MIBF-Net: Multi-modal Information Balanced Fusion Network for Clinical Diagnosis via Patient Narratives and Lesion Image

Zixuan Tang[1,5], Bai Sun[1,5], Shidan He[1,5], Yuan Hong[2], Dongdong Yu[3], Zhenzhong Liu[4], Mengtang Li[1], Bin Chen[3], and Shen Zhao[1,5]⋆

[1]School of Intelligent Systems Engineering, Sun Yet-Sen University, Shenzhen 518107, China
[2]Zhejiang Normal University, Hangzhou 321004, China
[3]Orthopedics Department, The First Affiliated Hospital of Zhejiang University, Hangzhou 310003, China
[4]Tianjin Key Laboratory for Advanced Mechatronic System Design and Intelligent Control, School of Mechanical Engineering, Tianjin University of Technology, Tianjin 300384, China
[5]Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou, 510006, China

Abstract. Accurate clinical diagnosis requires comprehensive analysis of medical imaging and patient narratives. However, current computer-aided diagnosis methods focus primarily on imaging modalities while neglecting the integration of patient-reported clinical narratives, due to the scarcity of high-quality patient narratives and the limitations in multimodal information fusion. To address these issues, we propose a dual-component framework consisting of: 1) a Retrieval Augmented Patient Narratives Generation Module (RANGM) that employs a retrieval-enhanced mechanism to guide pre-trained large language models in generating clinically plausible patient narratives; and 2) a Multimodal Information Balanced Fusion Network (MIBF-Net) incorporating our novel Information Balanced Fusion Attention (IBFA) module for effective cross-modal integration, along with a Modal Prediction-Divergent Loss (MP-Loss) to enhance the model's ability to diagnose samples that have ambiguous single modal prediction distribution. Owing to the plug-and-play design, our MIBF-Net can integrate with existing imaging-based state-of-the-art methods. Extensive experiments demonstrate significant performance improvements of 2.3%-4.6% on the HAM10000 dataset and 3.8%-6.4% on the ISIC2019 dataset. Our code is publicly available at https://anonymous.4open.science/r/MIBF-Net-2B52/.

Keywords: Multi-modal · Medical Image Classification · Deep Learning

⋆ Corresponding author

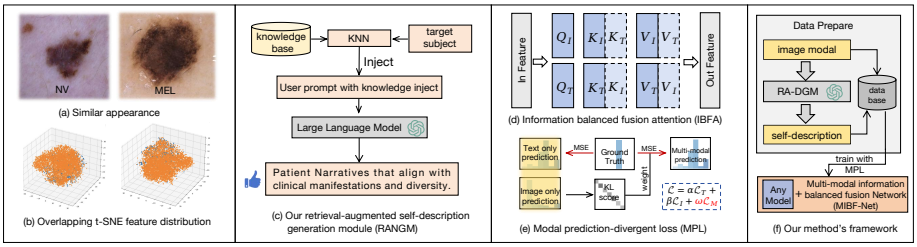


Fig. 1: (a)-(b) Some diseases, such as NV and MEL, exhibit highly similar visual appearances with overlapping feature distributions.(c) Our RANGM module enhances LLM generation capabilities by retrieving relevant medical knowledge through K-Nearest Neighbors (KNN) matching, enabling the generation of highly realistic patient narratives.(d)-(e) The meticulously designed Information Balanced Fusion Attention (IBFA) module achieves effective integration across modalities. Through the Modal Prediction-Divergent Loss (MP-Loss), we reinforce the model's focus on analyzing challenging samples with ambiguous modality information.(f) Our approach can serve as a plug-and-play solution that can be seamlessly integrated with any existing model to enhance their understanding of patient narratives and improve diagnostic accuracy.

