

Figure 1.The framework of our proposed method xxx consist of four main components : (a) The raw ultrasound video is initially departed into several clips. (b) The Multi-exports Mask Generation Network conduct the calculation of the token score (c) The encoder is trained by reconstructing the masked the cubic tokens according to the token score with decoder. (d) The final diagnostic result of a prostate ultrasound imaging clip is calculated using the token score and features extracted by the encoder through the classification head.

**3.Methodology**

In this section,we first present the motivation and overall framework of the proposed method xxx in Section 3.1. Next , we turn into the details of the proposed Multi-exports Mask Generation Network (MeMGN) in the pre-training phase and the additional prompt based on MMGN in the downstream classification fine-tuning phase in Section 3.2 and 3.3.

**3.1 Overview**

**3.1.1 Motivation**

Given a prostate ultrasound video contains frames , we aim to address the following two issues:

1. ***Insufficient key region representation pre-training leads to details missing****：*The mainstream masking methods lack the concentration on the lesion areas of prostate ultrasound images, which may influence the feature extraction in these key regions crucial for downstream tasks.
2. ***Direct applying general masked features leads to insufficient classification***: Most downstream classification methods using masked features directly on the pre-training encoder, which may cause the model to overlook generalizable features, weakening robustness and cross-domain performance.

To address these issues , we propose xxx that mask the tokens by MeMGN, guided by multiple pre-trained medical expert models and give additional prompt based on MeMGN in the downstream classification fine-tuning. This enables us to make the training of our encoder **representation extraction focus on key regions** , while paying more attention to **general features** during classification fine-tuning.

**3.1.2 Framework**

In this section , we will briefly introduce the overall framework of the proposed xxx in Fig.1. Given a prostate ultrasound video , which contains represents the frame, height, and width of the video. The objective of XXX is to focus on key regions in prostate ultrasound images to achieve highly accurate classification. The framework is divided into two stages: pretraining and fine-tuning. Here's a refined and academically appropriate version of your paragraph, with improved clarity and structure:

The objective of the pre-training phase in our XXX framework is to train the encoder to learn meaningful feature representations from video ultrasound images. The input video is divided into non-overlapping 3D cubic tokens of size . Then, the token is projected into a feature vector of dimension (set to 768 in our design). During the fine-tuning phase, the encoder extracts features from the input video are fed into a classification head to predict the final label of the video.

**3.2 Multi-exports guided Pre-training**

This section first provides a brief overview of the main framework in the pre-training phase, then delves into our proposed multi-expert mask generation network.

**3.2.1 Pre-training phase**

For an input video *V* with *T* frames, we adopt a data sampling strategy with a stride of 4 to reduce redundancy, resulting insampled frames. These frames are further divided into *L* clips *C*, each containing equals to 16 consecutive frames, where. Consequently, the input to the pre-training network is represented as *C* .

To tokenize the input, we apply a 3D convolution to each clip , reducing its dimensionality to . This operation uses a convolutional kernel of size *k*, stride *s*, and output channels *d*, and can be formulated as follows:

(1)

where *N* equals torepresents the token number.

After the patch embedding , the embeded tokens and the clip are given to the MeMGN , outputing the mask scores , can be represented as:

(2)

Then, we mask the top-k tokens with the highest scores, where and is the masking ratio. The remaining tokens are fed into the encoder of Vision Transformer for feature extraction and representation learning. This process can be expressed as:

(3)

where has the same dimensionality with , which equals to

Finally, the encoded visible tokens is appended with masked tokens to get the combined tokens , which is then fed into the decoder of the Vision Transformer to obtain the reconstruction result .

(4)

