Medical Language Mixture of Experts for Improving Medical Image Segmentation

Yaru Liu
School of Artificial Intelligence, Beijing University of Posts and Telecommunications
Beijing, China
liuyaru@bupt.edu.cn

Jiangbo Pei
School of Artificial Intelligence, Beijing University of Posts and Telecommunications
Beijing, China
jiangbop@bupt.edu.cn

Zhu He
School of Artificial Intelligence, Beijing University of Posts and Telecommunications
Beijing, China
hezhu@bupt.edu.cn

Guangjing Yang
School of Artificial Intelligence, Beijing University of Posts and Telecommunications
Beijing, China
ygj2018@bupt.edu.cn

Zhuqing Jiang*
School of Artificial Intelligence, Beijing University of Posts and Telecommunications
Beijing, China
jiangzhuqing@bupt.edu.cn

Qicheng Lao*
School of Artificial Intelligence, Beijing University of Posts and Telecommunications
Beijing, China
Shanghai Artificial Intelligence Laboratory
Shanghai, China
qicheng.lao@bupt.edu.cn

2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10822440

***Abstract*—Traditional medical image segmentation methods are mostly uni-modal approaches solely based on the image modality. Recently, the emergence of text-guided image segmentation methods, by utilizing text annotations to compensate for the quality deficiency in image data, has shown promise for improving medical image segmentation. Despite their success, these methods often experience inadequate utilization of beneficial text information, and have applicability issues in the missing text modality scenario. To address these limitations, in this paper, we propose a Medical Language Mixture of Experts (MLMoE), which introduces multiple sub-experts for extracting more diverse information from medical text. These different experts are then combined by a gating module, thus aggregating beneficial text information to assist the image segmentation. Furthermore, to guarantee its performance in the text-absent scenario, a virtual prompt based distillation module is proposed, which distills the valuable knowledge of MLMoE learned from available text information to the virtual prompt, as an alternative text input. Experimental results on two multi-modal medical segmentation datasets demonstrate the effectiveness of our opposed method, achieving state-of-the-art performance. Code will be available at: https://github.com/Rango-bit/MLMoE.git.**

***Index Terms*—Mixture of experts, Virtual prompt, Knowledge distillation, Medical image segmentation**

I. Introduction

Medical image segmentation is one of the most critical tasks in the field of medical image analysis. Traditional medical image segmentation methods [1]–[5] are mostly uni-modal approaches that rely solely on the image modality. Recently, the emergence of multi-modal pretraining approaches such as CLIP [6], ALIGN [7] and BriVL [8] have directed attention toward the significance of the text. Subsequently, a substantial amount of research [9]–[13] has emerged for text-guided image segmentation, utilizing information from medical reports to compensate for the quality deficiency in image data. In the field of medical image segmentation, Li et al. [12] introduced a language-driven image segmentation method LViT, where a hybrid CNN-Transformer architecture is designed to fuse image and text information through an early fusion mechanism. Zhong et al. [11] proposed utilizing independent image and text encoders to improve feature representations, which are fused at the mask decoding stage.

Despite their success, most of these methods often implicitly assume all information in the text modality is equally important and beneficial for medical image segmentation, or they require manual selection of important text information. However, we argue that this assumption is not always valid due to the significant gap between professional medical reports and the text encoder pretrained on non-medical data, and meanwhile, manual selection of text information is often prone to erroneous bias. Therefore, directly fusing undesirable text information with image information may cause unnecessary confusion and cannot fully maximize the enhancement effect. On the other hand, text information serves as an indispensable input in existing methods, presupposing its availability even during the inference phase. This limitation significantly restricts their applicability in the real world where the text modality is often missing in the test time.