1 Introduction

The integration of linguistic modalities (such as patient narratives) with medical imaging is essential for accurate disease diagnosis[24], as visual data alone often inadequately captures complex medical conditions. For instance, early-stage melanomas and nevi are challenging to differentiate visually [21,12,20] , as both exhibit irregular shapes and brownish-black pigmentation (Fig. 1(a)). This difficulty is further compounded by the significant overlap in their t-SNE distributions (Fig. 1(b)). Patient narratives, such as localized pain or lifestyle factors, provide critical contextual information that enhances diagnostic precision. Thus, combining linguistic and visual modalities is vital for robust disease diagnosis[3,6,18,29].

However, obtaining and using patient narratives is challenging [16][9]. Clinically, there is a lack of high-quality patient narratives data with image-text matching. Additionally, the use of existing text generation models is limited by poor medical knowledge generation quality and low diversity [23,2,27], making it challenging to simulate real clinical diagnostic scenarios. Meanwhile, there is a lack of effective multi-modal fusion methods and training approaches that focus on samples with predictive discrepancies [7,13,10]. Existing multi-modal information fusion methods exhibit information weighting bias when attending to multiple modalities, failing to balance attention between the two types of information [14,26,30,8,28]. For samples with significant discrepancies in single-modal predictions, there is no effective method to specifically supervise these samples [1].

To address these issues, we propose a novel probability recall augmented generation method (Fig. 1(c)) capable of generating realistic patient narratives. Additionally, we have designed an information balanced fusion attention-IBFA (Fig. 1(d)) and, based on this attention, implemented a multimodal classification method called MIBF-Net. We also devised a modal prediction-divergent loss (MP-Loss) to assist the model in focusing on the learning of samples with ambiguous single modal prediction distribution (i.e., lesion image modal and patient narratives modal are trend to predict different disease). In this paper, our contributions are fourfold:

- We are the first to combine patient narratives with lesion images for computer-aided diagnosis. Patient narratives provide information beyond the lesion images, such as pain perception. Compared to using lesion images alone for diagnosis, our method better meets clinical needs and aligns more closely with the logic of clinical diagnosis.
- To generate high quality patient narratives, we propose the Retrieval Augmented Patient Narratives Generation Module (RANGM), which can retrieve corresponding knowledge to enhance the professionalism and diversity of the patient narratives generated by the LLM.
- To utilize patient narratives and lesion images efficiently, we propose the Multi-modal Information Balanced Fusion Network (MIBF-Net). Thanks to our Information Balanced Fusion Attention (IBFA), MIBF-Net can fuse information from the two modalities in a balanced manner, without being dominated by single-modal judgments.
- To focus more on hard samples that have ambiguous single modal prediction distribution, we propose a meticulously designed modal prediction-divergent loss (MP-Loss) to supervise network training. This loss function guides the model pay more attention on these samples using KL divergency score, thereby improving classification performance for challenging samples.

2 Method

Retrieval Augmented Patient Narratives Generation Module (RANGM) overcomes the limitations of existing large language models (LLM) in downstream professional knowledge, enabling more comprehensive and diverse generation. We have compiled a substantial knowledge base by gathering information related to the diseases covered in our dataset. This knowledge base is then hashed using a BERT encoder as the hash function. To retrieve relevant knowledge, we employ the same BERT encoder for the disease names and their basic definitions that we wish to query. Through KNN (k-nearest neighbors) hashing matching with the knowledge base, we obtain the retrieval knowledge. To enhance diversity, we randomly select the top M pieces of knowledge with the highest matching scores as the final retrieval knowledge, which is then integrated into our meticulously designed prompt. Assume that the patient narratives, query tokens, and

