October 2025

VLM that uses pretrained BioBert weights for text part to use medical terminology text to help guide the segmentation in the provided image (use VLM to enhance image segmentation performance but on medical images and thus the text to guide the segmentation is medical related too)

Med-VLM: Enhancing Medical Image Segmentation Accuracy through Vision-Language Model

Yihao Zhao[1,2] Enhao Zhong[1,2] Cuiyun Yuan[1] Yang Li[1]
Man Zhao[1] Chunxia Li[3] Jun Hu[2] Wei Liu[4] Chenbin Liu[2]*

[1]National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital
& Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
[2]School of Electronic and Communication Engineering, Sun Yat-sen University
[3]Faculty of Health Sciences, Macau University
[4]Mayo Clinic

1942451444@qq.com, zhongenh@mail2.sysu.edu.cn, cuiyun.yuan@hotmail.com
liyang_ro@cicams-sz.org.cn, zhaoman@cicams-sz.org.cn, hujun25@mail.sysu.edu.cn
liu.wei@mayo.edu, chenbin.liu@gmail.com

Abstract

We proposed Med-VLM (Medical Vision-language Model), an innovative approach that leverages textual descriptions of organs to enhance segmentation accuracy in medical images. Existing medical image segmentation methods face several challenges: ***(1) Current medical segmentation models often fail to effectively incorporate valuable prior knowledge, such as detailed descriptions of organ locations and characteristics. (2) Most text-visual models prioritize target identification, rather than focusing on enhancing overall accuracy. (3) While some approaches attempt to use prior knowledge for accuracy enhancement, they often fall short in effectively incorporating pretrained models.*** *To overcome these limitations, Med-VLM introduced several key innovations:* ***low-rank adaptation, authoritative descriptions, BioBERT weights, and a feature mixer****. We conducted a comprehensive evaluation of MedVLM using three authoritative medical image datasets, covering the segmentation of various human body parts. Our method demonstrated superior performance compared to existing state-of-the-art approaches, including Lvit, MedSAM, SAM, and nnUnet. We designed a series of ablation experiments, which systematically assessed the contribution of each component of Med-VLM, providing insights into the model's performance characteristics.*

1. Introduction

Medical image segmentation plays a crucial role in radiation therapy, with organ delineation being a critical component[1, 2]. Traditionally, this process has been time-consuming, labor-intensive, and requiring high level of expertise. Automatic segmentation techniques have emerged as a solution to reduce workload, enhance consistency, and facilitate the analysis of large-scale datasets.

Convolutional neural networks (CNNs) initially domainated the field of medical image segmentation. [2–8] While these models achieved remarkable success in accurately segmenting target regions, they often suffered from task-specific design limitations. Their performance could deteriorate significantly when applied to new tasks or different imaging modalities. The introduction of transformers into image analysis in 2020 [9] marked a significant advancement in the field. Subsequent research [10, 11] adapted transformer networks for object detection and pixel-level precise segmentation. These models demonstrates superior generalization and transferability compared to traditional CNNs. However, transformer networks faced challenges related to the effective utilization of pre-trained weights. In this context, Ma et al.[12] adapted the Segment Anything Model (SAM) for the medical imaging field, addressing the unique challenges posed by medical images. This research inspired our approach: while transformer models have shown promise, there remains a need for enhanced strategies to fully exploit their capabilities. To this end, we not only adopted

*corresponding author

pre-trained image and text encoders, but also designed a Low-Rank Adaptation (LoRA)-based adapter[13] for the image encoder. Additionally, we utilized pre-trained weights from BioBERT[14] for the text encoder to tailor them specifically for medical imaging applications.

Numerous approaches have been explored to teach machines to comprehend the visual world through natural language[15–22]. Past research has demonstrated that Vision Transformers excel at mapping image representations to textual representation spaces. However, most existing models primarily focus on image-text dialogue and image-text matching tasks[15, 16, 18–20]. With the rise of generative artificial intelligence, some models have ventured into text-guided image generation[17, 22, 23], yet few have concentrated on text-guided segmentation specifically. Although the SAM model[11] claimed capabilities for text-guided segmentation, this feature is not present in its current released version. To address these gaps, we have developed our own text-image fusion architecture aimed at achieving high-precision segmentation enhanced by textual input. Our approach utilizes authoritative medical descriptions[24–33] as text inputs, incorporating critical information about organ boundaries and their relative positions. Furthermore, we have designed a query-based feature mixer[10] that thoroughly integrates image and text information before passing it to the decoder for output. This innovative architecture not only enhances segmentation accuracy but also represents a significant step forward in leveraging textual information within medical imaging contexts.

In summary, previous methods for text-guided delineation have the following limitations:
(1) Current medical automatic segmentation models do not effectively utilize prior knowledge, such as descriptions of organ locations.
(2) Most text-visual models aim to identify the target while segmenting, rather than improving accuracy.
(3) Some models attempt to use prior knowledge to enhance accuracy but do not incorporate pre-trained models.
To address these issues, our research has introduced low rank adaptation, authoritative descriptions, BioBERT weights, and a feature mixer to the segmentation model. This novel approach promises to significantly advance the field of medical image segmentation, offering improved accuracy, generalizability, and efficiency in radiation therapy planning and other medical applications.

