ABP: Asymmetric Bilateral Prompting for Text-guided Medical Image Segmentation

Xinyi Zeng[1], Pinxian Zeng[1], Jiaqi Cui[1], Aibing Li[1], Bo Liu[2], Chengdi Wang[3], Yan Wang[1,✉]

[1] School of Computer Science, Sichuan University, China
wangyanscu@hotmail.com
[2] Department of Computing, The Hong Kong Polytechnic University, China
[3] Department of Respiratory and Critical Care Medicine, West China Hospital, China

Abstract. Deep learning-based segmentation models have made remarkable progress in aiding pulmonary disease diagnosis by segmenting lung lesion areas in large amounts of annotated X-ray images. Recently, to alleviate the demand for medical image data and further improve segmentation performance, various studies have extended mono-modal models to incorporate additional modalities, such as diagnostic textual notes. Despite the prevalent utilization of cross-attention mechanisms or their variants to model interactions between visual and textual features, current text-guided medical image segmentation approaches still face limitations. These include a lack of adaptive adjustments for text tokens to accommodate variations in image contexts, as well as a deficiency in exploring and utilizing text-prior information. To mitigate these limitations, we propose Asymmetric Bilateral Prompting (ABP), a novel method tailored for text-guided medical image segmentation. Specifically, we introduce an ABP block preceding each up-sample stage in the image decoder. This block first integrates a symmetric bilateral cross-attention module for both textual and visual branches to model preliminary multi-modal interactions. Then, guided by the opposite modality, two asymmetric operations are employed for further modality-specific refinement. Notably, we utilize attention scores from the image branch as attentiveness rankings to prune and remove redundant text tokens, ensuring that the image features are progressively interacted with more attentive text tokens during up-sampling. Asymmetrically, we integrate attention scores from the text branch as text-prior information to enhance visual representations and target predictions in the visual branch. Experimental results on the QaTa-COV19 dataset validate the superiority of our proposed method.

Keywords: Medical Image Segmentation, Text-guidance, Bilateral Cross Attention, Asymmetric Prompting.

1 Introduction

Pulmonary diseases stand as a significant health challenge on a global scale, prompting extensive research into their diagnosis and treatment [1-2]. Within these endeavors, increasing attention has been focused on the diagnostic value of radiological information for pulmonary infectious diseases like COVID-19. Imaging modalities, such as X-rays, have become essential tools in this pursuit. However, traditional approaches to diagnosing pulmonary diseases through X-ray scans strongly rely on manual delineations of organs and lesions, which is quite labor-intensive and time-consuming. With the development of deep learning (DL), there has been a large amount of research leveraging deep neural networks to process radiology images efficiently [3-7]. Within this field, DL-based medical image segmentation, which autonomously generates lesion masks for infectious areas, stands out as a crucial application.

Traditional DL-based medical image segmentation methods primarily utilize convolutional or transformer-based neural networks to directly generate prediction masks for tumor or lesion areas [8-17]. For example, Zhou *et al.* [9] enhanced the UNet [8] architecture by introducing additional up-sampling nodes and skip-connections to mitigate the semantic gap. Chen *et al.* [12] pioneered the utilization of the Transformer to construct a UNet-like architecture for medical image segmentation, aiming to leverage the capabilities of the Transformer in capturing global contexts. However, these mono-modal methods often demand a considerable volume of annotated image data to achieve satisfactory performance. Besides, many medical datasets are accompanied by diagnostic textual descriptions or reports containing crucial information (e.g., descriptions of lesion severity and location), which can guide the segmentation process [18]. Yet, such complementary textual information is often disregarded in these studies.

