Engineering xxx (xxxx) xxx

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence—Article

Learning Conceptual Text Prompts from Visual Regions of Interest for Medical Image Segmentation

Zhu He [a], Haoran Zhang [a], Wentao Zhang [b], Shen Zhao [c], Qiqi Liu [d], Xiaohu Wu [e,*], Qicheng Lao [a,*]

[a] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China
[b] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China
[c] School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518107, China
[d] School of Engineering, Westlake University, Hangzhou 310030, China
[e] National Engineering Research Center of Mobile Network Technologies, Beijing University of Posts and Telecommunications, Beijing 100876, China

ARTICLE INFO

Article history:
Received 21 March 2025
Revised 6 February 2026
Accepted 7 April 2026

Keywords:
Conceptual text
Prompt learning
Knowledge distillation
Medical image segmentation

ABSTRACT

Vision–language segmentation models (VLSMs) are effective in medical image segmentation tasks. However, a major limitation of these models is their dependence on manually crafted textual inputs. Studies have used visual question answering to semiautomatically generate textual information. However, these methods encounter challenges such as error accumulation. Herein, we propose a method to learn conceptual text prompts directly from visual regions of interest (ROIs) for facilitating medical image segmentation. We extracted textual conceptual attributes from ROIs using a large multimodal model to derive coarse real-text prompts. A text latent space transformation module accepted the ROI images as input for generating fine-grained pseudo-text prompts to compensate for the lack of image detail perception in the abovementioned real-text prompts. These prompts were encoded into a unified text embedding. Thereafter, we applied a self-adding noise knowledge distillation method to transfer the knowledge from text embedding to the class token of the image encoder, enabling direct text-guided inference during testing while reducing error accumulation. Our approach minimized the need for manual prompt design by leveraging explicit discrete and implicit continuous text prompts to effectively guide visual segmentation. Extensive evaluation across 13 medical image segmentation datasets demonstrated that our model outperformed the state-of-the-art VLSMs and vision-based segmentation models, exhibiting superior segmentation accuracy.

1. Introduction

The recent rapid development of vision–language models (VLMs) has significantly enhanced the integration of information between visual and textual modalities, improving the effectiveness of information transfer between these modalities [1,2]. Researchers have improved the generalization capabilities of downstream image classification models by leveraging the robust summarization power of textual information, based on traditional visual models that rely on image feature extraction [3,4]. Subsequently, the application of VLMs has been extended from image-level recognition tasks (e.g., classification) to pixel-level recognition tasks (e.g., detection [5] and segmentation [6]), allowing highly precise object recognition facilitated by language guidance [7].

Medical image segmentation is a relatively complex visual recognition task [8]. Textual information from medical diagnostic reports can offer insights into the size, shape, location, and number of regions of interest (ROIs) in medical images [9,10], supporting vision–language segmentation models (VLSMs) in performing medical image segmentation tasks [11]. This textual information serves as explicit prior knowledge, constraining the focus of VLSMs on lesion areas and significantly correcting erroneous attention to irrelevant regions, improving the segmentation performance [12].

However, the application of VLSMs to medical image segmentation tasks presents significant challenges [13,14]. Fig. 1 shows three identified primary challenges: ① text labeling-based VLSMs require image-level text labels to effectively aid image segmentation tasks [15]. Alternatively, these models need to be trained on extensive data annotated with text and images to effectively

* Corresponding authors.
E-mail addresses: wuxhu11@gmail.com (X. Wu), qicheng.lao@bupt.edu.cn (Q. Lao).

https://doi.org/10.1016/j.eng.2026.04.006

Please cite this article as: Z. He, H. Zhang, W. Zhang et al., Learning Conceptual Text Prompts from Visual Regions of Interest for Medical Image Segmentation, Engineering, https://doi.org/10.1016/j.eng.2026.04.006

generalize similar data types without text labels [16]. In the absence of textual information, text labeling-based VLSMs revert to purely visual models and underperform as compared to visual segmentation models specifically designed for downstream visual tasks [17]. ② Vision question answering (VQA)-based VLSMs are used for generating textual descriptions from medical images, which is a common approach to mitigate the lack of textual labels [18]. However, this approach heavily relies on the ability of the VQA model to accurately interpret medical images [19] because any deviation in the generated text prompts might directly mislead visual segmentation and increase the likelihood of error accumulation [20,21]. In particular, during the inference phase, erroneous text prompts generated by the VQA method directly from the original images can mislead its reasoning ability [22]. ③ Prompt-based segmentation models are interactive models such as the segment anything model [23], using points, boxes, or masks as prompts. However, they require manual input of prompt information for each image during the testing phase.

In addition to the three aforementioned challenges, textual labels typically use concise language for describing a single type of lesion. Thus, the required length of textual descriptions increases exponentially for multiclass segmentation tasks. The difficulty in extracting information from long textual descriptions is another major limitation of VLSMs [24], further complicating their application in multiclass medical image segmentation tasks. Moreover, because of the discrete nature of semantic information, text descriptions can provide coarse representations of ROIs such as shape descriptions (e.g., circular or elliptical) and positional

