This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

MAMBA-Based Weakly Supervised Medical Image Segmentation with Cross-Modal Textual Information

Zhen Pan, Wenhui Huang*, and Yuanjie Zheng

School of Information Science and Engineering, Shandong Normal University, Jinan, China
whhuang.sdu@gmail.com

Abstract. In medical image segmentation, obtaining pixel-level annotated data is costly. While semi-supervised and weakly-supervised methods reduce annotation dependence, they still require some pixel-level annotations. In contrast, leveraging textual descriptions corresponding to medical images as supervisory information for segmentation is more promising. Textual descriptions are easier to acquire, as users only need to provide location and appearance details of lesions. We present TIFC-Mamba, a Mamba-based architecture for text-image fusion segmentation. The framework processes images and texts in parallel to establish cross-modal correspondences, aligning CLIP-encoded features through contrastive learning. The architecture employs a Mamba-based image encoder that reduces computational complexity compared to traditional Transformer models. We propose Mamba Fusion (MF) module integrates text and image features through Bi-Dimension Fusion (BiDF), enabling both intra-modal refinement and inter-modal interaction while preserving computational efficiency. Experiments on polyp and skin lesion datasets demonstrate competitive performance against fully supervised methods and state-of-the-art weakly-supervised approaches. Code and dataset will be available at https://github.com/PZalio/TIFCMamba.

1 Introduction

Medical image segmentation has become a critical tool for clinical diagnostics. With the development of deep learning, medical image segmentation typically relies on fully supervised paradigms, which require pixel-level annotations. However, the annotation of medical images requires expertise, and the cost of high-quality annotations is substantial, limiting the development of segmentation models [17, 18, 23]. In recent years, semi-supervised and weakly-supervised approaches have been proposed to reduce annotation costs. These methods, compared to fully-supervised approaches, still require some pixel-level labels for training, but alleviate the reliance on comprehensive annotations [10, 14]. Medical text annotations, however, are easier to obtain, as users only need to provide

[*] Wenhui Huang is the corresponding author.

lesion locations and descriptions. This makes medical text a promising solution for supervision in medical image segmentation [28,35].

Transformer-based attention mechanisms have significantly advanced multi-modal medical image segmentation by improving the fusion of image and text information [9,24]. However, the quadratic computational and memory requirements of full attention models pose challenges, especially when processing large images and lengthy text descriptions [15,26]. State Space Models (SSM), like Mamba [13] and its variant Vmamba [20], show promise due to their linear complexity and global receptive fields. However, SSMs are underexplored for multi-modal fusion, primarily focusing on multi-modal image fusion. ReMamba [33] suggests that traditional token concatenation methods are ineffective for Mamba, as its linear structure limits token interactions, leading to insufficient fusion and reduced performance.

In the text-supervised paradigm, only the semantic text corresponding to the image is used as supervisory information without any pixel-level mask annotations, and the training of the model is driven by the semantic or feature alignment of the text-image [16,25]. The approach of image-text alignment has been widely adopted in many works. Specifically, image-text alignment methods typically use an image encoder and a text encoder, aligning the two into a joint embedding space. In this way, zero-sample passing techniques can be used to allow both encoders to generate segmented outputs without specialised annotation [29]. This approach creates inconsistencies between training and testing. During training, image-text alignment is based on the entire image's semantic features, but during testing, the goal is to align text semantics with specific image regions. This misalignment may lead to suboptimal performance, as the model may not learn the relationship between local text semantics and image regions during training [19,36].

To address these challenges, we propose TIFCMamba, a novel text-supervised segmentation framework based on Mamba. The key contributions of our approach are as follows: 1) We introduce TIFCMamba, using medical text supervision and multi-modal contrastive learning to reduce annotation cost while avoiding the high computational cost of Transformer-based models. 2) We propose the Mamba Fusion Block with a bi-dimensional fusion mechanism, enhancing text-image feature interactions and addressing token fusion limitations in Mamba. 3) We introduce an image-text mutual alignment mechanism for precise alignment between image and text segments during training and testing.