2. Related works

Object detection & segmentation. In recent years, a variety of convolution-based methodologies have been developed to enhance object detection and segmentation capabilities. These approaches have laid the groundwork for significant advancements in the field, as evidenced by numerous studies that have explored their potential applications.[8, 15, 34–39]. With the emergence of transformers architectures, researchers have introduced innovative transformer-based techniques that excel in interactive segmentation tasks[10–12, 22, 40] Unlike prior methodologies that primarily focused on guiding segmentation through textual input or performing image-text matching post-segmentation, our research diverges from this path. Our objective is to leverage textual information to enhance the accuracy of segmentation.

Large multimodal model Multimodal learning has emerged as a powerful framework for addressing image-text matching challenges. In contrast to conventional convolutional models that predominatly utilize classification methods, multimodal learning demonstrates markedly enhanced generalization capabilities across diverse tasks[15–19]. It has also been successfully applied to video and image generation,showcasing its versatility and effectiveness[17, 22, 23]. Furthermore, multi-modal learning has found applications in both medical and natural image segmentation[11, 12, 22, 40], where it adeptly processes various prompt inputs such as points and bounding boxes. This integration of different data modalities enables models to achieve more precise and adaptable segmentation results, significantly improving performance across a range of applications.

Text-guided segmentation model Previous studies have explored the use of convolutional-based image encoders in conjunction with transformer-based text encoders to facilitate text-guided segmentation in medical imaging[41–44]. Building on the impressive performance of vision transformers in tasks involving image-text fusion, researchers have proposed several novel architectures designed to capitalize on these advancements. Additionally, a growing body of work has investigated the application of diffusion-based methods for segmentation, further expanding the toolkit available for enhancing segmentation accuracy in complex medical imaging scenarios[45, 46].

- image encoder
- text encoder
- 2 LORA adapters (i think one for image encoder and one for text encoder)
    - image and text encoders are pretrained models that have their weights frozen, the lora adapters, feature mixer and decoder have their weights updated during training
- feature mixer (mix image and text embeddings)
- mask decoder -> decode embeddings to create segmentation of the provided image (using info/context from text embedding and image embeddings)

3. Method

3.1. Model Architecture and Workflow

Our model comprises five main components: a transformer-based[9] image encoder, a BERT-based text encoder[15], two LoRA-based adapter[13], a query-based feature mixer[10], and a mask decoder. During training, the parameters of the image encoder and text encoder remain frozen,

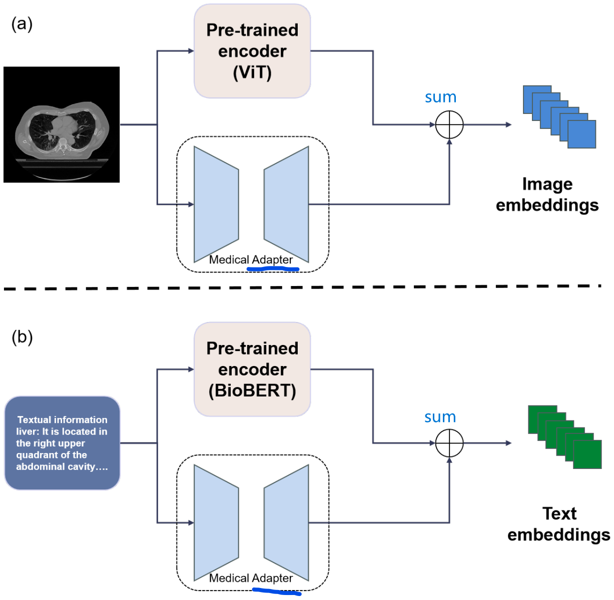
(a)
Pre-trained encoder (ViT)
sum
Image embeddings
Medical Adapter
(b)
Pre-trained encoder (BioBERT)
Textual information liver: It is located in the right upper quadrant of the abdominal cavity....
sum
Text embeddings
Medical Adapter

Figure 1. The structure of the image and text encoders

while the adapters, feature mixer, and mask decoder were trained. The model contains approximately 200 million parameters, with 5.2 million trainable parameters. Due to the limited availability of medical image datasets, particularly those with textual descriptions, training the entire model would likely lead to severe overfitting. To solve this issue, we initialized the image encoder and text encoder with weights from models pretrained on natural images[15]. We then employed adapters to fine-tune these components for our specific task. The model's input consists of a CT image slice and corresponding description provided by a clinical expert. This description includes the detailed information about the organ's anatomical position and its spatial relationship with surrounding structures. The model's output is a precise segmentation mask of the target organ or structure.

3.2. Image encoder

ViT is vision transformer

The architecture of the image encoder is illustrated in Fig.1.a. To balance the computational efficiency and performance, we used the ViT-base model as the image encoder. Our objective was to fine-tune these parameters to optimize them for the medical image segmentation tasks. Inspired by recent advancements in large language models, we implemented Low-rank Adaptation (LoRA)[13] to fine-tune our vision encoder. The LoRA adapter is defined as follows:

$$LoRA(I) = A \times B(I) \tag{1}$$

where I is the input image, A is initialized randomly and B is initialized as a zero matrix, with dimensions of 768 × 8 and 8 × 768, respectively. The approach offers a significant advantage: during initial training, the product of the two matrices is zero, thereby preserving the original weights and mitigating potential unexpected outcomes. As training progresses, the values in both matrices are adjusted through backpropagated gradients, allowing them to adapt to specific tasks. The formalized computation process can be expressed as:

image embedding is the sum of the image going through LorA and the pretrained ViT individually / separately

$$F_{im} = E_{image}(I) + L_{image}(I) \tag{2}$$

where I represents the input image, $E_{image}(.)$ denotes the image encoder function, $L_{image}(.)$ is the LoRA adapter, and F_{im} is the resulting image feature vector.