In recent years, multi-modal learning, especially in vision-language models [19], has garnered widespread attention from researchers for its robust generalization capabilities. As for medical image analysis, some pioneering efforts [20-26] have harnessed complementary textual modalities and have developed various text-guided medical image segmentation models, aiming to reduce the reliance on extensive medical image data and boost segmentation performance beyond image-only approaches. For instance, Li *et al.* [22] first proposed to leverage hybrid CNN and Transformer architecture to fuse medical images with textual information. Lee *et al.* [25] proposed a multi-modal fusion module that combines cross-attention with position attention, aiming to capture the associations between text and image for integration. Zhong *et al*. [26] proposed separate text and image encoders and designed a text-guided decoder to fuse features from both modalities during the decoding stage, achieving superior results.

Among these studies, the predominant approach involves the use of cross-attention or its variants [22-26] to model multi-modal interactions between feature representations (tokens) of radiology images and textual descriptions. This is typically achieved through a dual-branch architecture, which relies on a textual backbone encoder as the text branch to generate static text embeddings, with the expectation that these embeddings will effectively interact with image features through cross-attention modules on the image branch. Despite yielding superior results, these methods still encounter two limitations. Firstly, static text embeddings fail to facilitate the adaptive and progressive tuning of the contained text tokens to accommodate different semantic contexts in varied scales of image features. For instance, as visual features are gradually up-sampled towards the final output in the decoder, it is crucial to retain text tokens that contain more valuable information regarding segmentation targets, while removing tokens

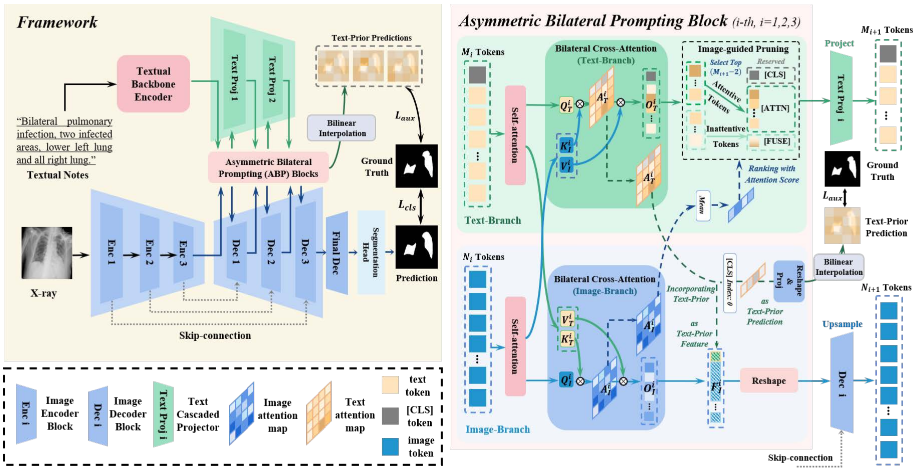

Fig. 1. The overall framework of the proposed ABP framework.

containing irrelevant content. Secondly, the simplistic architectural design of the text branch also prevents text embeddings from generating additional text-prior information beyond merely being sent to the image branch for cross-attention computation. This neglected text-prior information, in fact, contains valuable insights that can enhance image representations and target predictions.

To address the above limitations, we propose Asymmetric Bilateral Prompting (ABP), a novel method tailored for text-guided medical image segmentation. The core of ABP involves incorporating the ABP block preceding each up-sampling scale in the decoder. Compared to unilateral cross-attention adopted in previous dual-branch architectures, ABP exhibits improvements in two aspects. First, it integrates a symmetric bilateral cross-attention module in both image and text branches to explore and model preliminary multi-modal interactions. Second, it introduces two asymmetric operations to utilize information from the opposite branch to further guide modality-specific refinement within its own branch. Our main contributions can be summarized as:

1) We present Asymmetric Bilateral Prompting (ABP), a novel dual-branch framework for text-guided medical image segmentation. Notably, ABP first incorporates symmetric bilateral cross-attention for efficient inter-modality interaction and preliminary mutual refinement.
2) Building upon bilateral cross-attention, ABP then introduces two asymmetric operations. Within the text branch, we utilize attention scores from the image branch as rankings of attentiveness to prune and remove redundant tokens, ensuring that the image up-sampling progressively interacts with more attentive text tokens. Asymmetrically, in the image branch, we incorporate attention scores from the text branch as text-prior information to further enhance feature representations and target predictions.
3) The state-of-the-art performance on the Qata-COVID-19 dataset demonstrates the superiority of our proposed method.