2 Method

2.1 Preliminary of Mamba

The Structured State Space Model (SSM) [12,13] from control systems theory transforms an input sequence $x(t) \in \mathbb{R}$ into an output $y(t) \in \mathbb{R}$ via a hidden state $\mathbf{h}(t) \in \mathbb{R}^N$, governed by:

$$\mathbf{h}'(t) = \mathbf{A}\,\mathbf{h}(t) + \mathbf{B}\,x(t), \quad y(t) = \mathbf{C}^T\mathbf{h}(t) + D\,x(t), \tag{1}$$

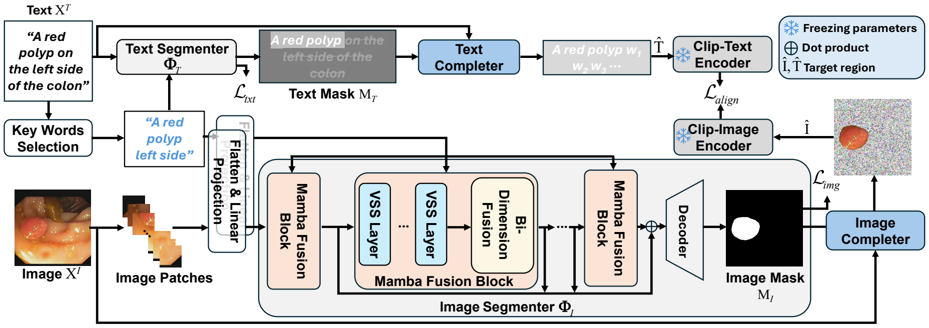

Fig. 1: TIFCMamba framework. Image and text inputs are processed by separate segmenters to generate masks and extract target regions. After random filling of non-target regions, a CLIP encoder aligns image-text features for text-supervised training of the image segmenter $\mathbf{\Phi}_I$.

where $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the state matrix, $\mathbf{B} \in \mathbb{R}^N$ the input matrix, $\mathbf{C} \in \mathbb{R}^N$ the output matrix, and $D \in \mathbb{R}$ is a skip connection (hereafter omitted, i.e., $D = 0$).

Since continuous-time systems are not directly amenable to digital computation, we discretize with time step Δ:

$$\mathbf{h}_t = \overline{\mathbf{A}}\,\mathbf{h}_{t-1} + \overline{\mathbf{B}}\,x_t, \quad y_t = \overline{\mathbf{C}}^T\,\mathbf{h}_t, \tag{2}$$

with $\overline{\mathbf{A}} = \exp(\Delta\mathbf{A})$, $\overline{\mathbf{B}} \approx \mathbf{B}$, and $\overline{\mathbf{C}} = \mathbf{C}$.

In the S4 [13] model, parameters $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \Delta)$ are learned via gradient descent but remain independent of the input, limiting contextual modeling. To address this, Mamba introduces the S6 mechanism, which makes $\mathbf{B}$, $\mathbf{C}$, and Δ input-dependent. Specifically, for an input sequence $\mathbf{x} \in \mathbb{R}^{B \times L \times C}$ (with batch size B, sequence length L, and feature dimension C), the parameters are computed as:

$$\mathbf{B} = \mathrm{Linear}(\mathbf{x}) \in \mathbb{R}^{B \times L \times N}, \mathbf{C} = \mathrm{Linear}(\mathbf{x}) \in \mathbb{R}^{B \times L \times N}, \tag{3}$$

$$\Delta = \mathrm{SoftPlus}\Big(\tilde{\Delta} + \mathrm{Linear}(\mathbf{x})\Big) \in \mathbb{R}^{B \times L \times C}. \tag{4}$$

Here, $\mathrm{Linear}(\cdot)$ denotes a linear transformation and $\mathrm{SoftPlus}(\cdot)$ ensures non-negativity for Δ; $\tilde{\Delta}$ is a learnable bias. This input-dependent design enhances the model's adaptability and its ability to capture contextual information.