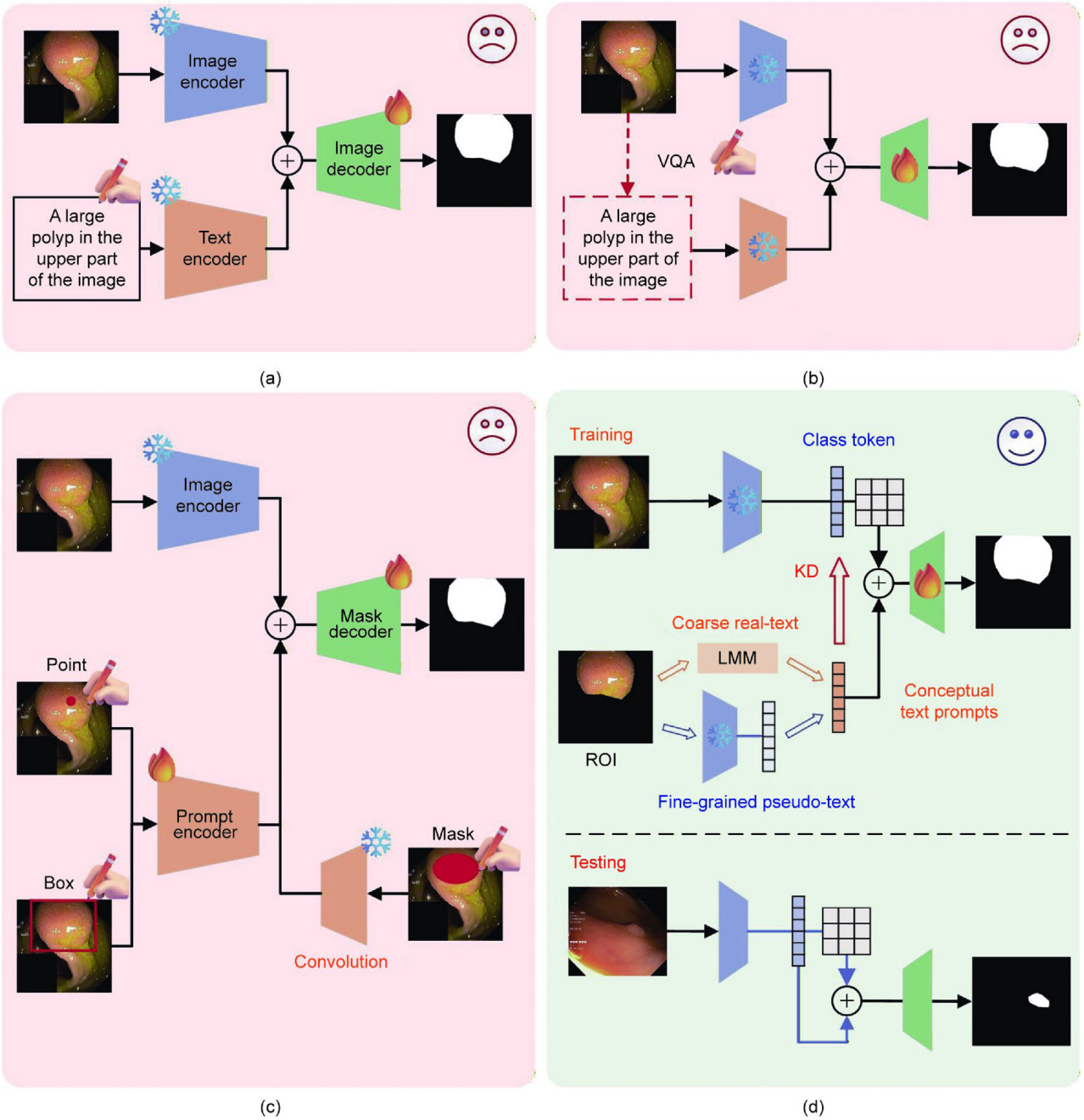

Fig. 1. Comparison of different prompt-based segmentation models: (a) text labeling-based VLSM, (b) vision question answering (VQA)-based VLSM, (c) prompt-based segmentation model, and (d) the proposed model. The proposed model learns conceptual text prompts from regions of interest (ROIs) during the training phase and applies these prompts during the testing phase through knowledge distillation (KD), minimizing manual intervention. LMM: large multimodal model.

information (e.g., upper left or lower right). This coarse representation lacks the precise ROI details required for the fine-grained task of medical image segmentation.

To address the aforementioned issues, we used large multimodal models (LMMs) to directly extract a conceptual text repository from images containing only medical ROIs. As LMMs align images and text within a shared semantic space [25,26], they can accurately extract conceptual attributes from ROIs. Thereafter, we combined these textual concepts to generate concise yet information-rich prompt texts. Using the vision–language alignment capability of VLSMs, we transferred the prompt information into the pure visual space, ensuring the effectiveness and efficiency of the model during the inference phase.

Fig. 1(d) illustrates the methodology used in this study. First, we used an LMM to extract text descriptors from ROIs and construct coarse real-text prompts. This approach preserves the powerful text generation capabilities of the LMM while ensuring the validity of the extracted text concepts. Second, we used a text latent space transformation module for extracting fine-grained pseudo-text prompts from ROIs. This additional information compensates for the absence of detailed visual perception in commonly used real-text prompts. Finally, we integrated the information from both prompts and developed a self-added noise knowledge distillation (SAN-KD) method to transfer the integrated information to the visual class ([CLS]) token, while reducing error accumulation during the distillation process [27]. Thus, the model can be used during the testing phase without requiring real ROI information.

Our proposed method uses an LMM for extracting general textual concepts from ROIs, producing more standardized and accurate text outputs than traditional VQA methods. Moreover, the fine-grained pseudo-text prompts address the subtle visual differences that real text cannot capture and encode ROI information for individual categories as a single token, enabling the segmentation task to be represented using only one token per class. This approach mitigates the challenges of long-text encoding in multiclass medical image segmentation tasks.

Our main contributions are summarized as shown below.

(1) The proposed method learns conceptual text prompts from ROIs for medical image segmentation tasks. This method integrates coarse real-text prompts and fine-grained pseudo-text prompts, effectively leveraging prior textual knowledge, preserving fine-grained visual details, and minimizing the need for manual intervention.

(2) A joint learning strategy for conceptual text knowledge distillation (KD) and visual segmentation is developed. The applicability and generalization of the class token's text-aligned information during the testing phase are enhanced through the proposed SAN-KD method.

(3) Extensive experiments across 13 diverse medical image datasets demonstrate that the learned pseudo-text prompts can effectively assist various medical image segmentation tasks, achieving higher state-of-the-art performance than those of current prevailing VLSMs.