3.3. Text encoder

The overall structure of the text encoder is illustrated in Fig.1.b. We selected BERT-base, an encoder-only architecture, as our text encoder[47, 48]. It is important to note that we utilized the pre-trained weights of BioBERT[14] instead of the standard BERT. The decision was made because BioBERT was trained on a corpus of 1M tokens from PubMed[49] and PMC[50], potentially offering superior performance in the medical domain compared to the standard BERT[47, 50]. To further optimize the text encoder for our specific task, we incorporated the same adapter mechanism as used in the image encoder. The text feature vector is computed as follows:

$$F_{text} = E_{text}(T) + L_{text}(T) \tag{3}$$

where F_{text} is the text feature vector, T is the input text, $E_{text}()$ represents the text encoder, and $L_{text}()$ denotes the text adapter.

3.4. Feature mixer

The structure of the feature mixer and the mask decoder is illustrated in Fig.2 Inspired by SAM[11], we designed a query-based image-text feature fusion module. It comprises a self-attention module[48], two cross-attention modules, and a feed-forward neural network[51].

After extracting the image and text information, denoted as F_{im} and F_{text} respectively, through the image encoder and text encoder, the text vector first passes through a self-attention module. Subsequently, it serves as a query in a cross-attention module with the image vector, followed by an multilayer perceptron (MLP). This process outputs the fused text vector $F_{fused\ text}$. This fused text vector then acts as a query in another cross-attention module with F_{im}, resulting in the fused image vector output $F_{fused\ im}$. Each computation in the above process is accompanied by skip connections to mitigate the gradient vanishing problem[52]. The formalized computation process is as follows:

F_im is the image embedding outputted from the image encoder

residual self attention transformer

$$F_{\text{text},1} = F_{\text{text}} + \text{attn}(F_{\text{text}}, F_{\text{text}}) \tag{4}$$

$$F_{\text{text},2} = F_{\text{im}} + \text{cross\_attn}(F_{\text{text},1}, F_{\text{im}}) \tag{5}$$

residual

cross attention transformer with the F_text1 as query and F_im output of image encoder as the key and va

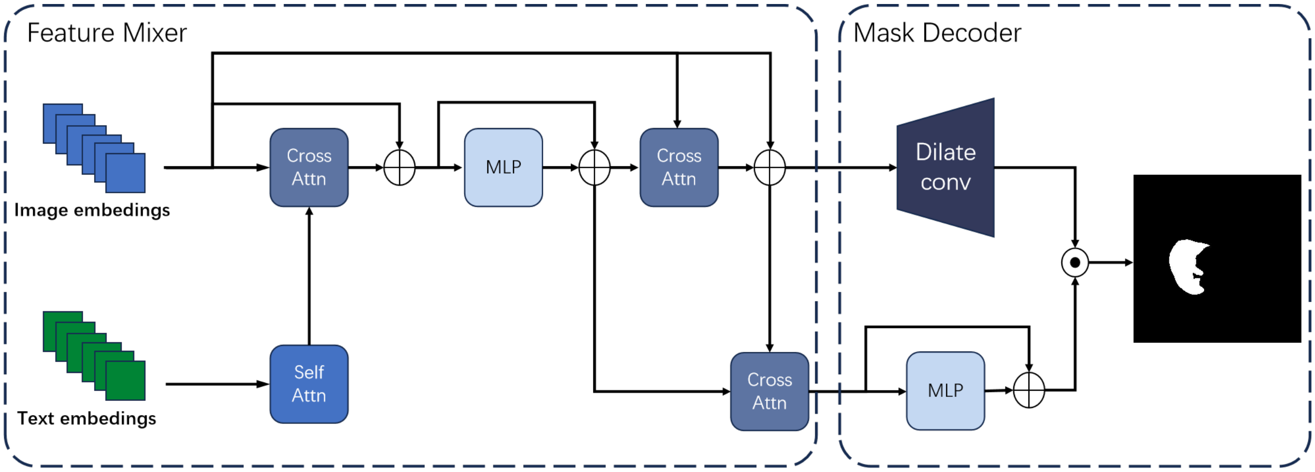

Figure 2. The structure of the feature mixer and mask decoder

F_im is output of image encoder

residual

$$\underline{F_{\text{fused text}}} = F_{\text{text},2} + \text{MLP}(F_{\text{text},2}) \tag{6}$$

$$\underline{F_{\text{fused im}}} = F_{\text{im}} + \text{cross\_attn}(F_{\text{fused text}}, F_{\text{im}}) \tag{7}$$

residual

where $attn$ represents the self-attention operation, $cross\_attn$ denotes the cross-attention operations, and MLP is the multilayer perceptron module.