2 Methodology

2.1 Architecture

The overview of the proposed ABP is illustrated in Fig. 1, which comprises an image branch and a text branch. During the encoding stage of the image branch, the X-ray image, denoted as $I \in R^{B\times H\times W\times C}$, is processed through a visual backbone encoder consisting of three encoder blocks (i.e., Enc 1 ~ Enc 3), to generate three encoded features at different scales. Concurrently, in the text branch, the associated textual description undergoes tokenization via a textual backbone encoder, obtaining the initial text tokens $T_1 \in R^{B\times M_1\times C_1}$, where M_1 represents the length of the encoded sequence, and C_1 denotes its channel dimension.

During the decoding stage, we utilize an image decoder consisting of 3 text-guided decoder blocks (i.e., Dec 1 ~ Dec 3) and a final decoder block to upsample the image features with the aid of textual information. Notably, before each guided decoder block, we represent the image features in token form and input both visual tokens and text tokens into a well-designed Asymmetric Bilateral Prompting (ABP) block to facilitate the fusion and interaction of visual and textual information. Assisted by ABP blocks, text tokens undergo specialized image-guided pruning to retain the most relevant and representative information to guide the segmentation, while image tokens benefit from text prior information to enhance its feature representation and target prediction. Upon processing in the ABP block, the text tokens are further projected by cascaded projectors (i.e., $Text\ Proj$ 1 ~ $Text\ Proj$ 2), constructed by multiple multi-layer perceptrons (MLP) hierarchically, to align the channel dimensions of text tokens with those of image tokens in the next scale. Meanwhile, the output image tokens are reshaped back to the original spatial scale. The text-guided decoder block performs joint upsampling with skip-connection features from the encoder by doubling the spatial scale while halving the channel dimensions. For the final decoder block, we do not precede it with an ABP block for guidance, but instead, pass its output through a segmentation head to obtain the final prediction mask for infection areas, denoted as $M \in R^{B\times 1\times H\times W}$.

2.2 Asymmetric Bilateral Prompting (ABP) Block

Before the i-th text-guided decoder block ($i = 1, 2, 3$), we incorporate an ABP block to process the image tokens $I_i \in R^{B\times N_i\times C_i}$ and text tokens $T_i \in R^{B\times M_i\times C_i}$, aiming to fully exploit their complementary information for mutual refinement on both the image and text branches. Here, $N_i = H_i \times W_i$ represents the length of flattened image tokens, and M_i represents the length of text token sequence. Within the ABP Block, we first introduce Symmetric *Bilateral Cross-attention* for the preliminary multi-modal fusion and interactions of visual and textual tokens. Additionally, two asymmetric operations, namely, *Image-guided Pruning* and *Text-prior Incorporation*, are separately applied to the text and image branches to enhance the learning of modality-specific information.

Bilateral Cross-attention. After passing through the self-attention layers, image tokens I_i and text tokens T_i are interacted via Symmetric *Bilateral Cross-attention*. Unlike previous attention-based prompt methods [22-26] that solely refine image features with text prompts using unilateral cross-attention, our bilateral prompt aims to enhance the features of one modality with those of the other modality, which is more conducive to harmonizing the discrepancy between visual and text tokens. Specifically, I_i and T_i are projected into Q_I^i, K_I^i, V_I^i, and Q_T^i, K_T^i, V_T^i by their corresponding projector $Proj_I^i(\cdot)$ and $Proj_T^i(\cdot)$, as formulated below:

$$Q_I^i, K_I^i, V_I^i \in R^{B\times M_i\times C_i} = Proj_I^i(I_i), \tag{1}$$

$$Q_T^i, K_T^i, V_T^i \in R^{B\times N_i\times C_i} = Proj_T^i(T_i). \tag{2}$$

Subsequently, multi-head cross-attention is performed separately on the image and text branches, where the query (Q) is derived from its own branch while the key (K) and value (V) are sourced from the opposite branch. This process is denoted as:

$$A_I^i \in R^{B\times M_i\times N_i} = Softmax\left(Q_I^i \cdot K_T^i/\sqrt{C_i}\right), o_I^i \in R^{B\times N_i\times C_i} = A_I^i \cdot V_T^i, \tag{3}$$

$$A_T^i \in R^{B\times N_i\times M_i} = Softmax\left(Q_T^i \cdot K_I^i/\sqrt{C_i}\right), o_T^i \in R^{B\times M_i\times C_i} = A_T^i \cdot V_I^i, \tag{4}$$

where o_I^i and o_T^i represent the results of cross-attention. Then, they are further fed to a feed-forward network to derive the final output of bilateral cross-attention, denoted as $O_I^i \in R^{B\times N_i\times C_i}$ and $O_T^i \in R^{B\times M_i\times C_i}$. Note that, the generated cross-attention score maps A_I^i and A_T^i are preserved to guide and facilitate subsequent asymmetric operations, further enabling modality-specific refinement in the opposite branch.

Text Branch - Image-guided Pruning. While bilateral cross-attention can effectively model preliminary interactions between multi-modal tokens, this symmetrical form of interaction may not comprehensively capture essential modality-specific information tailored to each modality's characteristics. For example, as visual features are progressively decoded and up-sampled toward the final output, it becomes crucial to retain text tokens containing pertinent information regarding segmentation targets (e.g., words indicating lesion locations) while discarding tokens containing extraneous content (e.g., padding tokens and punctuation used as placeholders) [27]. In the text branch of the ABP block, we tackle this issue by utilizing complementary information (eg. attention score map) from the image branch to aid in the selection and pruning of text tokens.

Specifically, we first apply the average operation to the cross-attention score map $A_I^i \in R^{B\times M_i\times N_i}$ from the image branch along the third dimension, yielding attentiveness values with the size of $B \times M_i$. Corresponding to the dimensions of $T_i \in R^{B\times M_i\times C_i}$, text tokens with higher attentiveness values are considered more relevant to the visual representations, thus expected to play a more significant role in the subsequent decoding and segmentation processes. Therefore, we organize the text tokens in O_T^i by ranking their attentiveness values in descending order, and preserve the top-$(M_{i+1} - 2)$ tokens as attentive tokens, represented as [ATTN]. Note that, the first token in O_T^i, represented as [CLS], is always kept as the class token that directly indicates the segmentation target, and thus remains unaffected in the pruning process. Regarding the remaining tokens, instead of straightforward omission, we choose to utilize their attentiveness values as weights in a weighted average operation to fuse them into a single token, denoted as [FUSE]. This approach ensures the retention of a more comprehensive set of information during the pruning process. Finally, the pruned tokens can be represented as $P_T^i \in R^{B\times M_{i+1}\times C_i} = ([\mathrm{CLS}][\mathrm{ATTN}_1] \dots [\mathrm{ATTN}_{M_{i+1}-2}][\mathrm{FUSE}])$. In this manner, we preserve tokens highly associated with the segmentation target while consolidating relatively irrelevant tokens into a fused one, thus making the pruned token sequence more representative and enhancing segmentation guidance.

After pruning, P_T^i is then processed by a cascaded textual projector ($Text\ Proj\ i$) and projected into $T_{i+1} \in R^{B\times M_{i+1}\times C_{i+1}}$ to serve as the input of the next ABP block. Particularly, for the third ABP block, we omit its pruning process and $Text\ Proj\ 3$, as the obtained textual tokens do not interact with visual features later on.