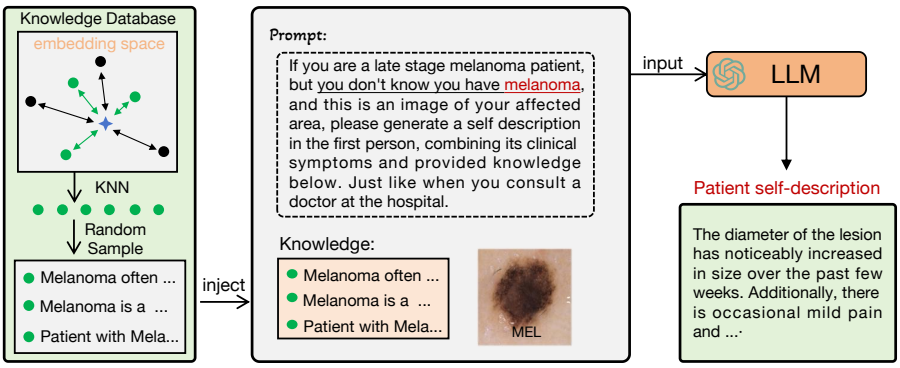


Fig. 2: The architecture of the proposed RANGM. It processes query disease embedding by performing KNN matching with encoded vectors from the knowledge base, randomly selecting M matches from the top K candidates to inject into the prompt, thereby guiding the LLM to generate high-quality patient narratives.

knowledge tokens are y, q, $\mathcal{H}$, the above process can be formulated as:

$$y = \arg\max_{y} \mathcal{LLM}\left(y \mid p, \mathcal{R}_{\text{final}} = \text{RandomSelect}\left(\underset{h_i \in \mathcal{H}}{\text{argtopK}}\left(\frac{q \cdot h_i}{\|q\|\|h_i\|}\right), M\right)\right) \tag{1}$$

Multi-modal information balanced fusion network (MIBF-Net). The pipeline of our MIBF-Net is shown in Fig. 3. The text modal and image modal input will firstly passing through their respective encoders to obtain their own features, F_t and F_i. These two features will first pass through two MLP classification networks to predict the Image only Prediction (IoP) and Text only Prediction (ToP), which are used for subsequent loss computation. Then, the two features are fed into the Information Balanced Fusion Attention (IBFA) module. Our IBFA concatenates the K and V vectors of the image modal and text modal, ensuring that during the attention query and weighting operations, the model can equally attend to information from its own modality and the queried modality. This achieves efficient feature fusion (i.e., balancing attention to inter-modal discrepancies while maintaining focus on intra-modal features). Meanwhile, thanks to the design of multi-head mechanism, our IBFA module is capable of decomposing features from different modalities into higher h dimensions, achieving a more granular level of fusion. We employed IBFA twice: once using the image query as the search input and another time using the text query as the search input, which can be formulated as:

$$\mathbf{F}_{i2t} = \mathcal{IBFA}(\mathbf{Q}_i, \mathbf{K}, \mathbf{V}) = \left[\sigma\left(\frac{\mathbf{Q}_i \mathbf{W}_j^{(q)} \cdot ([\mathbf{K}_i, \mathbf{K}_t]\mathbf{W}_j^{(k)})^\top}{\sqrt{d_k}}\right)[\mathbf{V}_i, \mathbf{V}_t] \cdot \mathbf{W}_j^{(v)}\right]_{j=1}^{h} \mathbf{W}^0 \tag{2}$$