This paper aims to address the aforementioned challenges. To fully leverage the potential of text modality for medical image segmentation, we propose a novel Medical Language Mixture of Experts (MLMoE). Inspired by the mixture of experts approaches [14]–[17], MLMoE introduces multiple sub-experts in text encoding to extract more diverse informa-

* Co-corresponding authors.

tion from medical reports. Subsequently, a gating module is proposed to aggregate beneficial text information from these experts to assist the image segmentation. This also eliminates the erroneous bias introduced by the manual text selection. Furthermore, to ensure model inference and maintain the model performance in the text-absent scenario during the test phase, we introduce a virtual prompt based distillation module, which distills the valuable knowledge of MLMoE learned from available text modality to a virtual prompt [18]. Our designed virtual prompt serves as an alternative text input, enabling model inference when text is unavailable, thus guaranteeing its performance in the text-absent scenario. Our main contributions are summarized as follows:

- We propose an MLMoE module that can selectively and efficiently mine a variety of specialized text information by automatically investigating the impact of information granularity in medical text.
- We introduce a virtual prompt based distillation module that distills the knowledge learned from the MLMoE module to virtual prompt and enables model inference in the text-absent scenario during the test time.
- Extensive experiments on multi-modality medical datasets demonstrate that our method achieves the best segmentation performance in both text-present and text-absent scenarios, compared to other state-of-the-art methods.

II. Method

The overview of our proposed method is shown in Fig. 1, which mainly introduces an MLMoE module for text-guided image segmentation to selectively and efficiently mine a variety of specialized text information (Section II-B); and a virtual prompt based distillation module to distill the valuable knowledge of MLMoE learned from available text information to a virtual prompt in the scenario where the text modality is not available (Section II-C).

A. Preliminary: Language Guided Medical Image Segmentation

In current prevailing approaches for text-guided medical image segmentation, the text information provides additional knowledge for improving the quality of segmentation masks. Given a medical image x_v and medical report x_t, they are first encoded into visual and text features: $f_v = \text{Encoder}_v(x_v)$, $f_t = \text{Encoder}_t(x_t)$. Then during the decoding phase, visual and text features are fused to obtain multi-modal features f_m. For the input visual features f_v, a multi-head self-attention layer is employed to enhance visual features, resulting in the output visual features f'_v. Subsequently, a multi-head cross-attention layer is utilized to fuse visual features f'_v and text features f_t to get multi-modal features f_m. The whole process can be summarized as:

$$\begin{aligned} f'_v &= f_v + LN(MSA(f_v)), \\ f_m &= f'_v + LN(MCA(\boldsymbol{f_t}, f'_v)), \end{aligned} \tag{1}$$

where $MSA(\cdot)$ denotes the multi-head self-attention layer, $MCA(\cdot)$ denotes the multi-head cross-attention layer, $LN(\cdot)$ denotes layer normalization. Then, the multi-model features f_m are upsampled and fused with low-level features from the encoding process to obtain the segmentation mask $\hat{y}$.

Despite its advantages, we observe that not all texts in medical reports are equally important and beneficial to visual models while manually selecting text information entails extra effort and is often inaccurate. To tackle these issues, we propose a medical language mixture of experts for efficiently and selectively leveraging text information that is beneficial to medical image segmentation.

B. Medical Language Mixture of Experts (MLMoE)

To facilitate the learning of diverse semantic information from highly specialized medical reports, we introduce MLMoE for processing the input text features. Let f_t be the text features. The MLMoE module projects f_t into distinct semantic spaces f_t^1, f_t^2, ..., f_t^N, where each projection represents an individual expert, enabling the learning of varied semantic information, such as the number, location, and size of medical lesions. The projection process could be formulated as:

$$f_t^i = \sigma(\text{Conv}(f_t \cdot W_t^i)), \quad i = 1, 2, \ldots, N, \tag{2}$$

where W_t^i is a learnable matrix, $\text{Conv}(\cdot)$ denotes the 1×1 convolution layer, and $\sigma(\cdot)$ denotes the ReLU activation function.