Image Branch - Text-prior Incorporation. Based on the relatively indirect interactions enabled by bilateral attention between the text and image branches, we also investigate a more direct approach to integrating text-prior information into the image branch, aiming to maximize the utilization of textual prompts in improving image decoding and segmentation. Recognizing that directly concatenating text tokens into image tokens may introduce the potential issue of domain discrepancy [28], we opt to use cross-attention score maps from the text branch as supplementary text-prior information, incorporating them into visual representations and target predictions through two different approaches.

To enhance visual representations, we use the cross-attention score map $A_T^i \in R^{B\times N_i\times M_i}$ from the text branch as *text-prior representations*, and concatenate it with the interacted image tokens $O_I^i \in R^{B\times N_i\times C_i}$ along the channel dimension, resulting in $F_I^i \in R^{B\times N_i\times(C_i+M_i)}$. As A_T^i contains information about the attentiveness of image features towards textual tokens, F_I^i can be seen as image tokens further enhanced by text-prior information. Then, F_I^i is reshaped to $B\times H_i\times W_i\times(C_i+M_i)$ and fed into $Dec\ i$ with skip connections, yielding the input for the next block.

In addition to utilizing text-prior information for feature-level enhancement, we propose constructing a *text-prior prediction* with A_T^i and introducing an auxiliary constraint at the objective level. Specifically, we extract the first element (index: 0) of the attention map A_T^i from the text branch, yielding $S_T^i \in R^{B\times N_i}$, which corresponds to the attention scores of the class token [CLS] in the sequence O_T^i. As the [CLS] is retained throughout the entire text token pruning and image decoding process, S_T^i contains information most representative of the segmentation target. Then we reshape S_T^i to the size of $B\times 1\times H_i\times W_i$ and pass it through an MLP-based projector ($MLP(\cdot)$), a sigmoid function $\sigma(\cdot)$, and a bilinear interpolation ($BI(\cdot)$) for expansion, obtaining the text-prior prediction $m_i \in R^{B\times 1\times H\times W}$ with spatial scales matching the ground truth $Y \in R^{B\times 1\times H\times W}$. Utilizing m_i at three scales, we construct the auxiliary loss as follows:

$$L_{aux} = \sum_{i=1}^{3} L_{dice-CE}(m_i, Y)\ ,\ \ m_i \in R^{B\times 1\times H\times W} = BI\left(\sigma\left(MLP\left(S_T^i\right)\right)\right), \quad (5)$$

where $L_{dice-CE}$ denotes the weighted combination of Dice loss and cross-entropy loss with a ratio of 1:1. By applying auxiliary loss, we compel the textual class token [CLS] to provide crucial text-prior information that furnishes the most helpful guidance for segmentation. This, in turn, facilitates the generation of the final prediction.

The overall objective function of our ABP method is defined as $L_{total} = L_{cls} + \gamma \cdot L_{aux}$, where the classification loss L_{cls} also employs the above Dice-CE loss for the final prediction M, and the hyper-parameter γ is utilized to balance the two loss terms.

Table 1. Quantitative comparison in terms of Dice, MIoU, Acc.

Type	Method	Dice	MIoU	Acc
Mono-modal segmentation methods	UNet [8]	82.99	70.92	95.84
	UNet++ [9]	83.69	71.96	96.08
	TransUNet [12]	83.91	72.79	96.13
	AttnUnet [16]	82.40	70.06	95.67
Text-guided segmentation methods	LViT [22]	84.92	73.79	96.24
	CPAM [25]	87.43	78.32	96.96
	AT [26]	89.78	81.45	97.52
	ABP (Ours)	**91.03**	**83.53**	**97.83**

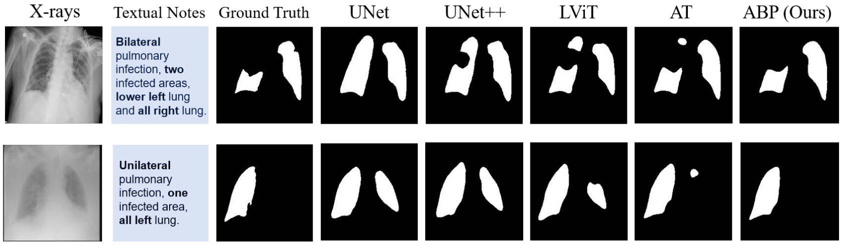
X-rays
Textual Notes
Ground Truth
UNet
UNet++
LViT
AT
ABP (Ours)
Bilateral pulmonary infection, two infected areas, lower left lung and all right lung.
Unilateral pulmonary infection, one infected area, all left lung.