3.5. Mask decoder

The mask decoder's primary function is to generate pixel-level segmentation masks from the fused image and text vectors. It comprises three main components: dilated convolutions, bi-directional transformer, and a MLP. The architecture is designed to efficiently process and combine spatial and semantic information. The dilated convolutions serves a dual purpose: reduce the number of channels, and increase the size of the feature map. The operation allows the network to expand the receptive field without losing spatial resolution, which is crucial for precise segmentation tasks. The bi-directional attention structure facilitates further fusion between the previously output text feature vector and the processed image feature map. This attention mechanism enables the model to capture long-range dependencies and context, enhancing the integration of textual and visual information. The mask decoder can be formalized as follows:

$$F_1 = \text{Dilated\_Conv}(\underline{F_{\text{fused im}}}) \tag{8}$$

$$F_2 = \text{cross\_attn}(\underline{F_{\text{fused im}}}, \underline{F_{\text{fused text}}}) \tag{9}$$

$$\text{Mask} = \text{MLP}(F_2) * F_1 \tag{10}$$

where $F_{fused\ im}$ is the fused image vector, and $F_{fused\ text}$ is the fused text vector, and $*$ denotes element-wise multiplication. The feature mixer, a key component in the fusion process, is stacked four times to ensure thorough blending of image and text features. During the training process, both the feature mixer and the mask decoder are optimized simultaneously.

3.6. Loss function

IMPORTANT: here they use BINARY cross entropy + dice (unweighted sum so theur coeffs are 1)

We utilized the unweighted sum of binary cross-entropy loss and Dice loss as the final loss function, as it has proven to be robust across various medical image segmentation tasks.

4. Experiment

4.1. Implementation Details

Our model architecture comprises of four key components: an image encoder with its adapter, a text encoder with its adapter, a text-image feature mixer, and a mask decoder. For the image encoder, we utilized a pretrained ViT-base[9], while the text encoder employs a pretrained BioBERT-base[14]. The model contains approximately 200 million parameters, of which 5.2 million are trainable. Training was conducted on a computing server equipped with eight NVIDIA A100 40GB GPUs. The training dataset consisted of approximately 47,000 images and 100,000 high-quality masks. We employed a batch size of 16 and trained the model for 150 epochs, resulting in a total training time of approximately 30 hours.

4.2. Dataset

For training and testing, we utilized data from three prominent medical imaging datasets: FLARE[53], SegTHOR[54], and MSD[55]. It is noteworthy that in the MSD dataset includes annotations for both organs and tumors, our study primarily focuses on the impact of organ location description on segmentation performance. Given the inherent uncertainty in tumor location, we restricted our analysis and training to the organ-specific portion of higher DSC is better, lower HD and ASD is better

Table 1. The comparison results on test datasets of five different methods, including our proposed approach, are summarized across all datasets used in this experiment. All data are presented as mean ± standard deviation, with the best result for each metric highlighted in bold.

Datasets		nnUnet	SAM	MedSAM	Lvit	Med-VLM(ours)
ALL	DSC	0.946±0.033	0.568±0.162	0.953±0.029	0.953±0.029	**0.976±0.029**
	HD_{95}	45.125±62.61	98.125±101.33	49.225±60.37	48.321±50.97	**21.86±39.95**
	ASD	13.105±13.94	24.275±27.055	12.205±16.675	11.86±8.99	**4.56±4.67**
FLARE	DSC	0.975±0.020	0.410±0.16	0.980±0.015	0.977±0.021	**0.990±0.015**
	HD_{95}	36.29±65.94	68.445±99.265	35.225±60.37	33.72±50.33	**23.08±40.15**
	ASD	13.11±20.775	17.105±38.08	10.255±20.105	9.97±15.334	**4.57±8.25**
SegTHOR	DSC	0.953±0.021	0.486±0.151	0.956±0.023	0.955±0.018	**0.981±0.014**
	HD_{95}	9.98±15.725	18.595±26.445	9.085±16.925	11.32±19.01	**3.1025±11.765**
	ASD	3.055±3.94	4.89±6.435	2.155±4.86	3.109±4.13	**1.55±2.51**
MSD	DSC	0.900±0.038	0.860±0.0798	0.908±0.04	0.901±0.055	**0.947±0.04**
	HD_{95}	72.555±58.325	85.06±56.215	65.61±63.195	67.16±57.124	**25.56±42.25**
	ASD	20.665±17.59	21.335±22.61	17.435±19.18	18.99±21.91	**5.21±9.95**

the MSD dataset. To enhance the segmentation process, we incorporated descriptive language regarding organ locations from authoritative medical texts[24–33].

4.3. Evaluation metrics

We adhered to the guidelines in Metrics Reloaded[56]. Our quantitative assessment of segmentation outcomes employs three primary metrics: the Dice Similarity Coefficient (DSC)[57], the 95th percentile Hausdorff distance (HD_{95})[58], and average surface distance (ASD)[59]. The DSC is a region-based segmentation metric designed to evaluate the overlap between expert annotation masks and segmentation results. It is defined by the following formula:

$$\text{DSC}(GT, AGC) = \frac{2|GT \cap AGC|}{|GT| + |AGC|} \tag{11}$$

where GT is the ground truth and AGC is the automatically generated contours.
HD_{95} and ASD are boundary-based metrics to evaluate the boundary consensus between expert annotation masks and segmentation results at a given tolerance. These metrics are defined as follows:
For a one-sided Euclidean distance from point set X to point set Y:

$$d(X \to Y) = \max_{x \in X} \min_{y \in Y}(d(X \to Y)) \tag{12}$$

The HD_{95} is calculated as:

$$\begin{aligned} &\text{HD}_{95}(GT, AGC) = \\ &\max_{95\%}(d(GT \to AGC), d(AGC \to GT)) \end{aligned} \tag{13}$$