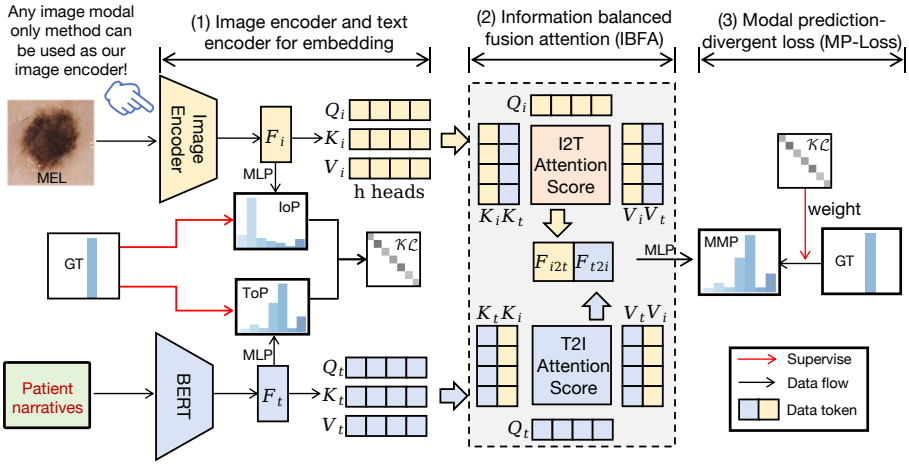


Fig. 3: The architecture of the proposed MIBF-Net: the input images and patient narratives are first encoded through encoders to obtain individual modality predictions (Image only prediction (IoP) and Text only Prediction ToP), then the encoded features are fused through IBFA for multimodal feature integration, with MLP subsequently predicting multimodal results (MMP), while MP-Loss loss optimizes network parameters by leveraging the distribution discrepancy between IoP and ToP along with MMP prediction distribution bias.

$$\mathbf{F}_{t2i} = \mathcal{IBFA}(\mathbf{Q}_t, \mathbf{K}, \mathbf{V}) = \left[\sigma\left(\frac{\mathbf{Q}_t\mathbf{W}_j^{(q)} \cdot ([\mathbf{K}_t, \mathbf{K}_i]\mathbf{W}_j^{(k)})^\top}{\sqrt{d_k}}\right)[\mathbf{V}_t, \mathbf{V}_i] \cdot \mathbf{W}_j^{(v)}\right]_{j=1}^{h} \mathbf{W}^0 \tag{3}$$

where $[\cdot]$ means concatenate, i represents image modal, t represents text modal and $\sigma(z) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$ is softmax function used to normalize the probability distribution of attention scores, emphasizing important information. Then the fused information $[\mathbf{F}_{i2t}, \mathbf{F}_{t2i}]$ from the two modalities is concatenated and fed into the final MLP classifier to get multi-modal predictions (MMP).

Modal prediction-divergent loss. We propose a loss function that leverages the average Kullback-Leibler (KL) divergence [17] between the classification predictions of the image and text modalities to dynamically weight the supervision of the final multimodal output. We denote the network parameters for predicting these three outputs as θ_i, θ_t, and $\theta_{i,t}$, respectively. As illustrated in Fig. 3, we employ two Multi-Layer Perceptrons (MLP) to generate the *image-only prediction* (i.e., IoP=$f(\mathbf{x};\theta_i)$) and the *text-only prediction* (i.e., ToP=$f(\mathbf{x};\theta_t)$). Simultaneously, the IBFA module fuses the information from both modalities to produce the *multi-modal prediction* (i.e., MMP=$f(\mathbf{x_i}, \mathbf{x_t};\theta_{i,t})$). Based on this framework, the modal prediction-divergent loss can be formulated as follows:

$$\mathcal{L}_{\text{MP-Loss}} = \mathbb{E}_{(\mathbf{x_i}, \mathbf{x_t}, y)\sim\mathcal{D}} \begin{bmatrix} \alpha \cdot ||y - f(\mathbf{x_i}; \theta_i)||_2 \\ + \beta \cdot ||y - f(\mathbf{x_t}; \theta_t)||_2 \\ + \gamma \cdot \mathcal{KL} \cdot ||y - f(\mathbf{x_i}, \mathbf{x_t}; \theta_{i,t})||_2 \end{bmatrix} \tag{4}$$

where, $\mathcal{KL}$ denotes the Kullback-Leibler divergence score calculated between the IoP and the ToP:

$$\mathcal{KL} = \frac{1}{2}\left(\sum_x f(\mathbf{x}; \theta_i) \log \frac{f(\mathbf{x}; \theta_i)}{f(\mathbf{x}; \theta_t)} + \sum_x f(\mathbf{x}; \theta_t) \log \frac{f(\mathbf{x}; \theta_t)}{f(\mathbf{x}; \theta_i)}\right) \tag{5}$$

This approach enhances learning for challenging samples where there is a significant information discrepancy between the image and text modalities.