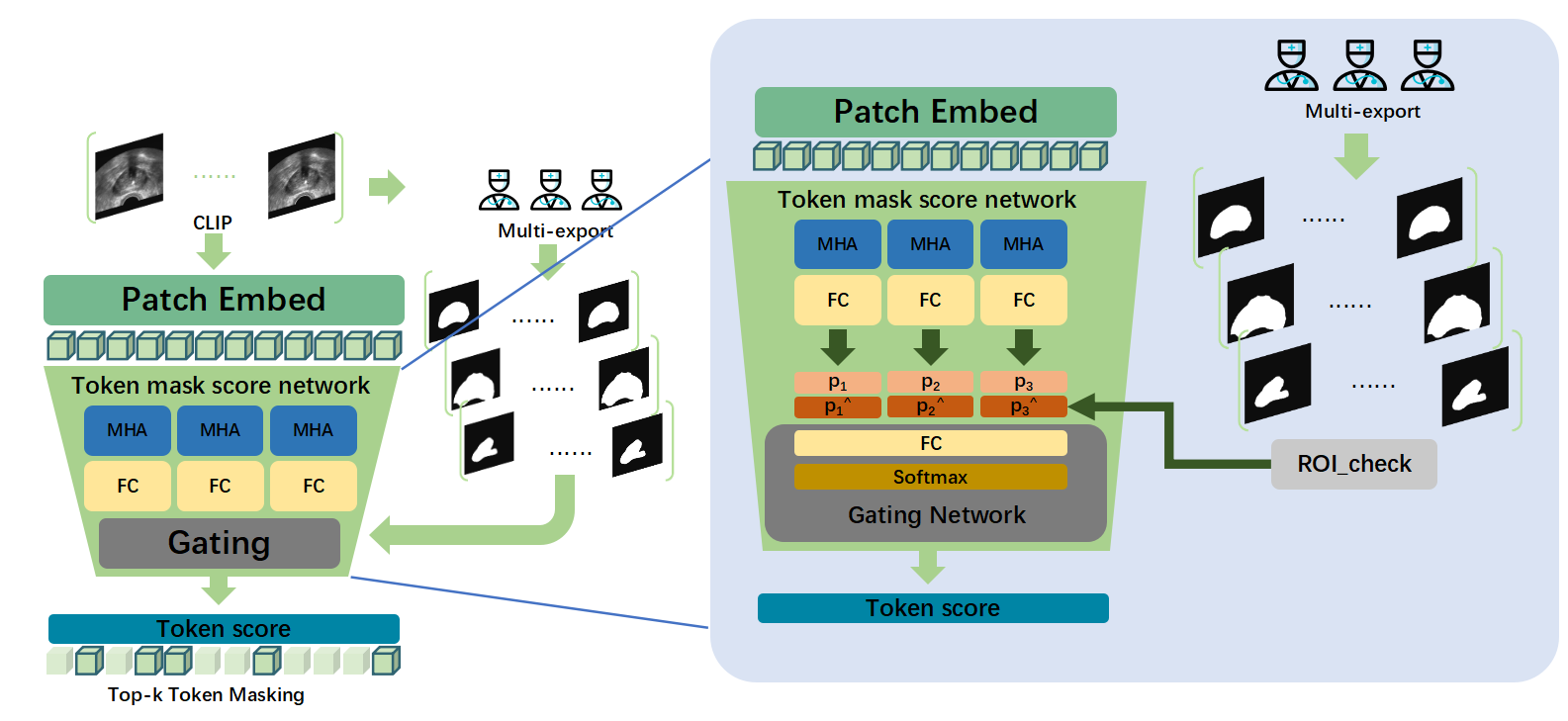


Figure 2.Detailed structure of the proposed MeMGN , including four steps : (a) The input cubic tokens perform Multi-head attention (MHA) operations at different tokens to obtain the basic token scores.(b) Raw images of input clip get segments from pretrained export models.(c) Some of the basic token scores are increased due to the segments provided by export models.(d) Three different exports’ token scores are fused by Gating Network to obtain the final output token scores.(这里我没有修改，你得先改图，图中的a,b,c,d标注了再对照着写相关的caption描述)

（改图2）

**3.2.2 Multi-exports Mask Generation Network**

As illustrated in Fig. 2, we introduce a multi-exports mask generation network (MeMGN) including three different token mask score network and their export model, effectively capture the importance of the lesion area and the relationship between lesion area and other regions. Specifically , given the input cubic tokens

, the MeMGN initially Make three identical copies of the cubic tokens for multi-exports network :

(5)

Where represents the copy operation and is the tokens for the i-th export , whereranges form 1 to 3. Inspired by the Masking probability network of [1] , we construct a muti-exports guided mask generation network that combines the priori knowledge from pretrained models instead of training a segment model with large-scale annotated images. For tokens , we use the i-th token mask score network to conduct the basic token score by performing multi-head attention , followed by fully connected layer. ​The network aggregates contextual information from the video and determines the importance between tokens, thereby guiding the subsequent masking process and compelling the model to learn representations of key high-scoring regions.It can be formulated as:

(6)

Where represents a series of multi-head attention blocks.Note that the multi-head attention’s output is the same shape as input .Then the output is fed into the fully connected layer to calculate the basic token score:

(7)

Where denotes the fully connected layer. Outcomes .Afterward , the raw images of the clip are collected to generate the priori knowledge by pretrained export models. Given the raw images , each export model give out The binarized result represents the segmentation result. To meet the token scores’ format of the cubic tokens, we perform patch processing on the binarized segmentation results. The overall steps of the patch processing can be described as :

(8)

(9)

Where denotes The set of elements at corresponding positions in the binarized segmentation result, checks whether all elements in a block contain 1, which corresponds to the segmented area. If so, it outputs 1, effectively performing an erosion operation on the segmentation region. .Then, we flatten all the values to obtain the final a priori knowledge proposal  *.*

Region priors then boost the probability score as follows:

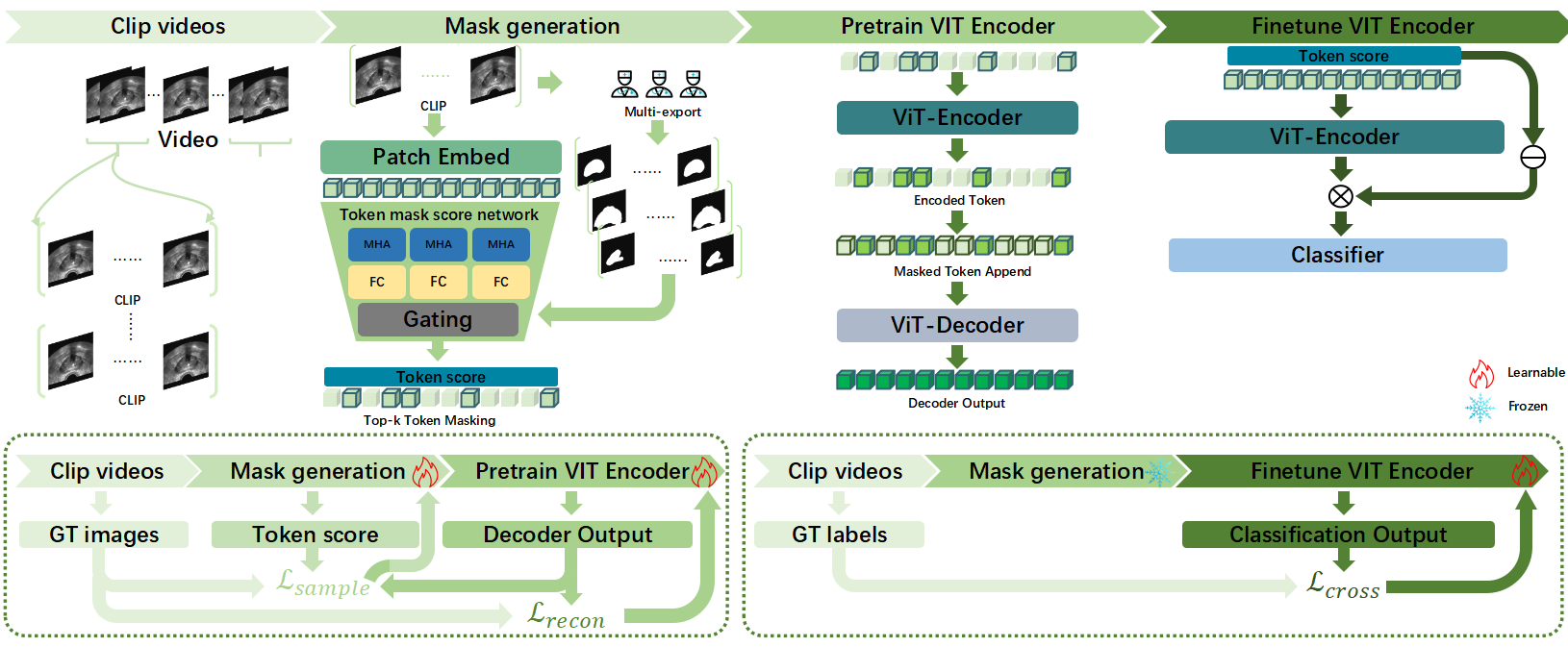
(10)