2. Related work

2.1. Vision language segmentation models

The advent of VLMs trained on large-scale paired image–text datasets, such as a large-scale image and noisy-text embedding (ALIGN) [3] and contrastive language–image pretraining (CLIP) [1], has considerably bridged the gap between visual and textual modalities. VLSMs, built using these models as pre-trained encoders, harness the rich semantic information in textual content. This enables them to achieve strong recognition performance on unseen visual categories and categories outside the training data distribution [28]. Following research on natural images, numerous VLMs have also been proposed for medical image segmentation tasks [29,30]. However, the training of VLSMs relies on large-scale image–mask–text triplets, thereby exacerbating the difficulty of annotating medical images [31]. Furthermore, because of substantial differences among medical image modalities, the VLSMs trained on a single dataset with segmentation masks and the corresponding textual descriptions are difficult to generalize to other modalities or even different datasets within the same modality [32]. Therefore, relatively robust and generalized methods are required for applying VLSMs to a broader range of medical image segmentation tasks.

2.2. Text prompt generation

Automatic generation of text prompts for target objects in medical images is an effective strategy to compensate for the absence of text modality, enabling the use of VLSMs for medical image segmentation [33,34]. Tomar et al. [35] enabled the proposed network to learn additional information from a predefined text prompt by introducing an auxiliary classification task. Poudel et al. [36] processed segmentation masks, utilized VQA, and extracted information from online medical journals to collect 14 attribute types related to the ROIs, including category name, shape, color, size, number, and location. They developed nine prompting methods from these attributes to investigate the effects of various types of prompts on the performance of VLSMs. The experimental results of the above studies demonstrated that while the VLSM outperforms the unimodal visual model on multiple datasets, the effects of different text prompts vary considerably across datasets. The optimal prompting method depends heavily on the downstream task and the particular VLSM used, making it difficult to draw general conclusions.

The generation of effective text prompts necessitates careful consideration of the intrinsic characteristics of medical images to provide attribute information that can effectively aid the segmentation task [37,38]. Moreover, to ensure the effectiveness of a text prompt, segmentation masks are required to generate text labels for both the training and testing sets. This approach is inherently inequitable to unimodal visual segmentation models, which do not use mask information during the testing phase. Furthermore, generating a text prompt to compensate for the absence of text modality entails considerable manual effort. The variability in prompt design can substantially affect model performance, thereby rendering this method unreliable.

2.3. VLSM for multiclass segmentation

Utilizing VLSMs for multiclass medical image segmentation requires lengthy prompt texts to provide textual assistance for each segmentation category. However, many VLMs face difficulties in extracting information from long texts. For instance, CLIP has a maximum sequence length limit (default of 77 tokens), with the practically effective length frequently limited to fewer than 20 tokens [24]. Similarly, the language meets vision transformer (LViT) model [39] restricts the input to ten tokens for the text encoder.

Although the self-attention mechanism can perceive information from all input tokens [40], the limitation on sequence length results in a loss of substantial information in the long text [41]. However, the length of the prompt text typically increases exponentially with the number of categories to be segmented, making it challenging to provide effective text prompts for multiclass segmentation tasks, consequently necessitating further research.

3. Method

In this section, we first outline the implementation of learning conceptual text prompts from ROIs, followed by the procedure for transferring these prompts to the CLS token using the proposed SAN-KD method. In the test phase, the CLS token replaces the entire text module for generating predicted embeddings when masks are absent. Next, we detail the approach to the joint learning of conceptual text prompts and medical image segmentation. The detailed architecture of our proposed model is illustrated in Fig. 2.

3.1. Learning conceptual text prompts

We designed our conceptual text prompt network (CoTexNet) model based on the CLIP segmentation (CLIPSeg) architecture [6], leveraging the visual and textual encoders of the CLIP model. Leveraging the pretraining of the CLIP model on a large number of image–text pairs, our model effectively preserves the image–text matching capability within the segmentation framework.

Given a medical image I, let n denote the number of categories to be segmented. In the training phase, we cover I with a mask $M_i (i = 1, 2, \ldots, n)$ corresponding to each segmentation category, producing ROI images R_i, which can be expressed as Eq. (1).

$$\tilde{R} = I \otimes (M_1, M_2, \ldots, M_n) \tag{1}$$

where $\tilde{R} = (R_1, R_2, \ldots, R_n)$ and $\otimes$ denotes element-wise multiplication. For each category i, we used R_i and M_i to generate attribute descriptors a_i^j using the LMM and input prompt P (Eq. (2)).

$$a_i^j = \mathrm{LMM}(P, \{M_i, R_i\}) \tag{2}$$

where j represents the number of attributes.