Fig. 2. Qualitative comparison with other state-of-the-art methods.

3 Experiments

Implementation Details. Our network is implemented using the PyTorch framework on a single RTX 3090 GPU and trained for 100 epochs with a batch size of 32. The network weights are updated using the Adam optimizer, with an initial learning rate set to 5×10^{-5} and a minimal rate of 1×10^{-6}. During training, a cosine decay scheduler is employed to adjust the learning rate. We explored various γ values ranging from 0 to 1.0 with a step of 0.2, finding 0.4 yields optimal results across three metrics. As for token lengths, we referred to LViT [22] and set three candidate values (36, 24, and 18) for M_1, and explored two descending ratios to determine M_2 and M_3: an arithmetic progression (1, 3/4, 2/4) and a geometric progression (1, 1/2, 1/4). We found that the arithmetic progression with M_1 set at 24 is optimal. Aligned with AT [26], the text encoder (CXR-BERT) is pretrained and fixed, while the image encoder (ConvNeXt-Tiny) is trained from scratch with its output dimension C1 set to 768.

Dataset. We utilize the publicly available QaTa-COV19 dataset [7] to evaluate the effectiveness of our proposed method. This dataset was collaboratively constructed and expanded by Qatar University and Tampere University. The current version of the dataset comprises 9258 COVID-19 chest X-rays, with 5716 samples designated for training, 1429 samples for validation, and 2113 for testing, following the same configuration in [26]. Each image is accompanied by a ground-truth mask indicating the lung lesion regions affected by COVID-19. Concerning the text-guided segmentation task in QaTa-COV19, substantial contributions have been made by Li *et al*. [22] and Zhong *et al.* [26], who expanded and enhanced the text annotations of the dataset.

Comparative Experiments. To evaluate the effectiveness of the proposed ABP, we conduct a comparative analysis against four mono-modal segmentation methods: UNet

Table 2. Quantitative comparison of ablation models in terms of Dice, MIoU, Acc.

Model	Modules	Description	Dice	MIoU	Acc
A		Mono-modal Baseline	84.65	74.36	96.25
B	Bilateral Attention	A + Image Branch cross-attention	87.26	80.22	96.91
C		B + Text Branch cross-attention	89.85	81.72	97.40
D	Text-branch Operation	C+ Token Pruning w/o [FUSE]	90.32	82.34	97.67
E		D + [FUSE] Token	90.62	82.85	97.73
F	Image-branch Operation	E + Text-prior Feature	90.88	83.19	97.76
G (Ours)		F + Text-prior Predcition (L_{aux})	**91.03**	**83.53**	**97.83**

[8], UNet++ [9], TransUNet [12], AttnUnet [16], and three multi-modal text-guided segmentation methods: LViT [22], CPAM [25], and Ariadne's Thread (AT) [26]. Note that, we reproduced CPAM with the enhanced annotations [26], and AT is the current State-Of-The-Art method for text-guided segmentation on QaTa-COV19. To ensure fairness, we directly adopted the results of some methods from AT [26] and reproduced other methods using their optimal hyperparameter settings under the same evaluation process. Table 1 presents the dice score (Dice), Mean Intersection over Union (MIoU), and prediction accuracy (Acc) of different methods for segmentation results on the test samples. Upon observation, our proposed ABP method consistently outperforms the other methods on average across all three evaluation metrics. Compared to the AT, our method shows significant improvements by 1.25% for dice score, 2.08% for MIoU, and 0.31% for Acc. We also conducted paired t-tests to verify our improvements. Results indicate that p-values on three metrics are all less than 0.05, indicating statistical significance. The above enhancements can be attributed to our efficient architecture, which incorporates both symmetric bilateral cross-attention for preliminary interactions and asymmetric modality-specific operations for further refinement.