The ASD is defined as:

$$\text{ASD}(\text{GT}, \text{AGC}) = \frac{1}{N_{\text{GT}} + N_{\text{AGC}}} \left(\sum_{x \in \text{GT}} \min_{y \in \text{AGC}} \|x - S(\text{AGC})\| + \sum_{y \in \text{AGC}} \min_{x \in \text{GT}} \|y - S(\text{GT})\| \right) \tag{14}$$

where HD_{95} is the longest bidirectional distance between the ground truth and automatically generated contours at the 95th percentile, N_{GT} and N_{AGC} are the numbers of pixels in the contour of ground truth and automatically generated contours, respectively. $S(GT) and S(AGC)$ represent the surfaces of the ground truth and automatically generated contours, respectively.

4.4. Comparison with other methods

To statistically analyze and compare the performance of the four methods mentioned (MedSAM[12], SAM[11], U-Net[1], and Lvit[60] specialist models), we conducted a comprehensive evaluation using multiple metrics on the test dataset. We calculated the mean and variance of these metrics for each method to assess their segmentation performance. This analysis aimed to determine whether any of the methods demonstrated statistically superior segmentation accuracy compared to the others, offering valuable insights into the comparative effectiveness of the evaluated methods.

5. Result

5.1. Statistical results

Tab.1 demonstrates the superiority of our method compared to other approaches. It presents the mean and variance of various performance indicators in the test set, providing a comprehensive comparative analysis across different datasets. The result high-light the improved accuracy and robustness of our proposed model.

Our approach shows significant improvements over the previous state-of-the-art methods, MedSAM, in both region-based and boundary-based segmentation metrics. The mean DSC similarity increased by 0.023, indicating better overall segmentation accuracy. More notably, our method achieved substantial enchancements in boundary delineation precision: the mean HD_{95} decreased by 23.265, with a variance reduction of 11.02; the mean ASD decreased by 7.3, with a variance reduction of 4.32. These results underscore the effectiveness of our approach in improving the precision of boundary delineation, a critical aspect in medical image segmentation.

6. Ablation experiments

To validate the effectiveness of our innovations, including the utilization of authoritative descriptions, the integration of image and text adapters, and the application of BioBERT pre-trained weights, we have designed a series of ablation experiments.

6.1. Effectiveness of authoritative description

We conducted an ablation experiment to evaluate the impact of inputting authoritative text descriptions on segmentation results. The model was retrained and tested using three different types of input: complex descriptions (our method), simple descriptions (only organ names), and no descriptions. This approach allowed us to assess how varying levels of descriptive detail affect the model's segmentation performance. The results of this experiment are presented in Tab.2.

Our findings indicate that merely adding text to label the organ does not significantly improve the contour delineation. However, the use of authoritative textual descriptions substantially enhances the accuracy of the automated contouring, demonstrating the value of detailed expert input in refining segmentation precision. These results suggests a promising direction for future research: the development of a set of authoritative guidelines specifically designed to be better understood by models, with the aim of enhancing automatic contouring. Such guidelines would leverage expert knowledge to improve the precision and reliability of automated segmentation processes.

6.2. The architecture of the text encoder

We utilized BERT, an encoder-only architecture, as our text encoder, hypothesizing that this architecture is advantageous for generating text embeddings. To validate our approach, we conducted comparative experiments using various architectures: GPT-2[37], LLAMA(both LLAMA2[61] and LLAMA3[62]), which are based on casual decoders, as well as GLM[63], which is based on a prefix decoder. It's noteworthy that previous researchers have fine-tuned BERT [49] [14] and GPT-2[64] on the PubMed dataset. We also conducted experiments using these fine-tuned weights and compared them with weights trained on general natural language corpora.

Table 2. The impact of description types on segmentation performance.

Description types	None	Simple	Complex
DSC	0.953	0.953	**0.976**
HD_{95}	48.31	48.28	**21.86**
ASD	18.41	17.81	**4.56**

The results of our ablation experiments are presented in the Tab.3. Our findings indicate that models with encoder-only architectures significantly outperform other models in terms of text embedding capabilities. Interestingly, we observed that models with a larger number of parameters did not show a significant improvement in segmentation accuracy compared to smaller models, validating the appropriateness of our chosen model size. Additionally, models fine-tuned using the PubMed database demonstrate superior performance compared to their pre-fine-tuned counterparts. This further validates our choice of the BERT architecture and BioBERT weights for our text encoder.

6.3. The efficacy of LoRA-based adapter

To demonstrate the effectiveness of our method using a LoRA-based adapter to fine-tune the pre-trained image encoder and text encoder for adapting it to medical images, we conducted ablation experiments under four conditions: not using the adapter (baseline), using the adapter only for the text encoder, using the adapter only for the image encoder, and using the adapter for both encoders. The result of these experiments are presented in Tab.4.

Our findings indicate that fine-tuning both the image encoder and the text encoder contributes significantly to improving segmentation accuracy. Notably, fine-tuning the image encoder yields better results than fine-tuning the text encoder using LoRA-based adapter.