After separating the text information based on the expert mechanism, we then focus on selecting suitable text information. In contrast to most MoE based works [14], [17], [19] that often generate the gating score based on a single modality, we propose to incorporate both text and image features as the input of the gating module. Specifically, the gating score g is produced by the following process:

$$g_i = \frac{\exp(s_i/\tau)}{\sum_{j=1}^{N} \exp(s_j/\tau)}, \quad s = \Psi(f_t \oplus f_v), \tag{3}$$

where $\oplus$ denotes the concatenation, Ψ is a linear projection layer that maps the input to a N-dimension vector s, s_i is the i_{th} dimension of vector s, g_i is gating score of the i_{th} expert and τ is a learnable temperature scalar. Then, the outputs of the N experts are linearly combined according to their gating scores, which can be formulated as:

$$f'_t = \textstyle\sum_{i=1}^{N} g_i \cdot f_t^i. \tag{4}$$

Similar to Eq. (1), we fuse the filtered effective text information f'_t and image information f'_v to output the segmentation mask f_m. The formula is updated to:

$$f_m = f'_v + LN(MCA(\boldsymbol{f'_t}, f'_v)). \tag{5}$$

To avoid the imbalance problem in MoE where only a small subset of experts are activated [20], we design a diversity loss (inspired by some clustering works [21], [22]) to increase the diversity of the expert weights: $L_{div} = \mathbb{E}_x(\mathcal{H}(g)) - \mathcal{H}(\mathbb{E}_x(g))$, where $\mathcal{H}$ denotes the entropy function and $\mathbb{E}$ denotes the

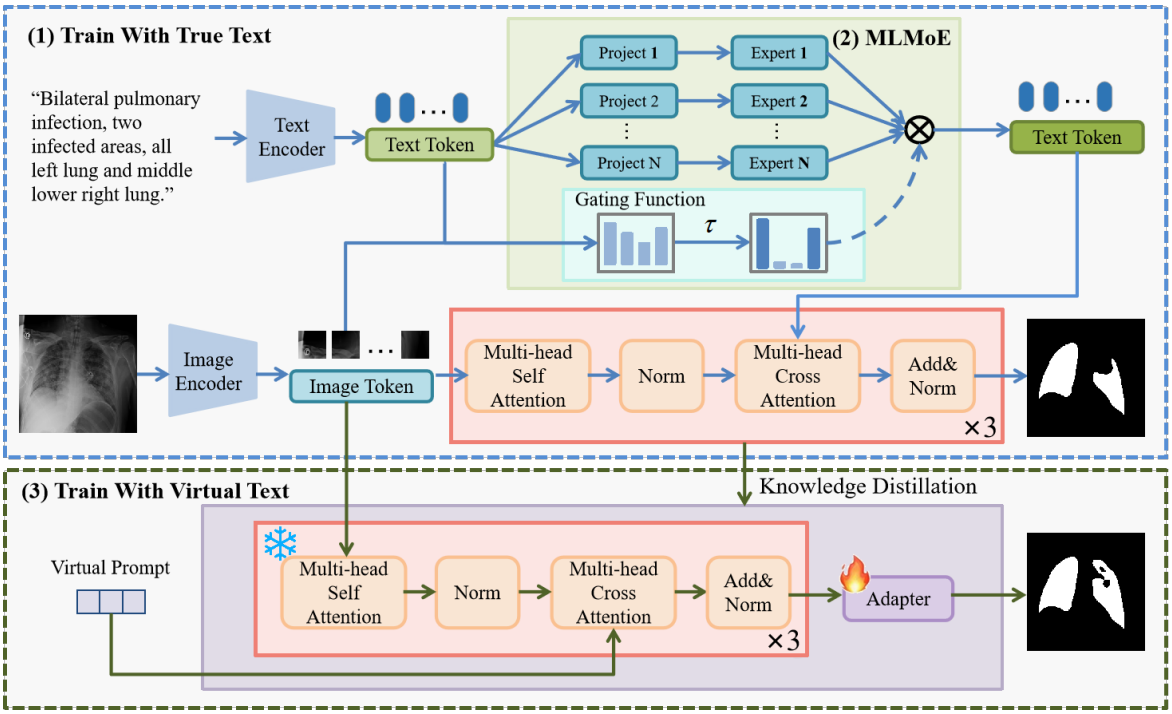

Fig. 1. The overview of our proposed method. (1) The training process in the text-present scenario. (2) Illustration of our MLMoE. (3) The training process in the text-absent scenario. The knowledge distillation operation transfers the knowledge learned from the MLMoE to virtual prompt.