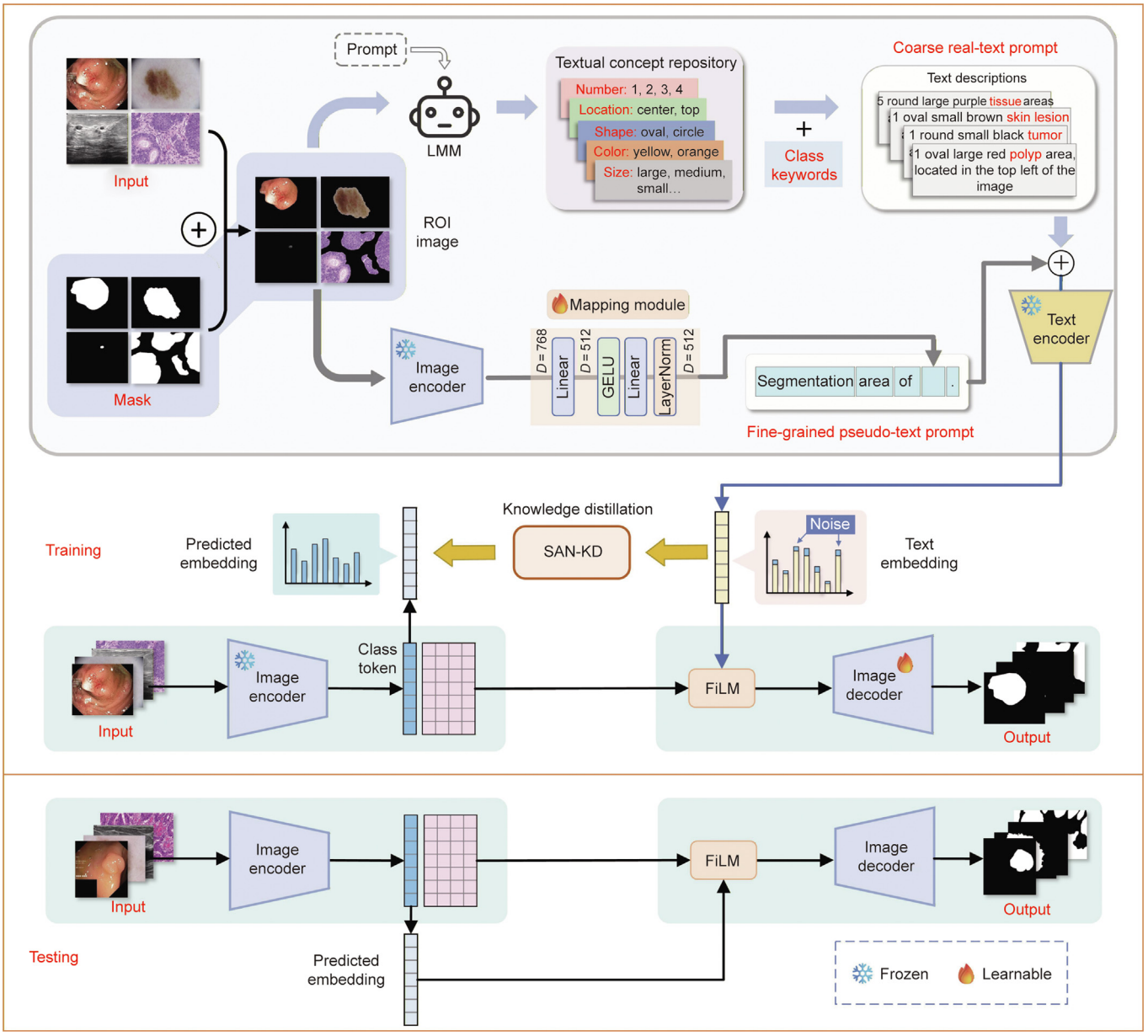

Fig. 2. Detailed structure of the CoTexNet model. The textual concept repository is generated by inputting ROI information and prompts into the LMM. During training, the model transfers textual information to the predicted embeddings derived from the class token using the proposed SAN-KD method, while the predicted embeddings are directly used for inference during testing. GELU: Gaussian error linear unit; FiLM: featurewise linear modulation; D: dimension.

Herein, we used the large language and vision assistant (LLaVA)-v1.5-13B model [42,43], which exhibits excellent vision–language understanding, to extract a_i^j from ROIs. Fig. 3 illustrates the prompt used as input into LLaVA-v1.5-13B and an example of the generated attribute descriptors. a_i^j is integrated with dataset-specific keywords using a prompt template: "[number] [shape] [size] [color] [keywords] area, located in the [location] of the image." This process produces a coarse real-text prompt, denoted as T_r. The generated prompt might be: "1 oval large red polyp area, located in the top left of the image." Similar to VQA methods, the dataset-specific keywords in the prompt template are manually provided. This design aligns with the semantic localization capabilities of VLSMs, guiding the precise segmentation of the target region.

Each R_i is fed into the image encoder and processed through a mapping module to obtain the corresponding output vectors. These vectors are concatenated with the prompt template to generate fine-grained pseudo-text prompts T_p, which can be expressed as Eq. (3).

$$T_p = t_0 \oplus \phi\left(\widetilde{R}\right) \tag{3}$$

where t_0 denotes the prompt template and $\oplus$ denotes the concatenation operation. We used "segmentation area of" as the text of the prompt template in our CoTexNet model, replacing the default "a photo of" used in the CLIP model. $\phi(\cdot)$ represents the image encoder and mapping module. Subsequently, T_p and T_r are provided as input into the text encoder E_T to obtain the final output text embedding T_e (Eq. (4)).

$$T_e = E_T(T_p \oplus T_r) \tag{4}$$

In summary, we designed a mapping module that projected image embeddings containing ROI information into the language embedding space, enabling visual features to be approximately aligned with the predefined textual prompts and establishing a semantic correspondence between coarse real-text prompts and fine-grained pseudo-text prompts. The module comprises two linear layers with Gaussian error linear unit (GELU) activation inserted between them, followed by a layer normalization layer to stabilize the output feature distribution. This design follows the projection concept commonly adopted in multimodal representation learning frameworks such as CLIP [1] and ALIGN [4], while incorporating additional nonlinearity and normalization to stabilize the feature distribution and improve crossmodal alignment.

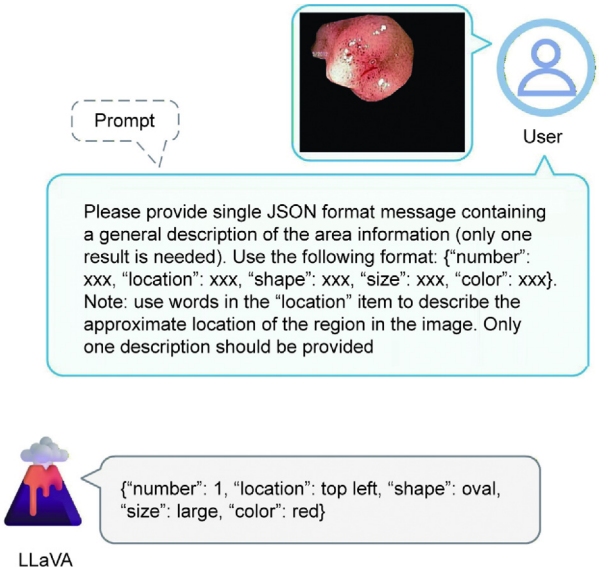

Fig. 3. Generation of conceptual textual prompts using the LLaVA model. JSON: JavaScript object notation.