6.4. The structure of the adapter

Among the current methods for fine-tuning pre-trained models, besides the LoRA approach mentioned in this paper, other methods have also demonstrated excellent performance, such as Prefix Tuning[65] and adapter tuning[66]. We compared these three fine-tuning approaches, and the results are presented in Tab.5. Our findings indicate that the LoRA tuning yielded the best outcomes, followed by adapter tuning, while prefix tuning produced the least effec-

Note these are just results of different models for 4 different examples that are for different organs each having a CT image and text prompt as input. To see overall performance on the datasets, look at Table 1.

higher DSC is better, lower HD and ASD is better

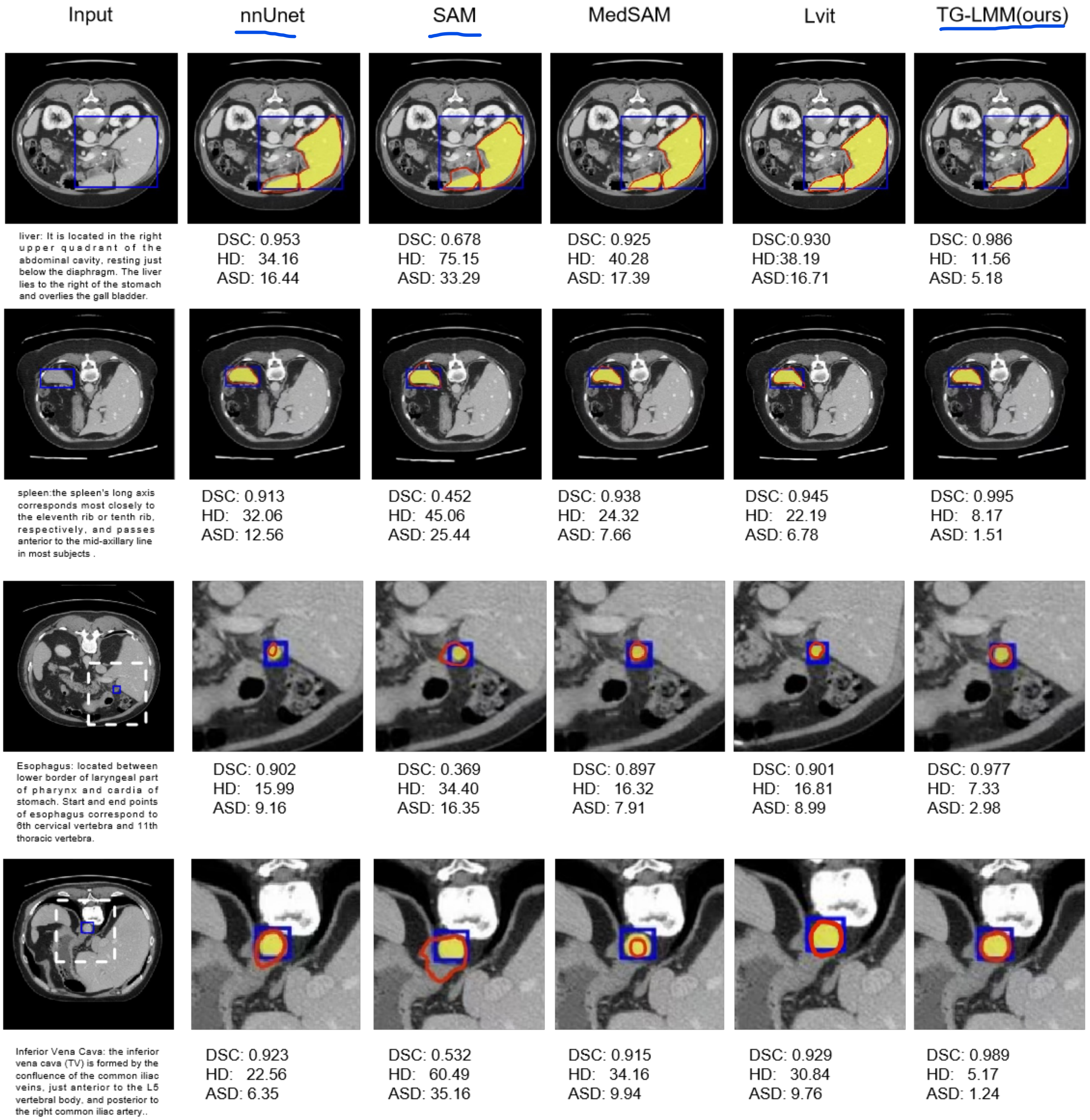
Input
nnUnet
SAM
MedSAM
Lvit
TG-LMM(ours)
liver: It is located in the right upper quadrant of the abdominal cavity, resting just below the diaphragm. The liver lies to the right of the stomach and overlies the gall bladder.
DSC: 0.953 HD: 34.16 ASD: 16.44
DSC: 0.678 HD: 75.15 ASD: 33.29
DSC: 0.925 HD: 40.28 ASD: 17.39
DSC:0.930 HD:38.19 ASD:16.71
DSC: 0.986 HD: 11.56 ASD: 5.18
spleen:the spleen's long axis corresponds most closely to the eleventh rib or tenth rib, respectively, and passes anterior to the mid-axillary line in most subjects .
DSC: 0.913 HD: 32.06 ASD: 12.56
DSC: 0.452 HD: 45.06 ASD: 25.44
DSC: 0.938 HD: 24.32 ASD: 7.66
DSC: 0.945 HD: 22.19 ASD: 6.78
DSC: 0.995 HD: 8.17 ASD: 1.51
Esophagus: located between lower border of laryngeal part of pharynx and cardia of stomach. Start and end points of esophagus correspond to 6th cervical vertebra and 11th thoracic vertebra.
DSC: 0.902 HD: 15.99 ASD: 9.16
DSC: 0.369 HD: 34.40 ASD: 16.35
DSC: 0.897 HD: 16.32 ASD: 7.91
DSC: 0.901 HD: 16.81 ASD: 8.99
DSC: 0.977 HD: 7.33 ASD: 2.98
Inferior Vena Cava: the inferior vena cava (TV) is formed by the confluence of the common iliac veins, just anterior to the L5 vertebral body, and posterior to the right common iliac artery..
DSC: 0.923 HD: 22.56 ASD: 6.35
DSC: 0.532 HD: 60.49 ASD: 35.16
DSC: 0.915 HD: 34.16 ASD: 9.94
DSC: 0.929 HD: 30.84 ASD: 9.76
DSC: 0.989 HD: 5.17 ASD: 1.24