2.2 Overall Framework

The core challenge of text-supervised segmentation is establishing semantic correspondences between images and text. As shown in Fig. 1, our TIFCMamba framework operates on a medical image-text dataset $D = \{(X_1^I, X_1^T), (X_2^I, X_2^T), \cdots, (X_i^I, X_i^T), \cdots (X_n^I, X_n^T)\}$, where images lack pixel-level labels and are only annotated with semantic descriptions. We jointly train an image segmenter $\mathbf{\Phi}_I$ and a text segmenter $\mathbf{\Phi}_T$ using contrastive learning to align the segmented regions across modalities.

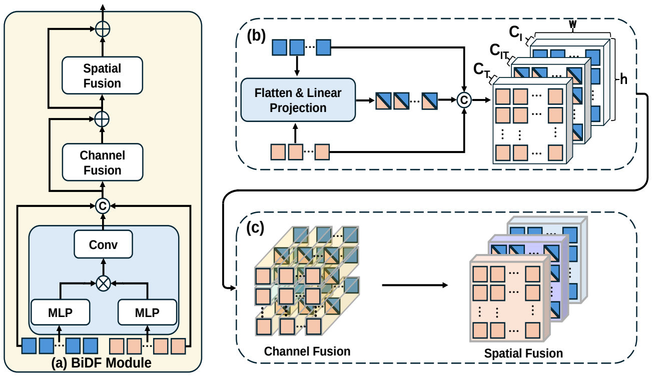

Fig. 2: BiDF module. Firstly, the different modal interactions of image-text are fused, then the three modal information is spliced in the channel dimension, and finally the different modalities are fully interacted by channel and space fusion in sequence.

For each image-text pair (X_i^I, X_i^T), a keyword selector [5] extracts keywords W_T from X_i^T (e.g., “A red polyp left side”). The image segmenter $\mathbf{\Phi}_I$ uses X_i^I and W_T to generate an image mask M_I, while the text segmenter $\mathbf{\Phi}_T$ processes X_i^T and W_T to produce a text mask M_T. The masked image $\hat{I}$ is obtained by cropping X_I with M_I and randomly padding the background, and similarly, a completed text $\hat{T}$ is constructed using M_T. Finally, CLIP [21]'s image encoder E_I and text encoder E_T extract features from $\hat{I}$ and $\hat{T}$, respectively, and contrastive learning aligns their representations.

2.3 Image-Text Segmentation

Mamba-Based Image Segmenter. As shown in Fig. 1, our image segmenter centers on the *Mamba Fusion Block*, which integrates visual and textual modalities. Intermediate features from each block are skip-connected to the decoder to produce the final segmentation mask. The fusion block comprises two modules. First, the *Visual State Space* (VSS) [20]module treats image features as token sequences and employs a 2D-Selective-Scan (SS2D) [20] mechanism that scans features in four directions to capture long-range spatial dependencies. Second, the *Bi-Dimension Fusion* (BiDF) module seeks to fuse text features with image features. It calculates the cross-correlation between the two modalitie: image features and text features, and subsequently propagates this information to the features of each image patch.

BiDF Module. As illustrated in Fig. 2(a), the BiDF module operates in two stages. In the first stage, the BiDF module first expands the text features $\mathbf{F}_T$ into $\hat{\mathbf{F}}_T \in \mathbb{R}^{h \times w \times C_T}$, then fuses the image features $\mathbf{F}_I$ with the expanded text features $\hat{\mathbf{F}}_T$ to obtain $\mathbf{F}_{IT} \in \mathbb{R}^{h \times w \times C_{IT}}$. This process allows each image patch to incorporate textual information. Finally, the three features are concatenated along the channel dimension. The process is formalized as follows:

$$\mathbf{F}_{cat} = \text{Concate}\left(\mathbf{F}_I, Conv(\mathbf{F}_I\mathbf{W}_I \cdot (\hat{\mathbf{F}}_T\mathbf{W}_T)^T), \hat{\mathbf{F}}_T\right), \tag{5}$$

where $\mathbf{W}_I \in \mathbb{R}^{C_I \times C_0}$ and $\mathbf{W}_T \in \mathbb{R}^{C_T \times C_0}$ are learnable parameters. $\mathbf{F}_I\mathbf{W}_I \cdot (\hat{\mathbf{F}}_T\mathbf{W}_T)^T \in \mathbb{R}^{h \times w}$ is transformed via a 1×1 convolution to obtain $\mathbf{F}_{IT} \in \mathbb{R}^{h \times w \times C_{IT}}$.