In T_p, each input R_i occupies one token, with the number of R_i equal to the number of categories to be segmented (n). Together with three tokens for t_0 and two special tokens ([SOS] and [EOS]) used by the text encoder, this results in a total of n + 5 tokens. Although the CLIP model has fewer than 20 valid tokens, this suffices for T_p in CoTexNet to encode the ROI of each segmentation category, as typical medical image segmentation tasks rarely exceed 15 categories and often involve only one or two. Therefore, in addition to compensating for the missing image detail information in the coarse real-text prompt, another important function of T_p is to incorporate all ROI-related image auxiliary information using fewer tokens, preventing the loss of prompt information for certain categories.

In the image encoding module of the vision part, I is processed by the image encoder E_I to obtain feature maps $F^s(s = 1, 2, \ldots, l)$ at different stages, with the CLS token C. l denotes the number of stages in the image encoder. Subsequently, F^l from the last stage and T_e are taken as inputs to the featurewise linear modulation (FiLM) module [44] for feature fusion. The process can be expressed using Eqs. (5) and (6).

$$C, F^s = E_I(I) \tag{5}$$

$$F' = \mathrm{FiLM}\left(F^l + \varphi(T_e)\right) \tag{6}$$

where $\varphi(\cdot)$ represents the Sigmoid function and the inverse Z-score operation. F' in Eq. (6) is subsequently fed into the image decoder, along with $F^s(s = 1, 2, \ldots, l-1)$, constructing the prediction mask M_p. This process can be expressed using Eq. (7).

$$M_p = D_I\left(F' \oplus F^{l-1} \oplus \ldots \oplus F^1\right) \tag{7}$$

where $D_I(\cdot)$ represents the image decoder. During the model training process, we froze the image and text encoders of CoTexNet to preserve the image–text matching properties. The parameters of the nonfrozen components were optimized by minimizing the segmentation loss between M_p (Eq. (7)) and the ground truth mask.

3.2. Self-adding noise for KD

Unlike traditional VQA methods that derive textual prompts directly from the original input image, we designed the ROI images using masks during the training phase and extracted reliable conceptual text prompts to assist downstream visual segmentation. During the testing phase, the absence of mask images prevented the direct application of the aforementioned method. As the ROI is inherently part of the input image, features extracted from the input image inherently contain ROI-specific information [45]. Moreover, the CLIP model, adopted as the backbone, supports bidirectional vision–text alignment.

We developed an SAN-KD method to transfer the knowledge encoded in the conceptual text embeddings to the CLS token C from the image encoder (Eq. (5)) to compensate for the absence of mask information. Furthermore, this method uses an internal self-noise technique to mitigate the impact of distillation errors on downstream visual segmentation.

The mentioned operations were performed on C to obtain the predicted embedding T_c (Eq. (8)).

$$T_c = \mathrm{Sigmoid}(\mathrm{Map}(C)) \tag{8}$$

where $\mathrm{Map}(\cdot)$ denotes the mapping module illustrated in Fig. 2. Subsequently, SAN-KD is used for transferring prompt information from T_e to T_c. First, adding noise from the distillation target data T_c to T_e, we obtain a new distillation target T':

$$T' = \text{Sigmoid}((1-\alpha)\cdot T_e + \alpha \cdot T_c) \tag{9}$$

where α denotes a learnable noise parameter with an initial value of 0.2. The Kullback–Leibler divergence function is used to measure the distillation error between T' and T_c:

$$L_{KL} = -\sum_{q=1}^{d} T'_q \cdot \left(\log T'_q - \log T_{c,q}\right) \tag{10}$$

where L_{KL} is the Kullback–Leibler divergence loss. d denotes the dimension of T_c. T' is adapted by performing an adaptive inverse Z-score operation to adjust their numerical distributions to be suitable as inputs to the FiLM module. The specific process can be expressed as follows:

$$\hat{T} = \text{Dropout}(T' \cdot \sigma^k + \mu^k) \tag{11}$$

where σ^k and μ^k denote learnable parameters with initial values of 1 and –0.002, respectively. $\hat{T}$ is the textual feature representation output by the text module during training. Dropout represents a dropout operation applied to output neurons, with a dropout rate of 0.2. In the test phase, the above parameters are directly applied to T_c to obtain $\hat{T}_c$:

$$\hat{T}_c = T_c \cdot \sigma^k + \mu^k \tag{12}$$

where $\hat{T}_c$ is the textual feature representation obtained during testing. Finally,$\hat{T}_c$ directly replaces $\varphi(T_e)$ in Eq. (6), serving as input to FiLM and facilitating inference during the testing phase.

CoTexNet uses $\hat{T}$ from the text embedding as input to FiLM during training and replaces $\hat{T}$ with $\hat{T}_c$ from the CLS token during testing, eliminating the need to use the ROI information and inferring the text module. However, during optimization of Eq. (10), there is an error for fitting $\hat{T}_c$ to $\hat{T}$, which further widens the performance gap between the training and test sets. Thus, we use two methods to reduce error accumulation—namely, self-adding noise from Eq. (9) and dropout from Eq. (11). The former introduces partial distillation errors to real-text inputs during the training phase to enhance the model's adaptability to these errors during testing, while the latter mitigates the model's overfitting to T_e.

3.3. Joint learning with segmentation

We used two losses: ① a segmentation loss for learning the complete conceptual text prompt and performing the segmentation task, and ② a distillation loss for aligning the real-text embeddings with the predicted embeddings. The specific computational procedure is expressed as follows:

$$W, b, T_e = \arg\min_{W,b} L_{seg}(f(I; W, b, T_e), M) \tag{13}$$

$$W', b' = \arg\min_{W',b'} L_{KL}\left(f'(T_c; W', b'), T_e\right) \tag{14}$$

where $f(\cdot)$ and $f'(\cdot)$ represent the parts of CoTexNet used for segmentation and distillation, respectively.W and W' denote the weight parameters. b and b' indicate the bias terms.L_{seg} denotes the sum of Dice and crossentropy losses. $M = \{M_i\}$ $(i = 1,2,\ldots,n)$ denotes the set of ground truth masks.