3 Experiment

Datasets. We conduct experiment on ISIC2019 [1] and HAM10000 [22]datasets to prove the effectiveness of proposed method. The ISIC2019 dataset contains 25,531 dermatological images, while the HAM10000 dataset comprises 10,015 dermatological images. Both datasets cover seven skin disease categories, including melanoma (MEL), nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AK), benign keratosis (BKL), dermatofibroma (DF), and vascular lesions (VASC). Additionally, the ISIC2019 dataset includes an extra category of squamous cell carcinoma (SCC). These datasets are characterized by significant class imbalance and high visual similarity between different categories, making them challenging benchmarks. For the experimental setup, we adopt a random split strategy, using 80% as the training set and 20% as the test set. we use Top-1 accuracy and F1-score to evaluate the model's performance. All images are resized to 224 pixles.

Implementation Details. We use BERT [4] as text encoder. For the image encoder, our approach can incorporate existing image-only diagnostic methods as our image encoder, enabling these state-of-the-art models to seamlessly understand patient narratives and perform clinical scenario-compliant diagnostic tasks. We employ GPT-4 as our Large Language Model (LLM), and API access can be obtained through an official subscription [2]. Since both datasets include records of the lesion locations, we have also integrated the lesion locations into the outputs generated by the LLM. For the loss function, we set $\alpha = 0.6$, $\beta = 0.4$, $\gamma = 0.1$. We utilize the Adam optimizer with a learning rate of 5e-5 to train our network with 50 epochs. Experiment is executed on four Nvidia RTX 3090 GPUs.

Comparative Experiments. Since our work is the first to utilize patient narratives as an auxiliary modality for image-based disease classification, meanwhile,

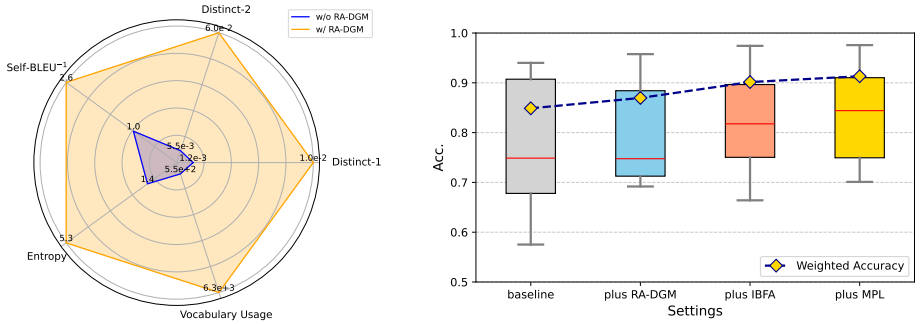
[1] https://challenge.isic-archive.com/landing/2019/
[2] https://openai.com/api/