expectation function. The first term $\mathbb{E}_x(\mathcal{H}(g))$ minimizes the entropy of the gating score on each sample x, encouraging the gating score to concentrate on a few experts for each sample. The second term $-\mathcal{H}(\mathbb{E}_x(g))$ encourages the diversity of the gate weights regarding all samples, which promotes the assignment of different experts to different samples. The loss function (L) for training in the text-present scenario is the sum of the dice loss (L_{dice}), the cross-entropy loss (L_{ce}), and the diversity loss (L_{div}), formulated by:

$$L = L_{dice} + L_{ce} + L_{div}. \tag{6}$$

C. Distilling MLMoE with Virtual Prompt and Adapter

In practice, medical multi-modality data is often imperfect due to lacking of text modality [18], [23]. In the missing-modality scenario, current language-guided medical image segmentation methods fail to work effectively due to the lack of indispensable text input. To maintain model performance in the text-absent scenario, we propose to distill the text knowledge acquired via our proposed MLMoE in the text-present scenario into those without text. To achieve this, we consider a learnable virtual prompt as the text input which mitigates the performance drop caused by missing modality. The virtual prompt is a randomly initialized and learnable vector with dimensions of 24×768. It is designed to capture useful text information distilled from MLMoE and learn a generalized text representation. This is inspired by the findings showing that prompts are good indicators for simulating different distributions of input [24]. Specifically, based on the virtual prompt, we can derive a virtual fused feature $\hat{f}_m$ by replacing the f_t in Eq. (1) with the virtual prompt. During the distillation process [25], the true fused feature f_m, which effectively integrates text and image information based on our MLMoE, is designated as the teacher. The virtual fused feature $\hat{f}_m$ is referred to as the student. The distillation loss is designed to minimize the distance between $\hat{f}_m$ and f_m, which can be formulated by:

$$L_{dis} = \mathbb{E}_x(\|f_m - \hat{f}_m\|_2), \tag{7}$$

where the term $\|\cdot\|_2$ measures the L_2 distance between $\hat{f}_m$ and f_m. $\mathbb{E}$ is the expectation function, indicating distillation application to all samples. The loss function (L_A) for training in the text-absent scenario is the sum of the dice loss (L_{dice}), the cross-entropy loss (L_{ce}) and the distillation loss (L_{dis}), formulated by:

$$L_A = L_{dice} + L_{ce} + L_{dis}. \tag{8}$$

Furthermore, to reduce training volume, we specifically employ an adapter module [26] within the decoder (only for the text-absent scenario). During distillation, only the adapter's parameters are updated, with all other parameters in both the encoder and decoder remaining fixed. The adapter module consists of a bottleneck structure and a residual connection, executing the transformation for $\hat{f}_m$ as follows: $\hat{f}_m = \hat{f}_m + \Phi_u(\Phi_d(\hat{f}_m))$. Here, Φ represents the projection operator defined by $\Phi(\hat{f}_m) = \hat{f}_m \cdot w + b$, where Φ_d and Φ_u represent the downsampling projection and the upsampling projection, respectively.

TABLE I
PERFORMANCE COMPARISONS OF OUR METHOD WITH MONO-MODAL AND MULTI-MODAL METHODS ON THE QATA-COV19 DATASET AND MOSMEDDATA+ DATASET. THE COMPARISON OF THE RESULTS OF THE FOLLOWING EXPERIMENTS ARE STATISTICALLY SIGNIFICANT (P<0.001).