The concatenated features $\mathbf{F}_{cat}$ are then fused in two stages: spatial fusion (using the VSS module's 2D selective scan) followed by channel fusion (with a State Space Model for a 1D scan), resulting in the final fused feature $\mathbf{F}_{fuse}$. As shown in Fig. 2(c), In the second stage, to enhance cross-modal interactions, we design two fusion mechanisms: one along the channel dimension and one along the spatial dimension. For the channel dimension, we adopt the 1D selective scan from VMamba, while for the spatial dimension, we use the 2D selective scan. The process is formalized as follows:

$$\mathbf{F}_{fuse} = Spatial(Channel(\mathbf{F}_{cat})). \tag{6}$$

Text Segmenter. Text segmenter $\mathbf{\Phi}_T$ that processes an input text X_i^T and a set of noun $\{N_j\}_{j=1}^J$ to generate a noun-specific word mask. A CLIP text encoder [21] augmented with two learnable multi-head attention layers extracts word features $\mathbf{x}_t = \tilde{E}_T(X_i^T) \in \mathbb{R}^{L \times C}$, where L is the number of tokens and C the feature dimension. For the given noun N_j (with embedding $\mathbf{n}_j \in \mathbb{R}^C$), word-specific logits are computed as $\ell_j = w \cdot (\mathbf{x}_t \cdot \mathbf{n}_j) + b$, with learnable parameters w and b and a dot product computed per token. Each word is assumed to belong either to one of the J noun-associated segments or to none; accordingly, a softmax over the J segments plus an extra "none" category yields the word mask $M_T = [m_i^t]_{i=1}^L$ defined by

$$m_i^t = \frac{\exp(\ell_{j,i})}{1 + \sum_{j'=1}^J \exp(\ell_{j',i})}, \quad i = 1, \ldots, L. \tag{7}$$

A pseudo-label vector $p \in \{0,1\}^L$ is then generated by setting $p_i = 1$ if word i attains the highest probability for one of the J segments and $p_i = 0$ otherwise; the text segmentation loss $\mathcal{L}_{txt}$ is defined as the cross-entropy between M_T and p, guiding $\mathbf{\Phi}_T$ to correctly segment the text. To align image and text modalities, our framework leverages contrastive learning between image regions and corresponding text segments. Specifically, CLIP encoders extract region embeddings $\mathbf{e}^I = E_I(\hat{I})$ from image regions $\hat{I}$ and word embeddings $\mathbf{e}^T = E_T(\hat{T})$ from text segments $\hat{T}$. For a batch of B triplets (each consisting of an image, its paired text, and a selected noun), we compute a similarity matrix $S \in \mathbb{R}^{B \times B}$, where each element $S_{i,j}$ is the cosine similarity between $\mathbf{e}_i^I$ and $\mathbf{e}_j^T$. A symmetric InfoNCE [21] loss is then applied:

$$\mathcal{L}_{align} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(S_{i,i}/\tau)}{\sum_{j=1}^B \exp(S_{i,j}/\tau)} + \log\frac{\exp(S_{i,i}/\tau)}{\sum_{j=1}^B \exp(S_{j,i}/\tau)}\right], \tag{8}$$

Table 1: Comparison with the SOTA method on the ClinicDB, ColonDB, LaribPolypDB, and ISIC2017 datasets, containing two fully supervised (FS) and six weakly supervised methods.