Fig. 2 showcases the qualitative segmentation results derived from the test samples. In the second row, textual notes reveal that the infection solely appears in the left lung. However, without this textual guidance, distinguishing whether the right lung is infected based solely on image content is challenging, resulting in poor performance of mono-modal methods. Notably, among the multi-modal methods incorporating text, only our ABP accurately segmented the left lung region without misdirected attention to the right lung, underscoring its exceptional ability to leverage textual guidance.

Ablation Study. To assess the contribution of each proposed module, we progressively create ablation models as follows. Initially, we remove the ABP blocks and solely utilize X-ray images for segmentation, establishing the mono-modal baseline referred to as Model-A. Then, we introduce textual descriptions as well as the image branch cross-attention to Model-A, creating the multi-modal baseline Model-B. Based on Model-B, we integrate text branch cross-attention to form the complete symmetric bilateral cross-attention as Model-C. Next, token pruning without the fusion of inattentive tokens is introduced, forming Model-D. Advancing from Model-D, we progressively incorporate fuse token, text-prior feature, and text-prior prediction (i.e., L_{aux}), yielding models Model-E, Model-F, and Model-G. As shown in Table 2, Model-A performs poorly due to the absence of textual assistance, while Model-B exhibits significant improvement with the incorporation of a simple image-branch cross-attention. Upon integrating complete bilateral cross-attention, the performance surpasses that of the AT method, which uses unilateral cross-attention, demonstrating the effectiveness of our bilateral design. The subsequent introduction of our asymmetric operations in Model-D to Model-G further enhances performance, highlighting the contributions of modality-specific processing to text-guided medical image segmentation.

4 Conclusion

In this paper, we proposed Asymmetric Bilateral Prompting (ABP), a novel method for text-guided medical image segmentation, aiming to overcome the key limitations in current methods (i.e., the lack of adaptive adjustments for text tokens and the insufficient use of text-prior information). ABP integrates symmetric bilateral cross-attention for preliminary interactions and two asymmetric operations for modality-specific refinement. By utilizing attention scores from both the image and text sides, we enabled adaptive adjustments for text tokens and enhanced the image representations. Experimental results on QaTa-COV19 dataset demonstrate the superiority of our method.

Acknowledgments. This work is supported by National Natural Science Foundation of China (NSFC 62371325, 62071314), Sichuan Science and Technology Program 2023YFG0025, 2023YFG0101, and 2023 Science and Technology Project of Sichuan Health Com-mission 23LCYJ002.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Lalmuanawma S, Hussain J, Chhakchhuak L: Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos, Solitons & Fractals 139, 110059 (2020)
2. Shi F, Wang J, Shi J, et al.: Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19. IEEE reviews in biomedical engineering 14, 4-15 (2020)
3. Wang K, Zhan B, Zu C, Wu X, et al.: Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. Med. Image Anal. 79, 102447 (2022)
4. Degerli A, Ahishali M, Kiranyaz S, et al.: Reliable covid-19 detection using chest x-ray images. In: IEEE International Conference on Image Processing, pp. 185-189. (2021)
5. Tang C, Zeng X, Zhou L, Zhou Q, et al.: Semi-supervised medical image segmentation via hard positives oriented contrastive learning. Pattern Recogn. 146: 110020 (2024)
6. Qiu Y, Liu Y, Li S, et al.: Miniseg: An extremely minimum network for efficient covid-19 segmentation. In: AAAI Conference on Artificial Intelligence, 35(6), pp. 4846-4854. (2021).
7. Tang P, Yang P, Nie D, et al.: Unified medical image segmentation by learning from uncertainty in an end-to-end manner. Knowl.-Based Syst. 241, 108215 (2022)
8. Ronneberger O, Fischer P, Brox T, et al.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) MICCAI 2015, Part III 18, pp. 234-241. Springer, Cham (2015)