Figure 3. The segmentation examples generated by our method, alongside comparisons with other methods, are presented. The blue bounding box denotes the input prompt, the yellow filled area represents the ground truth, and the red contour indicates the automatically generated segmentation.

tive results. The implementation details of this experiment are provided in the supplemental materials

7. Conclusion and discussion

7.1. Conclusion

We developed an innovative automatic segmentation model for medical images that leverages authoritative organ de-

Table 3. Comparison of different text encoder architectures.

Metric	GPT2	GPT2-large	BioGPT	LLAMA2	LLAMA3	GLM	BERT	BERT-large	BioBERT
DSC	0.933	0.939	0.943	0.949	0.946	0.933	0.970	0.969	**0.976**
HD_{95}	49.68	43.29	37.88	33.18	34.19	48.19	29.90	28.67	**21.8**
ASD	12.20	14.12	11.39	9.96	10.19	15.39	8.91	9.65	**4.56**

Table 4. Ablation study on fine-tuning the pre-trained text and image encoders.

Metric	Baseline	Text	Image	Both
DSC	0.956	0.961	0.969	**0.976**
HD_{95}	51.78	42.86	32.89	**21.86**
ASD	28.19	19.16	7.19	**4.56**

Table 5. Comparison of fine-tuning methods.

Metric	Prefix tuning	Adapter tuning	LoRA
DSC	0.959	0.966	**0.976**
HD_{95}	32.16	27.61	**21.86**
ASD	11.29	7.26	**4.56**

scriptions. Our approach incorporates four key innovations: a LoRA-based adapter for fine-tuning, medically fine-tuned text encoder weights, a query-based feature mixer, and the use of authoritative medical descriptions as input. These advancements have led to state-of-the-art performance in medical image segmentation tasks, as validated through extensive experiments on three authoritative datasets[53–55]. Comprehensive ablation studies further elucidated each component's contribution to the model's overall performance.

7.2. Discussion

Utilizing Authoritative Descriptions. A significant reason why this method outperforms MedSAM is its utilization authoritative descriptions as input, making the entire model training process resemble the learning journey of a medical student, thereby yielding impressive results. The quality of automated contouring has long been a significant challenge for researchers[67]. Our model achieves better robustness due to its richer input representations compared to previous models. This enables it to perform well even on images that were previously considered difficult to segment, as evidenced by the variance in the Tab.1

The encoder-only structured text encoder. Despite the emergence of a numerous large language models serving generative tasks, BERT with its encoder-only architecture remains the preferred choice for text embedding, as validated by our experiments. We hypothesize that the text-guided segmentation requires extracting specific representations from text, such as the location and category of organs. This process necessitates filtering out irrelevant information from the text input, a task for which the encoder architecture is more suited than the decoder. In addition, we found that increasing the size of the language encoder does not necessarily lead to significant performance improvements. Compared to the BERT-base model, the main enhancements in the BERT-large model lie in the dimensionality of word embeddings and the depth of the model. We posit that the vocabulary contained in our input text does not fully utilize the higher embedding dimensionality of the larger model. Furthermore, this task does not require high-level semantic information; shallow information such as spatial positions and contour features is sufficient.

Limitations: Due to limitations in system resources, we have not yet adopted a 3D image encoder, which may restrict our ability to extract information in the vertical direction. Apart from the authoritative descriptions mentioned in this paper, there are many other forms of prior knowledge, such as human anatomy atlas segmentation and authoritative instructional books. In future work, we hope to leverage all of these using knowledge graph methods to achieve a more accurate and versatile segmentation system.

References

[1] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 1, 5

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1

[3] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016.

[4] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi.

V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[5] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[7] Hao Chen, Qi Dou, Lequan Yu, Jing Qin, and Pheng-Ann Heng. Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images. *NeuroImage*, 170: 446–455, 2018.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 2

[9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 4

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 2

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 2, 3, 5

[12] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 1, 2, 5

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 3

[14] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. 2, 3, 4, 6

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[16] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022. 2

[17] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[20] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[21] Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. M6: A chinese multimodal pretrainer. *arXiv preprint arXiv:2103.00823*, 2021.