Method	ClinicDB		ColonDB		LaribPolypDB		ISIC2017		Supervised Mode
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	
ResUNet (2020) [11]	81.33	77.40	83.62	75.78	79.88	76.47	83.52	78.06	FS
SwinUnet-T (2021) [6]	86.64	82.24	85.90	83.76	88.43	80.05	85.47	81.35	FS
WeakPolyp (2023) [30]	84.30	81.56	86.67	79.89	82.79	80.60	85.41	82.67	Box
TCL (2023) [7]	84.35	80.89	85.02	81.67	85.58	78.32	85.46	81.90	Text
SimSeg (2023) [34]	85.17	80.38	84.92	80.16	85.60	79.73	87.28	82.49	Text
SimTxtSeg (2024) [32]	86.38	81.72	85.18	80.95	86.43	80.30	86.51	80.94	Text
CoDe (2024) [31]	86.98	82.45	86.58	81.45	87.35	81.07	87.09	83.51	Text
XCoOp (2024) [4]	86.55	82.43	85.73	80.18	88.31	80.61	86.36	81.52	Text
TIFCMamba-T	87.50	81.53	87.38	80.93	87.67	81.09	87.20	83.13	Text
TIFCMamba-S	88.07	83.93	87.67	81.45	87.81	81.77	87.79	83.59	Text
TIFCMamba-B	**88.24**	**84.22**	**87.74**	**82.56**	**88.92**	**82.43**	**87.95**	**83.76**	Text

with a learnable temperature parameter τ; note that even if the same noun is selected multiple times, the corresponding regions and text segments remain distinct, ensuring effective alignment.

Image and Text Completer. In order to avoid image-text blank regions from adversely affecting the encoded feature alignment when performing Clip image-text encoding, we introduced image-text completer respectively. We avoid the unfavorable effect by randomly complementing pixels or words in regions other than the image-text target region with the complemented image $\hat{I}$ and text $\hat{T}$, respectively: $\hat{I} = X^I \cdot M_I + \text{Fill}_I(1 - M_I)$, $\hat{T} = X^T \cdot M_T + \text{Fill}_T(1 - M_T)$.

Dual-Modal Contrastive Alignment Loss. For the image segmenter, we use the loss function $\mathcal{L}_{img}$ of TCL [7], which relies only on image-text pairs for training. The overall loss is a weighted sum of the image segmentation loss $\mathcal{L}_{img}$, text segmentation loss $\mathcal{L}_{txt}$, and the contrastive loss $\mathcal{L}_{align}$:

$$\mathcal{L}_{DMCA} = \lambda_{img}\mathcal{L}_{img} + \lambda_{txt}\mathcal{L}_{txt} + \lambda_{align}\mathcal{L}_{align}, \tag{9}$$

where we set the loss coefficients to $\lambda_{img} = 1.0$, $\lambda_{txt} = 1.0$, $\lambda_{align} = 0.5$.

3 Experiments

3.1 Experiment Setting

Datasets. We conducted experiments on polyp medical imaging datasets, augmented with textual cues to improve segmentation performance. For polyp segmentation, three publicly available colonoscopy datasets were used: CVC-ClinicDB [2], CVC-ColonDB [27], ETIS-LaribPolypDB [22], and ISIC2017 [3]. We used GPT-4 [1] to generate descriptions of the images in the dataset and adapted some of the textual descriptions, including the location, appearance, bounding box, and the proportion of the image occupied by the lesion, and made adjustments to some of the textual descriptions. The textual description of the dataset can be found in our code.