Then if the i-th token spatially lies within the candidate regions, then we inflate the score of the token by , where is a small tensor less than 2. Inspired by the multi-expert mechanism and the multi-expert consultations in medicine, we choose to fuse the scores of three export network through a learnable gating mechanism using a fully connected layer to calculate the weight for each export network rather than using a fixed mechanism such as averaging :

(11)

Then, perform pixel-wise weighted averaging to obtain the final score by follows :

(12)



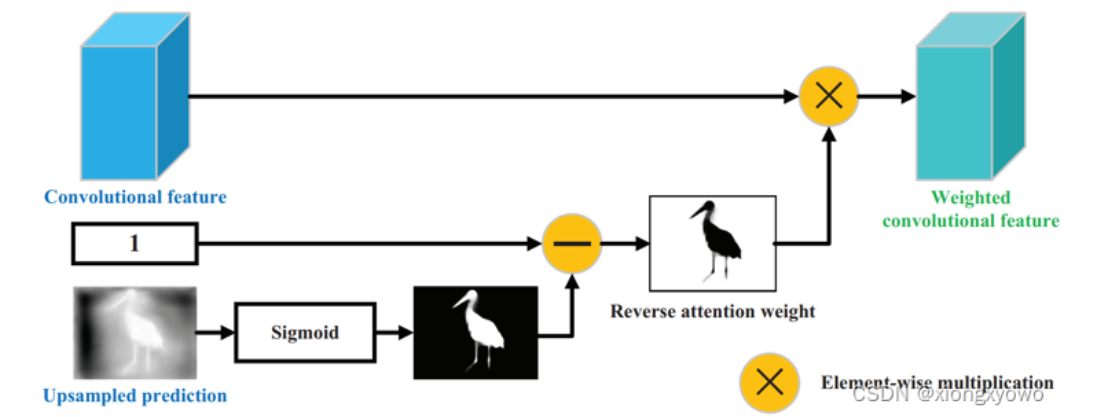


Figure 3.Fine-tuning phase

**3.3 Multi-exports guided fine-tuning**

After pretraining phase, the decoder is abandoned. The pretrained MeMGN and encoder are used in the fine-tuning phase. And this section explains how the fine-tuning phase uses the MeMGN and encoder in short.

**3.3.1 Fine-tuning phase**

Given a video *V* with label *y*. We sample the video as pre-training phase and get *L* clips *C.* The input of fine-tuning phase is also represented as *C ,* where denote the length of a clip, the height and the width of frames.

The clip *C* first is embedded to obtain the cubic token as the pretraining phase. Then, the token along with the Segments *S* are jointly input to the pre-trained MeMGN to generate the mask scores .

Afterwards the embedded tokens are transmitted to the pretrained encoder to obtain the basic feature , which can be described as follows:

(13)

Where represent the pretrained encoder blocks. With the guidance of the mask scores , the final feature of outcome comes from the proposed Multi-exports guided attention (MeGA) operation:

(14)

Where represents the Multi-exports guided attention operation. Finally the output of the final feature is used to make the classification result with a classification head composed of a fully connected layer and a softmax operation, which is formulated as:

(15)

where and represent fully connected layer and softmax operation.

**3.3.2 Multi-exports guided attention**

The Vision Transformer based network can provide reliable features, but fine-tuning for classification tasks, especially when the dataset is smaller (compared to the pre-training phase), tends to lead to overfitting. Feature extracted by the vision transformer encoder become more global as the depth increases , while the shallow features generalizable features and can better represent the textural and structural patterns of the image. Leveraging this property, inspired by[6], we perform a reverse attention map multiplication operation using the mask scores generated by MeMGN to enhance generalizable features.

As shown in Fig.3, the input mask scores generated by pretrained MeMGN is are first sign-flipped and then passed through the sigmoid function , yielding a normalized mask probability . Subsequently, the computed mask probability and the basic feature extracted by pretrained encoder perform an element-wise multiplication, which can be summarized as:

(16)

(17)

where denotes the feature of all the cubic tokens. In the end , the final feature is generated by averaging across all token dimensions to create a global feature guidance, representing the overall information of the input image, rather than using a single CLS token like vision transformer.