Table 1: Comparison of Different Methods on Two Datasets with Improvements

Experiment on HAM10000 Dataset											
Method	**Acc**	**F1**	**MEL**	**NV**	**BCC**	**AK**	**BKL**	**DF**	**VASC**	**SCC**	**Improv.**
Resnet50 [11]	0.8915	0.8911	0.6552	0.9763	0.8600	0.5714	0.7328	0.6875	0.8320	-	+3.15%
w/ ours	0.9230	0.9490	0.7034	0.9824	0.8979	0.7200	0.8384	0.8260	0.7812	-	
DaViT [5]	0.8740	0.8720	0.6453	0.9664	0.7449	0.6533	0.6332	0.6957	0.9062	-	+3.20%
w/ ours	0.9060	0.9053	0.6667	0.9704	0.8600	0.6571	0.8190	0.7500	0.9500	-	
HiFuse [15]	0.8540	0.8499	0.5407	0.9555	0.7449	0.6133	0.6245	0.6522	0.875	-	+4.60%
w/ ours	0.9000	0.8751	0.7703	0.9023	0.9259	0.9642	0.9432	0.9047	0.8928	-	
MedMamba [25]	0.8940	0.8814	0.3448	0.9911	0.8800	0.8857	0.7856	0.7500	0.9500	-	+3.00%
w/ ours	0.9240	0.9198	0.7500	0.9788	0.8571	0.7066	0.8384	0.7826	0.9687	-	
ConvNeXt [19]	0.9065	0.9059	0.5747	0.9719	0.9000	0.8286	0.8190	0.8750	0.9500	-	+2.30%
w/ ours	0.9295	0.8899	0.7707	0.9524	0.9021	0.857	0.9090	0.9523	0.9227	-	
Experiment on ISIC2019 Dataset											
Method	**Acc**	**F1**	**MEL**	**NV**	**BCC**	**AK**	**BKL**	**DF**	**VASC**	**SCC**	**Improv.**
Resnet50 [11]	0.8488	0.8488	0.7580	0.9188	0.9034	0.5783	0.7395	0.7111	0.9400	0.5752	+6.44%
w/ ours	0.9132	0.9116	0.8527	0.9757	0.9083	0.7617	0.8355	0.7010	0.9162	0.71264	
DaViT [5]	0.8934	0.8917	0.7947	0.9589	0.9271	0.7283	0.8069	0.6957	0.9355	0.7288	+3.40%
w/ ours	0.9274	0.9263	0.8613	0.9761	0.9484	0.7774	0.8860	0.7783	0.9359	0.7339	
HiFuse [15]	0.8741	0.8740	0.7859	0.9410	0.8480	0.7283	0.8340	0.7391	0.9335	0.6610	+4.17%
w/ ours	0.9158	0.914	0.8585	0.9766	0.8821	0.7888	0.8584	0.7886	0.8620	0.7475	
MedMamba [25]	0.8867	0.8847	0.7439	0.9581	0.9331	0.7717	0.8263	0.7826	0.9032	0.6441	+3.86%
w/ ours	0.9253	0.9290	0.8571	0.9706	0.9208	0.8288	0.9061	0.8195	0.8916	0.7825	
ConvNeXt [19]	0.9061	0.9045	0.8366	0.9697	0.9301	0.8043	0.7683	0.6522	0.9355	0.7627	+3.82%
w/ ours	0.9443	0.9430	0.8807	0.9877	0.9568	0.8316	0.8865	0.8865	0.9458	0.8660	

our method does not specify a particular image encoder, and thus, it can be considered as a plug-and-play module applicable to any existing state-of-the-art image classification method, such as Resnet50 [11], DaViT [5] , HiFuse [15], MedMamba [25] and ConvNeXt [19]. The effectiveness of our proposed method has been demonstrated through experiments on two datasets as shown in Tab. 1. For the HAM10000 dataset, our method achieves a top accuracy of 0.9295. Notably, the HiFuse method demonstrates a 4.6% improvement after integrating our module. On the ISIC2019 dataset, our approach attains a maximum accuracy of 0.9445, while ResNet50 shows a significant 6.44% enhancement when equipped with our module. Furthermore, substantial improvements are observed in both the F1-score and classification accuracy across individual disease categories.

Ablation study. The ablation experiments conducted on the ISIC dataset further demonstrate the effectiveness of our proposed method. First, we show that our RANGM module generates more diverse textual narratives. As illustrated in Fig. 4(a), compared to the baseline without RANGM, the samples generated with RANGM exhibit significant improvements across evaluation metrics including Distinct-N, Vocabulary Usage, and Self-BLEU. We have also randomly selected some samples, as shown in Fig. 5, and found that after applying RANGM, the generated patient narratives align more closely with clinical scenarios, and the relevant knowledge reflected is more accurate. Additionally, separate ablation studies on individual modules validate their contributions. Using ResNet-50 as the image encoder, the baseline achieves a Top-1 accuracy of 0.8488 without any



(a) Comparison of Text Generation Methods with and without RANGM

(b) Comparison of Text Generation Methods with and without our proposed module.