Type	Method	Test	QaTa-COV19		MosMedData+	
		w/o text	mIoU (%)	Dice (%)	mIoU (%)	Dice (%)
Mono-Modal	U-Net [1]	-	71.71	83.52	50.76	67.34
	U-Net++ [2]	-	71.91	83.66	51.75	68.20
	AttentionUnet [4]	-	69.67	82.13	51.44	67.93
	Swin UNETR [3]	-	67.25	80.42	40.63	57.79
	nnU-Net [27]	-	72.35	84.39	52.53	69.30
Multi-Modal	LViT [12]*	✓	66.94	80.20	49.23	65.98
	Zhong et al. [11] *	✓	66.56	79.92	61.21	75.94
	Ours (w/o text)	✓	**78.04±0.09**	**87.66±0.10**	**62.57±0.11**	**76.98±0.13**
	LViT [12]	✗	77.21	87.14	61.70	76.32
	Zhong et al. [11]	✗	82.66	90.24	61.97	76.52
	Ours	✗	**83.81±0.09**	**91.19±0.14**	**63.90±0.17**	**77.98±0.23**

* denotes using the class label 'pneumonia' as text input during the inference.

III. EXPERIMENTS

A. Setup

1) Datasets: We use QaTa-COV19 dataset [28] and MosMedData+ dataset [29], [30] to evaluate the performance of our method. The QaTa-COV19 dataset consists of 9258 COVID-19 chest X-ray radiographs with text annotations provided by [12]. The MosMedData+ dataset contains 2729 CT scan slices of lung infections and text annotations, with a similar text structure as the QaTa-COV19 dataset.

2) Implementation details: We train the models using AdamW optimization with a batch size of 32 for the QaTa-COV19 dataset and 24 for the MosMedData+ dataset. The initial learning rate is set to 3e-3 for the QaTa-COV19 dataset and 8e-3 for the MosMedData+ dataset. We use the early stop mechanism if the model performance does not improve for 20 epochs. Following previous work [11], we split the original train set into the train (80%) and validation (20%) sets. All images are cropped to 224 $\times$ 224 pixels. For metrics, we use two standard image segmentation metrics: the Dice score and the mIoU metric.

B. Comparison Experiments

1) Our proposed method achieves the best performance in both text-present and text-absent evaluations: We compare our results with five Mono-Modal baselines including U-Net [1], U-Net++ [2], Attention Unet [4], Swin UNETR [3] and nnU-Net [27] and two Multi-Modal baselines including LViT [12] and Zhong et al. [11]. The quantitative results are presented in Table I. Note that our *text-absent evaluation* means that the text modality is missing during the inference (denoted as 'Test w/o text' in the table), and for evaluating multi-modal methods that do require text, we use the class label 'pneumonia' as their text input in text-absent evaluations. In text-present evaluations, our experimental results on the MosMedData+ dataset show that our proposed method achieves better performance than the previous state-of-the-art methods including mono-modal and multi-model methods. Specifically, our method improves

TABLE II
ABLATION STUDY OF OUR METHOD ON THE QATA-COV19 DATASET IN BOTH TEXT-PRESENT AND TEXT-ABSENT SCENARIOS.

Scenario	Method	mIoU(%)	Dice(%)
Testing with Text	w/o MLMoE	82.66	90.24
	w/o τ	83.35	90.92
	w/o f_v	83.28	90.88
	Ours	**83.81**	**91.19**
Testing w/o Text	w/o adapter	76.93	86.96
	w/o distillation	77.52	87.33
	Ours	**78.04**	**87.66**

TABLE III
ABLATION STUDY ON THE NUMBER OF EXPERTS USING THE QATA-COV19 DATASET.