[22] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 2

[23] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 2

[24] Francesco Alessandrino, Aleksandar M Ivanovic, Daniel Souza, Amin S Chaoui, Jelena Djokic-Kovac, and Koenraad J Mortele. The hepatoduodenal ligament revisited: cross-sectional imaging spectrum of non-neoplastic conditions. *Abdominal Radiology*, 44:1269–1294, 2019. 2, 5

[25] LM Biga, S Dawson, A Harwell, R Hopkins, J Kaufmann, M LeMaster, P Matern, K Morrison-Graham, D Quick, J Runyeon, et al. Anatomy & physiology. openstax/oregon state university, 2021.

[26] Jarrod A Collins, Julie-Vanessa Munoz, Toral R Patel, Marios Loukas, and R Shane Tubbs. The anatomy of the aging aorta. *Clinical anatomy*, 27(3):463–466, 2014.

[27] Demetrios Demetriades, Kenji Inaba, and George Velmahos. *Atlas of surgical techniques in trauma*. Cambridge University Press, 2020.

[28] Detlev Drenckhahn and Jens Waschke. *Taschenbuch anatomie*. Elsevier Health Sciences, 2020.

[29] Keith L Moore, Arthur F Dalley, and MR Agur. Clinically oriented anatomy. 7. izdanje, 2014.

[30] D Neil Granger, Lena Holm, and Peter Kvietys. The gastrointestinal circulation: physiology and pathophysiology. *Comprehensive Physiology*, 5(3):1541–1583, 2011.

[31] Shuja Rizvi, Chase J Wehrle, and Mark A Law. Anatomy, thorax, mediastinum superior and great vessels. In *StatPearls [Internet]*. StatPearls Publishing, 2023.

[32] Omesh Singh and Srinivasa Rao Bolla. Anatomy, abdomen and pelvis, prostate. 2019.

[33] Susan Standring, Harold Ellis, J Healy, D Johnson, A Williams, P Collins, and C Wigley. Gray's anatomy: the anatomical basis of clinical practice. *American journal of neuroradiology*, 26(10):2703, 2005. 2, 5

[34] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019. 2

[35] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[36] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[37] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 6

[38] J Redmon. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 2

[40] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3635–3647, 2024. 2

[41] Shahina Kunhimon, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. Language guided domain generalized medical image segmentation. *arXiv preprint arXiv:2404.01272*, 2024. 2

[42] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023.

[43] Go-Eun Lee, Seon Ho Kim, Jungchan Cho, Sang Tae Choi, and Sang-Il Choi. Text-guided cross-position attention for segmentation: Case of medical image. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 537–546. Springer, 2023.

[44] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20730–20740, 2022. 2

[45] Zheyuan Zhang, Lanhong Yao, Bin Wang, Debesh Jha, Elif Keles, Alpay Medetalibeyoglu, and Ulas Bagci. Emit-diff: Enhancing medical image segmentation via text-guided diffusion model. *arXiv preprint arXiv:2310.12868*, 2023. 2

[46] Zhiwei Dong, Genji Yuan, Zhen Hua, and Jinjiang Li. Diffusion model-based text-guided enhancement network for medical image segmentation. *Expert Systems with Applications*, 249:123549, 2024. 2

[47] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[48] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[49] Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48, 2017. 3, 6

[50] Wonjin Yoon, Chan Ho So, Jinhyuk Lee, and Jaewoo Kang. Collabonet: collaboration of deep neural networks for biomedical named entity recognition. *BMC bioinformatics*, 20:55–65, 2019. 3

[51] Allan Pinkus. Approximation theory of the mlp model in neural networks. *Acta numerica*, 8:143–195, 1999. 3

[52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[53] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Ma, Adamo Young, Cheng Zhu, Kangkang Meng, Xin Yang, Ziyan Huang, et al. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. *arXiv preprint arXiv:2308.05862*, 2023. 4, 8

[54] Tao He, Junjie Hu, Ying Song, Jixiang Guo, and Zhang Yi. Multi-task learning for the segmentation of organs at risk with label dependence. *Medical Image Analysis*, 61:101666, 2020. 4

[55] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022. 4, 8

[56] Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. Metrics reloaded: recommendations for image analysis validation. *Nature methods*, 21(2):195–212, 2024. 5

[57] László Orlóci. Ordination by resemblance matrices. In *Ordination of plant communities*, pages 239–275. Springer, 1978. 5

[58] Henry Blumberg. Hausdorff's grundzüge der mengenlehre. 1920. 5

[59] Tobias Heimann, Bram Van Ginneken, Martin A Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, György Bekes, et al. Comparison and evaluation of methods for liver segmentation from ct datasets. *IEEE transactions on medical imaging*, 28(8):1251–1265, 2009. 5

[60] Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. Lvit: language meets vision transformer in medical image segmentation. *IEEE transactions on medical imaging*, 2023. 5

[61] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 6

[62] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 6

[63] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*, 2021. 6

[64] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pretrained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022. 6

[65] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning, 2021. 6

[66] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 6

[67] Yihao Zhao, Cuiyun Yuan, Ying Liang, Yang Li, Chunxia Li, Man Zhao, Jun Hu, Ningze Zhong, and Chenbin Liu. One class classification-based quality assurance of organs-at-risk delineation in radiotherapy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4898–4906, 2024. 8