Fig. 4: Ablation experiment results. (a) After employing our RANGM, the metrics used to evaluate the diversity of model generation have shown significant improvement. (b) The box plot indicates the accuracy rates of various categories under different settings, with the line representing the change in accuracy across all samples. As our designs are sequentially incorporated, the classification performance continuously improves, which verifies the effectiveness of our method.

specialized design. After integrating the RANGM module (which supplements patient narratives), the accuracy increases to 0.8696, demonstrating that incorporating non-visual information (e.g., pain symptoms, lifestyle habits) through clinical narratives effectively enhances classification performance. Further replacing conventional cross-attention with our IBFA module for modality fusion elevates accuracy to 0.9015, confirming that IBFA's dual attention mechanism (modeling both intra-modal and inter-modal relationships) enables more effective information aggregation. Finally, employing the MP-Loss loss to strengthen supervision on ambiguous single-modality prediction distrubution samples let accuracy reaches 0.9132. Meanwhile, as shown in Fig. 4(b), the box plot analysis reveals improvements in both per-class accuracy and median accuracy, indicating MP-Loss's capability to better supervise challenging samples.

4 Conclusion

For the first time, we have integrated patient narratives with lesion images for disease diagnosis. We have developed an efficient method for generating patient narratives and designed a plug-and-play MIBF-Net and MP-Loss loss, which can assist existing image-only methods in understanding patient narratives and achieving diagnoses that align with clinical scenarios.

Acknowledgment. The research is funded by The National Key Research and Development Program Inter-governmental Special Project for International Science and Technology Innovation Cooperation (2022YFE0112500), Shenzhen

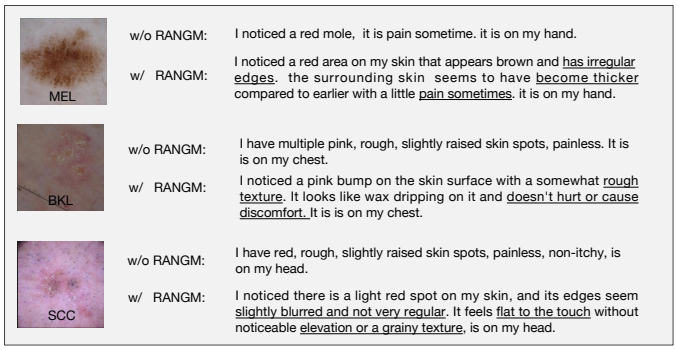


Fig. 5: We randomly selected three samples to demonstrate the improvement in effectiveness brought by using RANGM. After applying RANGM, the model's output becomes more realistic, includes more details, and better aligns with clinical scenarios.

Science and Technology Program (JCYJ20240813151224032, JCYJ20240813151102004), and Shenzhen Medical Research Fund (B2402030).

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

References

1. Muhammad Adeel Azam, Khan Bahadar Khan, Sana Salahuddin, Eid Rehman, Sajid Ali Khan, Muhammad Attique Khan, Seifedine Kadry, and Amir H Gandomi. A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Computers in biology and medicine*, 144:105253, 2022.
2. Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*, pages 354–372. PMLR, 2021.
3. Yin Dai, Yifan Gao, and Fayu Liu. Transmed: Transformers advance multi-modal medical image classification. *Diagnostics*, 11(8):1384, 2021.
4. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.
5. Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *European conference on computer vision*, pages 74–92. Springer, 2022.
6. Jiao Du, Weisheng Li, Ke Lu, and Bin Xiao. An overview of multi-modal medical image fusion. *Neurocomputing*, 215:3–20, 2016.
7. Junwei Duan, Jiaqi Xiong, Yinghui Li, and Weiping Ding. Deep learning based multimodal biomedical data fusion: An overview and comparative review. *Information Fusion*, page 102536, 2024.

8. Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864, 2020.
9. Rachel Grob, Mark Schlesinger, Lacey Rose Barre, Naomi Bardach, Tara Lagu, Dale Shaller, Andrew M Parker, Steven C Martino, Melissa L Finucane, Jennifer L Cerully, et al. What words convey: the potential for patient narratives to inform quality improvement. *The Milbank Quarterly*, 97(1):176–227, 2019.
10. Zhe Guo, Xiang Li, Heng Huang, Ning Guo, and Quanzheng Li. Medical image segmentation based on multi-modal convolutional neural network: Study on image fusion schemes. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 903–907. IEEE, 2018.
11. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
12. Monica Hessler, Elmira Jalilian, Qiuyun Xu, Shriya Reddy, Luke Horton, Kenneth Elkin, Rayyan Manwar, Maria Tsoukas, Darius Mehregan, and Kamran Avanaki. Melanoma biomarkers and their potential application for in vivo diagnostic imaging modalities. *International journal of molecular sciences*, 21(24):9583, 2020.
13. Bing Huang, Feng Yang, Mengxiao Yin, Xiaoying Mo, and Cheng Zhong. A review of multimodal medical image fusion techniques. *Computational and mathematical methods in medicine*, 2020(1):8279342, 2020.
14. Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019.
15. Xiangzuo Huo, Gang Sun, Shengwei Tian, Yan Wang, Long Yu, Jun Long, Wendong Zhang, and Aolun Li. Hifuse: Hierarchical multi-scale feature fusion network for medical image classification. *Biomedical Signal Processing and Control*, 87:105534, 2024.
16. Catherine Hutchison and May McCreaddie. The process of developing audiovisual patient information: challenges and opportunities. *Journal of clinical nursing*, 16(11):2047–2055, 2007.
17. Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.
18. Chunyu Liu, Yixiao Jin, Zhouyu Guan, Tingyao Li, Yiming Qin, Bo Qian, Zehua Jiang, Yilan Wu, Xiangning Wang, Ying Feng Zheng, et al. Visual–language foundation models in medicine. *The Visual Computer*, pages 1–20, 2024.
19. Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.
20. Thomas E Matthews, Ivan R Piletic, M Angelica Selim, Mary Jane Simpson, and Warren S Warren. Pump-probe imaging differentiates melanoma from melanocytic nevi. *Science translational medicine*, 3(71):71ra15–71ra15, 2011.
21. Mojdeh Rastgoo, Rafael Garcia, Olivier Morel, and Franck Marzani. Automatic differentiation of melanoma from dysplastic nevi. *Computerized Medical Imaging and Graphics*, 43:44–52, 2015.
22. Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.
23. Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*, 56(3):1–52, 2023.

24. Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876, 2022.
25. Yubiao Yue and Zhenzhang Li. Medmamba: Vision mamba for medical image classification. *arXiv preprint arXiv:2403.03849*, 2024.
26. Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020.
27. Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37, 2023.
28. Ruicheng Zhang, Haowei Guo, Zeyu Zhang, Puxin Yan, and Shen Zhao. Gamedsnake: Gradient-aware adaptive momentum evolution deep snake model for multi-organ segmentation, 2025.
29. Ruicheng Zhang, Yu Sun, Zeyu Zhang, Jinai Li, Xiaofan Liu, Au Hoi Fan, Haowei Guo, and Puxin Yan. Marl-mambacontour: Unleashing multi-agent deep reinforcement learning for active contour optimization in medical image segmentation, 2025.
30. Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. *ACM computing surveys*, 56(9):1–36, 2024.