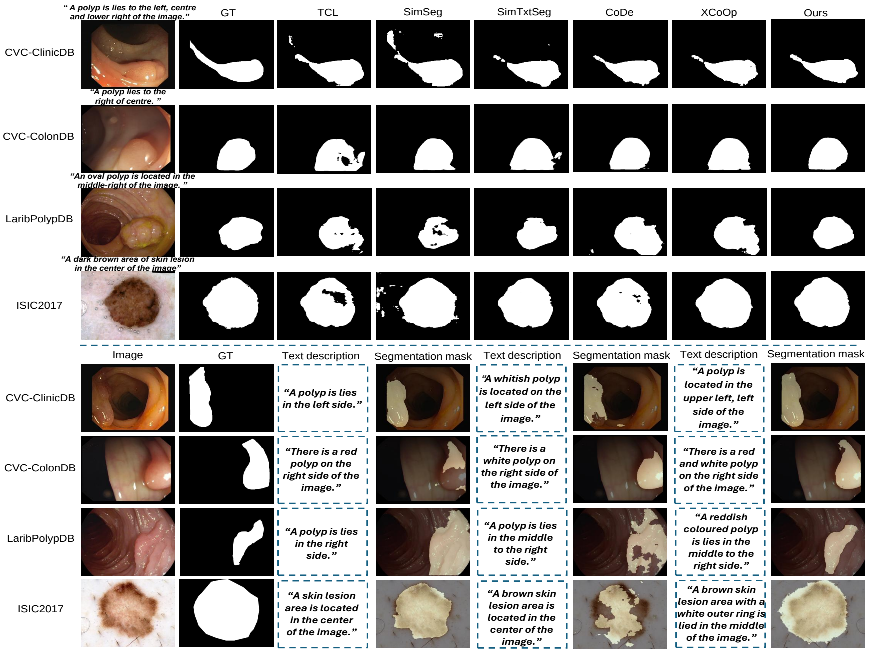
" A polyp is lies to the left, centre and lower right of the image."
GT
TCL
SimSeg
SimTxtSeg
CoDe
XCoOp
Ours
CVC-ClinicDB
"A polyp lies to the right of centre. "
CVC-ColonDB
"An oval polyp is located in the middle-right of the image. "
LaribPolypDB
"A dark brown area of skin lesion in the center of the image"
ISIC2017
Image
GT
Text description
Segmentation mask
Text description
Segmentation mask
Text description
Segmentation mask
CVC-ClinicDB
"A polyp is lies in the left side."
"A whitish polyp is located on the left side of the image."
"A polyp is located in the upper left, left side of the image."
CVC-ColonDB
"There is a red polyp on the right side of the image."
"There is a white polyp on the right side of the image."
"There is a red and white polyp on the right side of the image."
LaribPolypDB
"A polyp is lies in the right side."
"A polyp is lies in the middle to the right side."
"A reddish coloured polyp is lies in the middle to the right side."
ISIC2017
"A skin lesion area is located in the center of the image."
"A brown skin lesion area is located in the center of the image."
"A brown skin lesion area with a white outer ring is lied in the middle of the image."

Fig. 3: Qualitative comparison of segmentation results with different models under text supervision, and the effect of different text descriptions on the segmentation results of TIFCMamba-T.

Implementation details. Performance was evaluated using the mean Intersection over Union (mIoU) and mean Dice coefficient. All datasets were split into training and testing sets with a 7:3 ratio, each image is resized to 480×480. We designed three image segmenter variants of our TIFCMamba structure based on VMamba. The training process consisted of two stages, firstly initial pre-training on the CC3M [8] dataset using the Adam optimizer (with two randomly selected nouns per image-text pair), followed by fine-tuning on the polyp and skin dataset with the text branch parameters frozen. Experiments were executed on four NVIDIA 3090 GPUs with a batch size of 4 and a learning rate of 1×10^{-5}.

3.2 Results and Analysis

In Table 1, we compare the top five existing weakly-supervised models with two fully-supervised models. Among them, WeakPolyp [30] utilizes bounding boxes as supervision, while SimTxtSeg [32], TCL [7], CoDe [31], and XCoOp [4] employ text as supervision. Compared to fully-supervised methods, which entail higher annotation costs, our approach achieves comparable segmentation performance. When compared to other state-of-the-art weakly-supervised segmentation models, our model demonstrates improvements in mDice and mIoU by +1.26% and +1.77%, +1.16% and +0.89%, +0.61% and +1.36%, and +0.67% and +0.25%

Table 2: Quantitative analysis of TIFCMamba-B on fusion modes.

Fusion Mode		Polyp	ISIC2017
Spatial	Channel	mDice	mDice
×	×	63.57	58.39
✓	×	72.36	70.95
×	✓	79.86	77.49
✓	✓	88.24	87.95

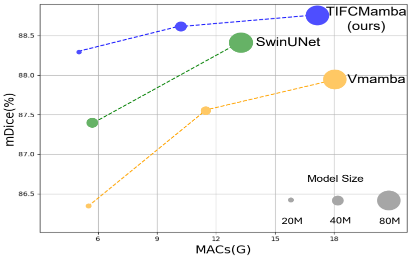

Fig. 4: Comparison of model efficiency on Polyp's three datasets. The SwinUnet is a purely Transformer-based model.

on the ClinicDB, ColonDB, LaribPolypDB and ISIC2017 datasets, respectively. A qualitative analysis of the segmentation results compared to the five text-supervised models is illustrated in Fig. 3.