Theoretically, the aforementioned optimization processes can be conducted in steps: training on the segmentation task (Eq. (13)) before fitting the conceptual text prompt (Eq. (14)). However, in the testing phase, the model needs to use T_c instead of T_e as the output of the text module, which causes the distillation error in Eq. (10) to directly affect the segmentation performance, resulting in error accumulation. This approach hinders the generalization of the conceptual text prompt, learned during training, to the test set. To further address this issue, we used joint learning for fitting and segmentation. In conjunction with the calculation process of T' in Eq. (9), the joint optimization formula is expressed as follows:

$$W, W', b, b' = \arg\min_{W,W',b,b'} \left[L_{seg}(f(I; W, b, T'), M) + L_{KL}(f'(T_c; W', b'), T_e)\right] \tag{15}$$

During the training process of CoTexNet, the combined embedding T', which integrates T_e and T_c, is used to replace the original T_e as the output of the text module. This approach introduces the distillation error of T_c as additional noise during segmentation task training. This bidirectional interaction affects the performance of the model on the segmentation task based on the degree of T_c fitting, while T_c fitting adjusts as T_e changes. As the segmentation task training stabilizes, the value of T_e ceases to undergo significant changes, allowing T_c to stabilize the fitting task, and ultimately terminating joint training.

4. Materials

4.1. Datasets

We conducted segmentation experiments on 13 medical image segmentation datasets, comprising 11 single-category and two multicategory segmentation datasets. These datasets covered a variety of imaging modalities. In particular, colon polyp datasets—Kvasir [46], ClinicDB [47], and the BKAI-IGH Neopolyp-Small Polyp dataset (BKAI) [48]—were derived from photographic imaging. The International Skin Imaging Collaboration (ISIC) [49] dataset contains dermatology images, while the diabetic foot ulcer (DFU) [50] dataset includes foot lesion images. For ultrasonography, we used the Digital Database Thyroid Image (DDTI) [51] dataset for thyroid lesions and the breast ultrasound image (BUSI) [52] dataset for breast lesions. The gland segmentation in colon histology images challenge contest (GlaS) [53] dataset, which focuses on colon histology, is based on microscopic imaging. We also included two computed tomography (CT)-based datasets: ① municipal medical hospitals in Moscow (MosMed) [54,55] for pulmonary infection segmentation, and ② combined healthy abdominal organ segmentation (CHAOS) [56] for liver segmentation. The Qatar University and Tampere University (QaTa) [57] dataset is based on coronavirus disease 2019 (COVID-19) lesion area segmentation in X-ray imaging. We included two multicategory cardiac segmentation datasets: ① automated cardiac diagnosis challenge (ACDC) [58] based on magnetic resonance imaging (MRI), and ② cardiac acquisitions for Multistructure Ultrasound Segmentation (CAMUS) [59] based on ultrasound imaging.

Table 1 presents the distribution of images across the training, validation, and test sets for these 13 datasets and specifies whether patient identifier (ID) information is included. For datasets containing patient information, all images from a given patient are assigned to only one training, validation, or test set. Datasets without patient IDs and those lacking officially defined test sets are randomly divided into training, validation, and test sets in an 8:1:1 ratio.

4.2. Implementation details

During model training, we used the Adam optimizer with a cosine annealing learning rate schedule. The maximum learning rate was set to 3×10^{-4}, the batch size to 8, and the training process was capped at 500 epochs. Early stopping was applied if the validation loss did not decrease for 50 consecutive epochs. All experiments were executed on two NVIDIA GeForce RTX 4090 graphics processing units (GPUs) with 24 GB of random access memory (RAM).

Table 1
Image splits in the 13 datasets and patient ID information.

Dataset	Class	Train	Validate	Test	Patient ID
Kvasir	1	800	100	100	×
BKAI	1	800	100	100	×
ClinicDB	1	490	61	61	×
BUSI	1	624	78	78	×
DDTI	1	511	63	63	×
DFU	1	854	78	78	×
ISIC	1	810	90	379	×
GLaS	1	100	33	32	√
MosMed	1	2183	273	273	×
CHAOS	1	1884	455	535	√
QaTa	1	5716	1429	2113	×
CAMUS	3	1600	200	200	√
ACDC	3	1124	406	372	√

Class indicates the number of classes of the areas to be segmented. × indicates that patient ID information is not included in the dataset, while √ indicates that it is included.

We used LLaVA-v1.5 during the offline generation of coarse real-text prompts once for each training image, and the generated prompts were cached for future use. Thus, LLaVA did not participate in model training or inference and introduced negligible computational overhead during actual training and deployment stages.

5. Experiments

5.1. Model performance comparison

The proposed model was compared with state-of-the-art medical image segmentation models. These included univisual models such as a completely convolutional UNet (ACC-UNet) [60], channel-wise cross fusion attention and transformer (CFATrans) [61], transformers for fully convolutional denseNets (TFCNs) [62], and nnSAM [63], and VLSMs such as CLIPSeg, CLIP-driven referring image segmentation (CRIS) [64], and LViT. The experimental results are presented in Table 2. For the models requiring text input, three methods were used for generating the text prompt: ① (*P*1) providing only the category label name, ② (*P*2) providing the category label name with a brief one-sentence description, and ③ (*P*3) using the VQA method [65] to extract information such as shape, location, and number from the original image to generate the prompts. As VLSMs (CLIPSeg, CRIS, and LViT) required text input for inference, we used the placeholder template "a photo of something" in text-free experiments, which contained no meaningful information, as the text input to retain the text encoding module. Further details on the generation of the text prompt are provided in Appendix A Section S1.