Number	mIoU(%)	Dice(%)
1	82.66	90.24
2	83.64	91.06
4	**83.81**	**91.19**
8	83.55	90.98
16	83.49	90.88

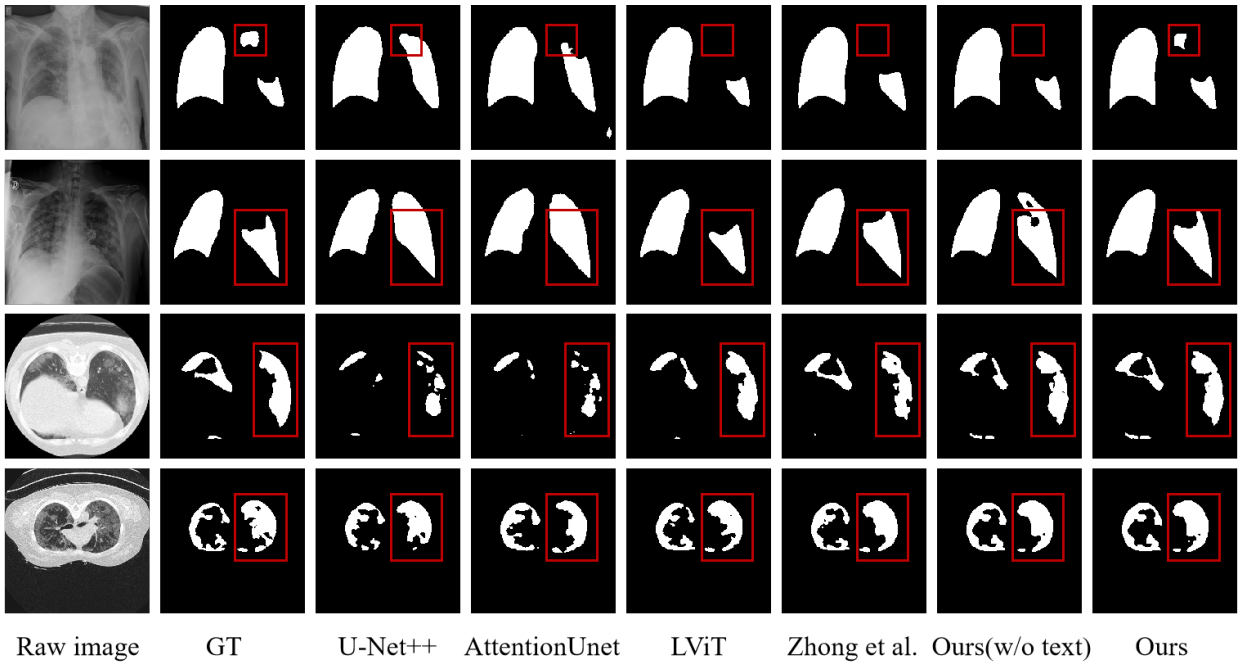

Fig. 2. Representative of qualitative results on the QaTa-COV19 (top) and MosMedData+ (bottom) datasets.

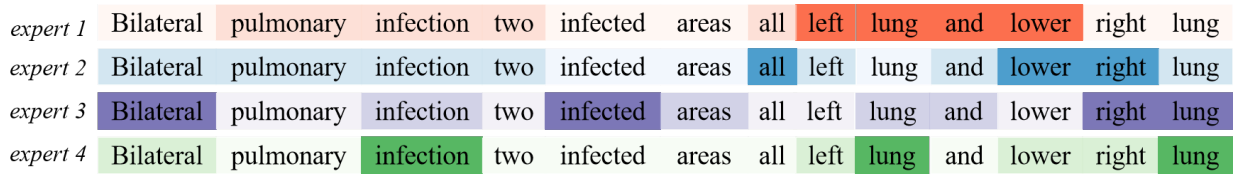

Fig. 3. Visualizing the text information learned by our method. The shades of color represent the degree of attention. Darker shades indicate a higher level of attention.

TABLE IV
ABLATION STUDY OF DIFFERENT PROMPTS IN THE TEXT-ABSENT SCENARIO.