3.3 Ablation Study

Impact of textual descriptions. We compared the effect of using different text descriptions on the segmentation performance of our model. As shown in Fig. 3, different text descriptions have a large impact on the segmentation results, especially when the text is inaccurately describing the position of the foreground in terms of orientation.In addition when the text description is too redundant it also affects the segmentation performance of the model.

Impact of fusion mode. We compared the effect of the cross-modal feature fusion approach we used on the segmentation performance when image-text feature fusion is used. As shown in Table 2, it can be seen that when using Spatial fusion and Channel fusion alone respectively, both have a positive effect on the model performance, but channel fusion contributes more to the segmentation performance improvement.

Comparison on model efficiency. We compared the model efficiency of our model with VMamba, SwinUnet on polyp datasets. Since our model uses an additional text segmenter, we only compared the efficiency of our image segmenter with the above two models during the testing phase. As shown in Fig. 4, our TIFCMamba-T achieves the best balance of mDice vs. MACs and Model Size.

4 Discussion and Conclusion

In this paper, we propose a textual weakly supervised medical image segmentation model based on Mamba, which implements the Mamba structure in cross-modal fusion of medical images. The Mamba architecture avoids the problem of over-complexity of the attention mechanism in Transformer, and at the same time, we use easily-accessible text descriptions as the weakly-supervised supervisory information, and achieve segmentation performance comparable to that of full supervision on the dataset of polyps and skin lesions, which signifies the promising prospect of using text supervision in medical image segmentation. However, our method still has problems such as relying on more accurate text descriptions, the complexity of the cross-modal fusion module, and the fact that the image Decoder does not consider cross-modal fusion features. In the future, we will work on solving these problems.

Acknowledgments. This work is supported by funds from the National Natural Science Foundation of China (62476156, 62003196, 62076249, 62072289, and 62073201), the Young Talent of Lifting Engineering for Science and Technology in Shandong (SDAST2024QTA091), the Youth Innovation Technology Project of Higher Education in Shandong Province (2023KJ193) and the Provincial Natural Science Foundation of Shandong Province of China (ZR2020QF032).

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

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics **43**, 99–111 (2015)
3. Berseth, M.: Isic 2017-skin lesion analysis towards melanoma detection. arXiv preprint arXiv:1703.00523 (2017)
4. Bie, Y., Luo, L., Chen, Z., Chen, H.: Xcoop: Explainable prompt learning for computer-aided diagnosis via concept-guided context optimization. In: MICCAI. pp. 773–783. Springer (2024)
5. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc." (2009)
6. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: ECCV. pp. 205–218. Springer (2022)
7. Cha, J., Mun, J., Roh, B.: Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In: CVPR. pp. 11165–11174 (2023)
8. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: CVPR. pp. 3558–3568 (2021)
9. Che, T., Zheng, Y., Yang, Y., Hou, S., Jia, W., Yang, J., Gong, C.: Sdof-gan: Symmetric dense optical flow estimation with generative adversarial networks. IEEE Transactions on Image Processing **30**, 6036–6049 (2021)
10. Chen, Z., Zheng, Y., Gee, J.C.: Transmatch: A transformer-based multilevel dual-stream feature matching network for unsupervised deformable image registration. IEEE Transactions on Medical Imaging **43**(1), 15–27 (2024)

11. Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C.: Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing **162**, 94–114 (2020)
12. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
13. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021)
14. Han, Z., Huang, W.: Discrete residual diffusion model for high-resolution prostate mri synthesis. Physics in Medicine and Biology (2024)
15. Huang, W., Gu, J., Duan, P., Hou, S., Zheng, Y.: Exploiting probabilistic siamese visual tracking with a conditional variational autoencoder. In: ICRA. pp. 14213–14219. IEEE (2021)
16. Huang, W., Gu, J., Ma, X., Li, Y.: Correlation filter-based self-paced object tracking. In: ICRA. pp. 4437–4442. IEEE (2017)
17. Huang, W., Gu, J., Ma, X., Li, Y.: End-to-end multitask siamese network with residual hierarchical attention for real-time object tracking. Applied Intelligence **50**, 1908–1921 (2020)
18. Li, W., Huang, W.: Joint optic disc and cup segmentation with parallel cooperative diffusion model. In: BIBM. pp. 2017–2020. IEEE (2023)
19. Liu, F., Huang, W.: A diffusion model-based joint dual-task network for low-quality retinal image enhancement and vessel segmentation. In: BIBM. pp. 2107–2110. IEEE (2023)
20. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. Advances in Neural Information Processing Systems **37**, 103031–103063 (2025)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
22. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. International Journal of Computer Assisted Radiology and Surgery **9**, 283–293 (2014)
23. Song, J., Zheng, Y., Xu, C., Zou, Z., Ding, G., Huang, W.: Improving the classification ability of network utilizing fusion technique in contrast-enhanced spectral mammography. Medical Physics **49**(2), 966–977 (2022)
24. Sun, M., Huang, W., Zhang, H., Shi, Y., Wang, J., Gong, Q., Wang, X.: Temporal contexts for motion tracking in ultrasound sequences with information bottleneck. Medical Physics (2023)
25. Sun, M., Huang, W., Zheng, Y.: Instance-aware diffusion model for gland segmentation in colon histology images. In: MICCAI. pp. 662–672. Springer (2023)
26. Sun, M., Wang, J., Gong, Q., Huang, W.: Enhancing gland segmentation in colon histology images using an instance-aware diffusion model. Computers in Biology and Medicine **166**, 107527 (2023)
27. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE Transactions on Medical Imaging **35**(2), 630–644 (2015)
28. Wang, J., Ge, X., Shi, Y., Sun, M., Gong, Q., Wang, H., Huang, W.: Dual-modal information bottleneck network for seizure detection. International Journal of Neural Systems **33**(01), 2250061 (2023)
29. Wang, J., Zheng, Y., Ma, J., Li, X., Wang, C., Gee, J., Wang, H., Huang, W.: Information bottleneck-based interpretable multitask network for breast cancer classification and segmentation. Medical Image Analysis **83**, 102687 (2023)

30. Wei, J., Hu, Y., Cui, S., Zhou, S.K., Li, Z.: Weakpolyp: You only look bounding box for polyp segmentation. In: MICCAI. pp. 757–766. Springer (2023)
31. Wu, J.J., Chang, A.C.H., Chuang, C.Y., Chen, C.P., Liu, Y.L., Chen, M.H., Hu, H.N., Chuang, Y.Y., Lin, Y.Y.: Image-text co-decomposition for text-supervised semantic segmentation. In: CVPR. pp. 26794–26803 (2024)
32. Xie, Y., Zhou, T., Zhou, Y., Chen, G.: Simtxtseg: Weakly-supervised medical image segmentation with simple text cues. In: MICCAI. pp. 634–644. Springer (2024)
33. Yang, Y., Ma, C., Yao, J., Zhong, Z., Zhang, Y., Wang, Y.: Remamber: Referring image segmentation with mamba twister. In: ECCV. pp. 108–126. Springer (2024)
34. Yi, M., Cui, Q., Wu, H., Yang, C., Yoshie, O., Lu, H.: A simple framework for text-supervised semantic segmentation. In: CVPR. pp. 7071–7080 (2023)
35. Zheng, Y., Sui, X., Jiang, Y., Che, T., Zhang, S., Yang, J., Li, H.: Symreg-gan: Symmetric image registration with generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 5631–5646 (2022)
36. Zheng, Y., Yang, Y., Che, T., Hou, S., Huang, W., Gao, Y., Tan, P.: Image matting with deep gaussian process. IEEE Transactions on Neural Networks and Learning Systems (2022)