The proposed CoTexNet model achieves optimal segmentation performance across 11 single-category datasets presented in Tables 2–4, demonstrating the effectiveness of our approach for learning conceptual text prompts from ROIs. Additionally, the suboptimal segmentation results appear for CLIPSeg and LViT models using text prompts, suggesting that VLSMs with text labels have higher potential than univisual models. However, suboptimal results vary with different prompt types without a clear pattern. In certain cases, the choice of prompts significantly affects performance. CLIPSeg exhibits a 17.41% lower Dice score for *P*3 than *P*1 on the GlaS dataset, while the CRIS exhibits an 11.22% lower Dice score for *P*2 than *P*1 on the DFU dataset. This indicates that VLSM performance is sensitive to prompt type, making it challenging to identify a generalized prompting method. Our proposed method adopts the same visual architecture as the CLIPSeg model and effectively supports downstream vision tasks by learning conceptual text labels from ROIs. It achieves the best segmentation performance across all eight datasets, demonstrating the effectiveness and robustness of this approach. We provide additional discussion and experimental results on the computational cost and model size comparisons in Appendix A Section S2.

Table 5 presents a performance comparison of the segmentation model on the CAMUS and ACDC multicategory segmentation datasets. The proposed model consistently outperforms others on these datasets. However, the visual unimodal model CFATrans outperforms the other VLSMs in terms of the dice coefficient (Dice) score on the CAMUS dataset, achieving suboptimal performance. On the ACDC dataset, suboptimal intersection over union (IoU) values are observed across different category segmentations using various VLSMs with different prompts, suggesting that VLSMs with explicit text inputs struggle to effectively assist segmentation tasks for each category, underperforming in multicategory segmentation tasks. In contrast, our proposed method of learning conceptual text prompts from ROIs demonstrates highly comprehensive and superior performance for multicategory segmentation tasks.

In Appendix A Section S3, we further evaluate the robustness of our framework by generating coarse real-text prompts using multiple open-source LMMs, including medical-domain finetuned models (VILA-M3-8B [66] and LLaVA-Med-v1.5-7B [67]) and general-domain models (DeepSeek-VL2-4B [68], Qwen2-VL-7B [69], LLaVA-v1.5-7B [43], and LLaVA-v1.5-13B [43] used in our main experiments). The results consistently show that CoTexNet maintains stable segmentation performance across different LMMs, demonstrating strong robustness to variations in prompt-generation models.

5.2. Visualized segmentation results

We compared the proposed CoTexNet model with the CLIPSeg, CRIS, and LViT models using *P*3 prompts across 11 single-category segmentation datasets to demonstrate the impact of the VQA method on VLSMs (Fig. 4). The text generated by the VQA method directly from medical images can mislead model recognition, resulting in reduced stability. "Rectangle" and "upper" were provided in the DDTI sample, leading the model to incorrectly predict the circular ROI in the upper left corner as a large rectangular region above the image. While there is a discrepancy between the prediction results of the proposed model and mask in this sample, the discrepancy is smaller compared to the other three VLSMs. In the sample from the ClinicDB dataset, the *P*3 prompt incorrectly provides the location keyword "center," causing the CLIPSeg and CRIS to predict no foreground region, while the LViT incorrectly predicts a portion of the center region. These visualization results of our proposed model indicate that segmentation stability outperforms other VLSMs.

Fig. 5 illustrates the results of two multicategory segmentations for the ACDC and CAMUS datasets. Owing to the relatively fixed

Table 2
Performance comparison of the proposed and existing methods on the Kvasir, BKAI, ClinicDB, and BUSI datasets.

Model	Text	Kvasir				BKAI				ClinicDB				BUSI			
		Dice	IoU	HF95	PA	Dice	IoU	HF95	PA	Dice	IoU	HF95	PA	Dice	IoU	HF95	PA
ACC-UNet	×	88.57	81.18	22.84	96.59	83.63	77.98	28.46	97.11	84.56	81.16	19.74	97.56	68.32	63.36	31.47	94.21
CFATrans	×	88.87	81.51	22.14	96.88	86.90	80.16	24.17	98.04	86.99	82.95	16.20	97.87	67.34	61.81	32.07	94.37
TFCNs	×	87.22	80.06	23.61	96.38	81.44	76.68	30.34	97.02	85.65	81.64	17.92	97.51	71.29	66.44	29.37	95.08
nnSAM	×	89.69	82.67	21.83	96.67	88.74	82.16	19.20	98.67	90.61	84.53	14.19	98.23	71.09	65.06	29.86	95.04
CLIPSeg	×	89.13	82.29	22.34	96.58	87.21	81.13	22.71	98.17	91.36	85.24	13.42	98.36	67.09	62.14	33.37	94.56
	*P*1	89.23	82.37	22.12	96.71	87.53	81.42	22.04	98.22	90.86	84.85	13.91	98.19	69.13	64.13	32.02	94.82
	*P*2	89.78	82.85	21.74	96.94	86.29	79.85	22.89	98.13	91.33	85.07	13.57	98.31	68.59	63.54	32.58	94.65
	*P*3	88.62	81.86	22.37	96.81	87.75	81.55	21.48	98.47	91.41	85.36	13.17	98.38	68.73	63.28	32.47	94.66
CRIS	×	85.89	78.89	25.21	95.92	84.06	79.09	23.47	97.66	82.86	75.22	21.81	97.05	67.06	62.11	33.21	94.58
	*P*1	84.87	77.27	26.18	95.67	85.47	79.76	24.18	97.94	88.60	82.62	14.05	98.11	66.17	61.70	33.64	94.49
	*P*2	84.26	77.59	26.48	95.51	81.91	77.07	29.34	97.04	85.61	79.75	17.68	97.64	65.47	60.64	34.27	94.28
	*P*3	87.65	80.60	23.27	96.54	83.88	78.56	28.34	97.48	86.85	79.05	16.28	97.81	64.64	59.32	35.81	94.02
LViT	×	86.29	80.92	24.28	96.11	88.71	82.56	19.34	98.61	85.15	82.60	18.05	97.47	68.48	62.13	32.17	94.58
	*P*1	86.37	80.18	24.38	96.12	89.34	83.23	18.11	98.87	85.70	82.88	17.34	97.69	67.56	59.22	33.57	94.21
	*P*2	85.81	79.59	25.43	95.84	87.05	81.18	22.14	98.24	86.63	83.62	16.57	97.84	67.93	61.48	32.67	94.34
	*P*3	86.30	80.50	23.81	96.08	86.78	80.35	23.57	98.17	88.03	84.05	14.28	98.04	71.66	66.58	29.94	95.18
Ours	—	**91.32**	**84.93**	**20.42**	**97.37**	**90.66**	**84.02**	**16.43**	**99.08**	**92.64**	**87.62**	**12.56**	**98.54**	**76.07**	**69.13**	**28.43**	**95.61**

Four evaluation metrics are given for each dataset: dice coefficient (Dice), intersection over union (IoU), 95% Hausdorff distance (HF95), and pixel accuracy (PA). The best results are highlighted in bold, while the second-best results are underlined. × denotes that no textual information is used during training and testing, while — denotes that textual information is used during training only.