9. Zhou Z, Rahman Siddiquee M M, Tajbakhsh N, et al.: Unet++: A nested u-net architecture for medical image segmentation. In: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Proceedings 4, pp. 3-11. Springer, Cham (2018)
10. Huang H, Lin L, Tong R, et al.: Unet 3+: A full-scale connected unet for medical image segmentation. In: IEEE international conference on acoustics, speech and signal processing, pp.1055-1059. (2020)
11. Nguyen T, Hua B S, Le N.: 3d-ucaps: 3d capsules unet for volumetric image segmentation. In: de Bruijne, M. et al. (eds.) MICCAI 2021, Part I 24, pp. 548-558. Springer, Cham (2021)
12. Chen J, Lu Y, Yu Q, et al.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306. (2021)
13. Yan X, Tang H, Sun S, et al.: After-unet: Axial fusion transformer unet for medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 3971-3981. (2022)
14. Cao H, Wang Y, Chen J, et al.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision, pp. 205-218. Springer, Cham (2022)
15. Hatamizadeh A, Nath V, Tang Y, et al.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop, pp. 272-284. Springer, Cham (2021)
16. Oktay O, Schlemper J, Folgoc L L, et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999. (2018)
17. Zeng X, Zeng P, Tang C, et al.: DBTrans: A Dual-Branch Vision Transformer for Multi-Modal Brain Tumor Segmentation. In: Greenspan, H., et al. (eds.) MICCAI 2023, pp. 502-512. Springer, Cham (2023)
18. Uppal S, Bhagat S, Hazarika D, et al.: Multimodal research in vision and language: A review of current and emerging trends. Information Fusion 77, 149-171 (2022)
19. Chen F L, Zhang D Z, Han M L, et al.: Vlp: A survey on vision-language pre-training. Machine Intelligence Research 20(1), 38-56 (2023)
20. Zhang Z, Yao L, Wang B, et al.: EMIT-Diff: Enhancing Medical Image Segmentation via Text-Guided Diffusion Model. arXiv preprint arXiv:2310.12868. (2023)
21. Wang P, Chung A C S. DoubleU-net.: colorectal cancer diagnosis and gland instance segmentation with text-guided feature control. In: European Conference on Computer Vision, pp. 338-354. Springer, Cham (2020)
22. Li Z, Li Y, Li Q, et al.: Lvit: language meets vision transformer in medical image segmentation. IEEE Transactions on Medical Imaging, vol. 43, no. 1, 96-107 (2023)
23. Tomar N K, Jha D, Bagci U, et al.: TGANet: Text-guided attention for improved polyp segmentation. In: Wang, L. et al. (eds.) MICCAI 2022, pp. 151-160. Springer, Cham (2022)
24. Poudel K, Dhakal M, Bhandari P, et al.: Exploring transfer learning in medical image segmentation using vision-language models. arXiv preprint arXiv:2308.07706. (2023)
25. Lee G E, Kim S H, Cho J, et al.: Text-Guided Cross-Position Attention for Segmentation: Case of Medical Image. In: Greenspan, H., et al. (eds.) MICCAI 2023, pp. 537-546. Springer, Cham (2023)
26. Zhong Y, Xu M, Liang K, et al.: Ariadne’s Thread: Using Text Prompts to Improve Segmentation of Infected Areas from Chest X-ray Images. In: Greenspan, H., et al. (eds.) MICCAI 2023, pp. 724-733. Springer, Cham (2023)
27. Kim S, Shen S, Thorsley D, et al.: Learned token pruning for transformers. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 784-794. (2022)

28. Ma J, Guo S, Zhang L.: Text prior guided scene text image super-resolution. IEEE Transactions on Image Processing 32, 1341-1353 (2023)
29. Boecking B, Usuyama N, Bannur S, et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision, pp. 1-21. Springer, Cham (2022)
30. Liu Z, Mao H, Wu C Y, et al.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11976-11986. (2022)