Prompt	QaTa-COV19		MosMedData+	
	mIoU (%)	Dice (%)	mIoU (%)	Dice (%)
pneumonia	77.36	87.23	62.35	76.81
a photo of infected areas	77.95	87.61	62.41	76.86
a photo of lung	77.90	87.58	62.07	76.60
virtual prompt	**78.04**	**87.66**	**62.57**	**76.98**

the Dice score by 1.46% and the mIoU score by 1.93% compared to the suboptimal model. A similar trend is observed for the QaTa-COV19 dataset. In text-absent evaluations, our proposed method achieves an 87.66% Dice score and 78.04% mIoU score on the QaTa-COV19 dataset, and 76.98% Dice score and 62.57% mIoU score on the MosMedData+ dataset. Notably, experimental results on the MosMedData+ dataset show that our method (denoted as 'Ours (w/o text)') surpasses the performance of the previous multi-model methods, nearly equivalent to our method in the text-present scenario during the inference stage.

We present the qualitative results of our method and other state-of-the-art methods in Fig. 2. As shown in the Fig. 2, compared to previous mono-modal methods and multi-modal methods, our proposed method generates more accurate segmentation results where the mis-segmentation errors are greatly reduced in the text-present scenario. In addition, the segmentation results of our method in the text-absent scenario closely approximate, and occasionally surpass the performance of previous multi-modal models with the text modality.

2) Visualization of experts: To fully understand the text information attended to by different experts, we visualize the text information learned by various medical language experts on the QaTa-COV19 dataset. Fig. 3 presents the text activation maps from different experts within our proposed MLMoE. It can be observed that different experts learn diverse semantics.

For instance, expert 1 pays closer attention to positional information, while expert 4 focuses more on information related to diseases.

C. Ablation Study

Here we conduct ablation study to evaluate the effectiveness of different components proposed in our approach, including the MLMoE, the temperature scalar of the gating function, and whether to integrate the vision features in the gating in the text-present scenario, as well as the distillation module and adapter used in the text-absent scenario on the QaTa-COV19 dataset. The results are organized in Table II. Note that we use 'w/o τ' to indicate the MLMoE module without temperature scalar τ, and 'w/o f_v' indicates the gating function without visual features in the MLMoE module. The results demonstrate that our choices are optimal. For expert numbers, we conduct ablation using different numbers of experts as shown in Table III. As can be seen from the Table III, the best performance is achieved when the number of experts is 4. Specially, when using only one expert, text features are mapped to a single feature space, and fused with visual features, which is lower than multiple experts we used.

To substantiate the effectiveness of our proposed virtual prompt in the text-absent scenario, we employ disease names (pneumonia) and general text descriptions (e.g., a photo of infected areas) for comparative analysis of the results. The results in Table IV show that using the virtual prompt eliminates the need to select input text based on images or lesions and ensures the best performance.

D. Model Parameters and Inference Time

Our model has 36.4M trainable parameters, slightly more than Zhong et al. [11] (32.7M), on which ours is based. However, this marginal increase enables the model to excel in the text-absent scenario, resulting in a performance boost (e.g., 7.46% Dice on the QaTa-COV19). Moreover, our model maintains competitive inference speeds. The inference time of our model is 7.72ms in the text-present scenario and 6.71ms in the text-absent scenario, which is comparable to that of Zhong et al. (6.78ms).

IV. Conclusion

In this paper, we propose a medical language mixture of experts (MLMoE) module that enhances the support of medical text for image segmentation. In addition, we introduce a virtual prompt based distillation module to ensure model inference and maintain the model performance in the text-absent scenario during the test time. Experiments demonstrate that our approach achieves state-of-the-art performance on both QaTa-COV19 dataset and MosMedData+ dataset in text-present and text-absent scenarios. In future work, we will focus on applying the MLMoE module to large medical foundation models for segmentation.

References

[1] O. Ronneberger, P. Fischer, and T. 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*. Springer, 2015, pp. 234–241.

[2] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[3] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI Brainlesion Workshop*. Springer, 2021, pp. 272–284.

[4] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[5] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[7] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.