**4.Datasets**（说明：红色是你的文字，蓝色是我进行修改的）

The training data used in this study are mainly collected from Shanghai Tenth People’s Hospital affiliated with Tongji University, consisting of 652 ultrasound video cases with a total of 89,355 frames. Among them, 338 cases are suspected of having lesions and 314 cases are identified as T1 stage, resulting in a relatively balanced distribution. To evaluate the generalization ability of the proposed model, an external validation set is constructed using 182 non-invasive prostate ultrasound videos collected from three additional medical centers: Bengbu People's Hospital, Ningbo No. 2 Hospital, and Fudan Zhongshan Hospital. The dataset is of large scale and high quality, covering multiple hospitals and a variety of ultrasound equipment, thereby providing a robust foundation for model training and evaluation.

To remove patient privacy information and other irrelevant elements such as watermarks, we employ the open-source tool LabelImg to annotate and crop the regions of interest. Furthermore, to meet the resolution requirements of the input data, all frames from the same video case are uniformly resized. Specifically, the longer side of each frame is scaled to 512 pixels, while the shorter side is padded with black borders to achieve a final resolution of 512×512 pixels. As a result, we obtain a total of 89,355 frames for training and internal validation, and 42,691 frames for multi-center external validation.

As described in Section 3.2 and 3.3 , We applied three pretrained medical segmentation models as our multi-experts guidance to guide the generation of mask in the pretraining phase and the feature extraction in the fine-tuning phase. We first utilize the Medical Segment Anything 2[2] to generate segment for the guidance. However, using ChatGPT[3] to generate the role prompt for MedSAM2. The overall generation process of the segment for the first export can bu summarized as follows:

(9)

Where represents the Medical Segment Anything 2 model. and are the raw image of ultrasound video and role prompt given by ChatGPT. For the other two expert models, we chose medical segmentation models nnUNet[4] and deeplab-v3[5] with pretrained weights which comes from ultrasound imaging of regions such as the gallbladder, breast, and others. The process by which these two pretrained experts provide segmentation guidance can be described as:

(10)

(11)

Where and represent the nnUNet2 model and Deeplabv3 model.

Finally, we saved the segments *S* that given by our chosen pretrained expert models as images for pre-training and fine-tuning phases.

**5.Experiments and analysis**

**5.1 Experimental Setups**

3.4.1 Multi-exports Mask Generation Network

To jointly constrain the MeMGN network and the visual representation model, we design sampling loss, smoothness loss, and reconstruction loss.

**Masking Reconstruction Loss.** We adopt the Mean Squared Error (MSE) loss to constrain the reconstructed values and the ground truth, which is formulated as follows:

, (4)

where and denote the predicted token and the ground-truth values of the token.represent the number of masked tokens.

**Smoothness Loss.** To mitigate block artifacts caused by overfitting or discontinuities in the image, we introduce a smoothness loss. Specifically, we apply smoothness constraints in both the vertical and horizontal directions to reduce abrupt changes between neighboring pixels, encouraging the generation of more natural and coherent images. The loss is formulated as follows:

(5)

(6)

**Token Sampling Loss.** To maximize sampling accuracy, inspired by AdaMAE, we designed a sampling loss. This loss applies stronger constraints on high-information regions prone to errors while relaxing constraints in low-information areas, encouraging the network to assign higher probability scores to high-information regions. The loss function is defined as follows:

, (8)

where denote the probability of the token to be masked. In the pre-training

**Reference**

1. Basu S, Gupta M, Madan C, et al. Focusmae: Gallbladder cancer detection from ultrasound videos with focused masked autoencoders[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 11715-11725.
2. Zhu J, Hamdi A, Qi Y, et al. Medical sam 2: Segment medical images as video via segment anything model 2[J]. arXiv preprint arXiv:2408.00874, 2024.
3. Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report[J]. arXiv preprint arXiv:2303.08774, 2023.
4. Isensee F, Jaeger P F, Kohl S A A, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation[J]. Nature methods, 2021, 18(2): 203-211.
5. Yurtkulu S C, Şahin Y H, Unal G. Semantic segmentation with extended DeepLabv3 architecture[C]//2019 27th signal processing and communications applications conference (SIU). IEEE, 2019: 1-4.
6. Chen S, Tan X, Wang B, et al. Reverse attention for salient object detection[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 234-250.
7. Tong Z, Song Y, Wang J, et al. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training[J]. Advances in neural information processing systems, 2022, 35: 10078-10093.