023.

11. Wang H, Guo S, Ye J, *et al. Sam-med3d: Towards General-purpose Segmentation Models for Volumetric Medical Images. arXiv preprint arXiv: 2310.15161*; 2024.

12. Wu J, Ji W, Liu Y, *et al. Medical Sam Adapter: Adapting Segment Anything Model for Medical Image Segmentation. arXiv preprint arXiv: 2304.12620*; 2023.

13. Gong S, Zhong Y, Ma W, *et al.* 3dsamadapter: Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation. *Med Image Anal.* 2024;98:103324.

14. Xie B, Tang H, Duan B, Cai D, Yan Y. *Masksam: Towards Auto-prompt Sam with Mask Classification for Medical Image Segmentation. arXiv preprint arXiv: 2403.14103*; 2024.

15. Li C, Khanduri P, Qiang Y, Sultan RI, Chetty I, Zhu D. *Autoprosam: Automated Prompting Sam for 3d Multi-Organ Segmentation. arXiv preprint arXiv: 2308.14936*; 2024.

16. Zhang Y, Jiao R. *Towards Segment Anything Model (sam) for Medical Image Segmentation: A Survey. arXiv preprint arXiv: 2305.03678*; 2023.

17. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. *Nat Commun*. 2024;15(1):654.

   doi: 10.1038/s41467-024-44824-z.

18. Shaharabany T, Dahan A, Giryes R, Wolf L. *Autosam: Adapting Sam to Medical Images by Overloading the Prompt Encoder. arXiv preprint arXiv: 2306.06370*; 2023.

19. Na S, Guo Y, Jiang F, Ma H, Huang J. *Segment any Cell: A Sam-Based Auto-Prompting Finetuning Framework for Nuclei Segmentation. arXiv preprint arXiv: 2401.13220*; 2024.

20. Min B, Ross H, Sulem E, *et al.* Recent advances in natural language processing via large pre-trained language models: a survey. *ACM Comput Surv*. 2024;57(1):1-45.

   doi: 10.1145/3605943

21. Radford A, Kim JW, Hallacy C, *et al.* Learning transferable visual models from natural language supervision. *arXiv preprint arXiv*:2103.00020;2021.

22. Jia C, Yang Y, Xia Y, *et al.* Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv*:2102.05918; 2021.

23. Zou X, Yang J, Zhang H, *et al.* Segment everything everywhere all at once. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, editors. *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023,* New Orleans, LA, USA; 2023.

24. Wang X, Zhang X, Cao Y, Wang W, Shen C, Huang T. *Seggpt: Segmenting Everything in Context. arXiv Preprint arXiv: 2304.03284*; 2023.

25. Oquab M, Darcet T, Moutakanni T. *Dinov2: Learning Robust Visual Features without Supervision. arXiv Preprint arXiv: 2304.07193*; 2024.

26. Wang Y, Zhou W, Mao Y, Li H. *Detect any Shadow: Segment Anything for Video Shadow Detection. arXiv preprint arXiv: 2305.16698*; 2023.

27. Deng R, Cui C, Liu Q, *et al. Segment Anything Model (sam) for Digital Pathology: Assess Zero-Shot Segmentation on Whole Slide Imaging. arXiv preprint arXiv*: *2304.04155*; 2023.

28. He S, Bao R, Li J, *et al. Accuracy of Segmentanything Model (sam) in Medical Image Segmentation Tasks. arXiv preprint arXiv: 2304.09324*; 2023.

29. Hu C, Li X. *When Sam Meets Medical Images: An Investigation of Segment Anything Model (Sam) on Multi-Phase Liver Tumor Segmentation. arXiv preprint arXiv: 2304.08506*; 2023.

30. Zhou T, Zhang Y, Zhou Y, Wu Y, Gong C. *Can Sam Segment Polyps? arXiv preprint arXiv: 2304.07583*; 2023.

31. Cheng J, Ye Y, Deng Z, *et al. Sam-med2d. arXiv preprint arXiv: 2308.116184*; 2023.

32. Lei W, Wei X, Zhang X, Li K, Zhang S. *Medlsam: Localize and Segment Anything Model for 3D CT Images. arXiv preprint arXiv*: *2306.14752*; 2024.

33. Yang Y, Wu X, He T, Zhao H, Liu X. *Sam3d: Segment Anything in 3D Scenes. In: International Conference on Computer Vision*; 2023.