[8] Y. Huo, M. Zhang, G. Liu, H. Lu, Y. Gao, G. Yang, J. Wen, H. Zhang, B. Xu, W. Zheng *et al.*, "Wenlan: Bridging vision and language by large-scale multi-modal pre-training," *arXiv preprint arXiv:2103.06561*, 2021.

[9] N. K. Tomar, D. Jha, U. Bagci, and S. Ali, "Tganet: Text-guided attention for improved polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 151–160.

[10] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu, "Denseclip: Language-guided dense prediction with context-aware prompting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 082–18 091.

[11] Y. Zhong, M. Xu, K. Liang, K. Chen, and M. Wu, "Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest x-ray images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 724–733.

[12] Z. Li, Y. Li, Q. Li, P. Wang, D. Guo, L. Lu, D. Jin, Y. Zhang, and Q. Hong, "Lvit: language meets vision transformer in medical image segmentation," *IEEE transactions on medical imaging*, 2023.

[13] Z. Wang, Y. Lu, Q. Li, X. Tao, Y. Guo, M. Gong, and T. Liu, "Cris: Clip-driven referring image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 686–11 695.

[14] B. Mustafa, C. Riquelme, J. Puigcerver, R. Jenatton, and N. Houlsby, "Multimodal contrastive learning with limoe: the language-image mixture of experts," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9564–9576, 2022.

[15] S. Shen, Z. Yao, C. Li, T. Darrell, K. Keutzer, and Y. He, "Scaling vision-language models with sparse mixture of experts," *arXiv preprint arXiv:2303.07226*, 2023.

[16] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 5232–5270, 2022.

[17] H. Akbari, D. Kondratyuk, Y. Cui, R. Hornung, H. Wang, and H. Adam, "Alternating gradient descent and mixture-of-experts for integrated multimodal perception," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[18] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multimodal transformers robust to missing modality?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 177–18 186.

[19] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat *et al.*, "Glam: Efficient scaling of

language models with mixture-of-experts," in *International Conference on Machine Learning*. PMLR, 2022, pp. 5547–5569.
[20] S. Pavlitska, C. Hubschneider, L. Struppek, and J. M. Zöllner, "Sparsely-gated mixture-of-expert layers for cnn interpretability," in *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2023, pp. 1–10.
[21] K. Ghasedi Dizaji, A. Herandi, C. Deng, W. Cai, and H. Huang, "Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5736–5745.
[22] J. Pei, A. Men, Y. Liu, X. Zhuang, and Q. Chen, "Evidential multi-source-free unsupervised domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
[23] Y.-L. Lee, Y.-H. Tsai, W.-C. Chiu, and C.-Y. Lee, "Multimodal prompting with missing modalities for visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 943–14 952.
[24] K. Saito, K. Sohn, X. Zhang, C.-L. Li, C.-Y. Lee, K. Saenko, and T. Pfister, "Prefix conditioning unifies language and label supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2861–2870.
[25] S. Jung, D. Lee, T. Park, and T. Moon, "Fair feature distillation for visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 115–12 124.
[26] G. Luo, M. Huang, Y. Zhou, X. Sun, G. Jiang, Z. Wang, and R. Ji, "Towards efficient visual adaption via structural re-parameterization," *arXiv preprint arXiv:2302.08106*, 2023.
[27] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.
[28] A. Degerli, S. Kiranyaz, M. E. Chowdhury, and M. Gabbouj, "Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 2306–2310.
[29] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," *European Radiology Experimental*, vol. 4, no. 1, pp. 1–13, 2020.
[30] S. P. Morozov, A. Andreychenko, N. Pavlov, A. Vladzymyrskyy, N. Ledikhova, V. Gombolevskiy, I. A. Blokhin, P. Gelezhe, A. Gonchar, and V. Y. Chernina, "Mosmeddata: Chest ct scans with covid-19 related findings dataset," *arXiv preprint arXiv:2005.06465*, 2020.