34. Chen C, Miao J, Wu D, *et al.* Ma-sam: Modality-agnostic sam adaptation for 3D medical image segmentation. *Med Image Anal*. 2024;98:103310.

35. Pan J, Lin Z, Zhu X, Shao J, Li H. *St-adapter: Parameter-Efficient Image-to-Video Transfer Learning. arXiv preprint arXiv: 2206.13559*; 2022.

36. Muksimova S, Umirzakova S, Baltayev J, Cho YI. Rl-cervix. net: A hybrid lightweight model integrating reinforcement learning for cervical cell classification. *Diagnostics* (*Basel*). 2025;15(3):364.

37. Jia M, Tang L, Chen BC, *et al. Visual Prompt Tuning. arXiv Preprint arXiv: 2203.12119*; 2022.

38. Radford A, Kim JW, Hallacy C, *et al.* Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning. PMLR*; 2021. p. 8748-8763.

39. Jia C, Yang Y, Xia Y, *et al.* Scaling up visual and vision-language representation learning with noisy text supervision. In: *International Conference on Machine Learning. PMLR*; 2021, pp. 4904–4916.

40. Dosovitskiy A. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929*; 2020.

41. Ding H, Liu C, Wang S, Jiang X. Vision-language transformer and query generation for referring segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*; 2021. p. 16321-16330.

42. Li Y, Zhang J, Teng X, Lan L, Liu X. *Refsam: Efficiently Adapting Segmenting Anything Model for Referring Video Object Segmentation. arXiv Preprint arXiv: 2307.00997*; 2024.

43. Heller N, Isensee F, Trofimova D. *The kits21 Challenge: Automatic Segmentation of Kidneys, Renal Tumors, and Renal Cysts in Corticomedullary-phase ct. arXiv Preprint arXiv: 2307.01984*; 2023.

44. Bilic P, Christ P, Li HB, *et al.* The liver tumor segmentation benchmark (LiTS). *Med Image Anal*. 2023;84:102680.
doi: 10.1016/j.media.2022.102680

45. Antonelli M, Reinke A, Bakas S, *et al.* The medical segmentation decathlon. *Nat Commun*. 2022;13(1):4128.
doi: 10.1038/s41467-022-30695-9

46. Zhuang X, Li L, Payer C. Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge. *Med Image Anal*. 2019;58:101537.
doi: 10.1016/j.media.2019.101537

47. Landman B, Xu Z, Iglesias J, Styner M, Langerak T, Klein A. Miccai multi-atlas labeling beyond the cranial vault-workshop and challenge. Vol. 5. In: *Proceeding MICCAI Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge*; 2015. p. 12.

48. Tang Y, Yang D, Li W, *et al. Self-supervised Pre-training of Swin Transformers for 3d Medical Image Analysis. arXiv preprint arXiv:2111.14791*; 2022.

49. Ji Y, Bai H, Yang J, *et al. Amos: A Large-scale Abdominal Multiorgan Benchmark for Versatile Medical Image Segmentation. arXiv preprint arXiv:2206.08023*; 2022.

50. Isensee F, Petersen J, Klein A, *et al. nnU-net: Self-Adapting Framework for u-net-Based Medical Image Segmentation. arXiv preprint arXiv: 1809.10486*; 2018.

51. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Wells WM 3rd, Frangi AF, editors. *Medical Image Computing and Computer-Assisted Intervention - MICCAI*, Nassir Navab, Joachim Hornegger; 2015.

52. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth H, Xu D. *Swin unetr: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. arXiv preprint arXiv: 2201.01266*; 2022.

53. Zhou HY, Guo J, Zhang Y, *et al.* nnformer: Volumetric medical image segmentation via a 3D transformer. *IEEE Trans Image Process.* 2023;32:4036-4045.
doi: 10.1109/TIP.2023.3293771

54. Shaker A, Maaz M, Rasheed H, *et al. Unetr++: Delving into Efficient and Accurate 3D Medical Image Segmentation. arXiv Preprint arXiv: 2212.04497*; 2024.

55. Lee HH, Bao S, Huo Y, Landman BA. *3D ux-net: A Large Kernel Volumetric Convnet Modernizing Hierarchical Transformer for Medical Image Segmentation. arXiv Preprint arXiv: 2209.15076*; 2023.