Table 3
Performance comparison of the proposed and existing methods on the DDTI, DFU, ISIC, and GlaS datasets.

Model	Text	DDTI				DFU				ISIC				GLaS			
		Dice	IoU	HF95	PA	Dice	IoU	HF95	PA	Dice	IoU	HF95	PA	Dice	IoU	HF95	PA
ACC-UNet	×	68.56	61.11	33.52	82.83	77.72	66.78	9.03	98.98	88.21	80.93	17.58	94.34	74.72	60.16	38.71	73.49
CFATrans	×	74.07	66.93	29.82	94.12	76.26	65.49	9.22	98.85	89.31	80.77	13.51	94.48	76.58	62.24	37.47	75.23
TFCNs	×	75.34	67.04	28.79	74.33	69.39	58.68	11.67	98.41	87.31	79.75	15.33	93.85	69.83	56.75	39.82	66.97
nnSAM	×	75.75	69.50	28.46	74.35	77.20	66.72	8.98	98.92	86.50	78.57	15.79	93.02	86.41	76.91	29.01	83.84
CLIPSeg	×	76.08	68.09	27.61	94.37	78.80	68.10	8.53	99.06	91.52	84.35	12.83	95.06	83.01	71.44	33.84	79.98
	*P*1	75.29	67.24	28.16	94.33	78.87	68.26	8.48	99.07	91.60	84.44	12.64	95.11	87.10	77.68	28.33	85.06
	*P*2	76.27	68.34	28.34	94.39	78.41	67.53	8.89	99.06	91.46	84.13	12.77	95.08	84.09	73.22	31.05	81.57
	*P*3	75.19	67.07	28.48	94.31	79.02	68.34	7.57	99.11	91.58	84.30	12.86	95.13	69.69	54.15	40.59	66.89
CRIS	×	74.49	66.13	30.16	94.18	69.26	58.81	11.86	98.36	87.28	79.76	15.23	93.84	61.84	45.42	43.28	59.17
	*P*1	69.60	62.06	33.24	93.01	73.36	63.93	10.07	98.77	82.52	72.68	17.89	91.23	74.53	59.80	38.16	73.82
	*P*2	68.27	60.07	34.19	92.85	62.14	51.21	14.68	97.34	88.70	81.63	14.51	94.30	79.07	66.21	36.21	77.18
	*P*3	69.80	62.19	33.07	93.04	67.15	56.70	12.34	98.27	80.91	71.49	19.57	90.42	82.25	70.55	35.82	79.86
LViT	×	74.23	68.93	30.46	94.15	78.66	68.42	8.21	99.04	88.97	82.40	14.31	94.26	83.91	72.42	35.17	80.24
	*P*1	75.59	69.24	29.67	94.27	78.90	68.73	8.07	99.15	89.53	82.51	13.47	94.67	83.53	72.58	34.59	80.38
	*P*2	72.40	64.64	31.22	93.54	78.73	68.64	8.33	99.12	89.06	82.65	13.82	94.38	85.02	73.87	32.64	82.54
	*P*3	76.13	69.53	28.37	94.38	78.69	68.52	8.48	99.08	88.85	82.37	14.56	94.21	86.41	76.33	29.37	83.91
Ours	—	**77.23**	**70.14**	**26.64**	**94.56**	**81.61**	**70.71**	**6.82**	**99.64**	**92.48**	**85.97**	**12.31**	**95.89**	**88.36**	**78.82**	**26.13**	**86.77**

Table 4
Performance comparison of the proposed and existing methods on the MosMed, CHAOS, and QaTa datasets.

Model	Text	MosMed				CHAOS				QaTa			
		Dice	IoU	HF95	PA	Dice	IoU	HF95	PA	Dice	IoU	HF95	PA
ACC-Unet	×	70.99	65.05	26.88	99.33	82.47	85.62	12.82	99.25	77.87	72.33	35.56	96.69
CFATrans	×	68.86	62.68	27.92	99.27	80.31	84.74	13.21	99.17	77.91	72.69	35.91	96.72
TFCNs	×	64.78	59.27	30.82	99.18	80.14	84.68	16.73	99.04	77.22	67.70	36.72	95.80
nnSAM	×	67.46	63.65	28.51	99.24	79.35	83.95	16.87	99.01	78.66	74.37	36.14	96.98
CLIPSeg	×	72.08	65.82	25.81	99.48	82.08	85.50	12.21	99.22	78.26	73.88	33.17	96.97
	*P*1	72.30	65.87	25.34	99.49	82.16	85.59	12.01	99.26	78.12	73.68	34.22	96.93
	*P*2	72.51	65.92	25.28	99.50	81.96	85.43	12.86	99.21	78.92	74.12	33.15	97.01
	*P*3	69.82	63.57	26.92	99.31	80.36	84.53	13.09	99.18	76.88	73.28	39.65	96.18
CRIS	×	68.35	61.72	27.69	99.25	75.34	82.47	18.35	98.56	77.50	73.12	34.05	96.88
	*P*1	69.52	63.03	27.11	99.29	76.58	83.23	17.26	98.61	77.06	72.25	36.08	96.48
	*P*2	69.63	63.12	26.88	99.30	75.82	82.81	18.02	98.58	77.36	72.53	35.66	96.60
	*P*3	70.58	64.25	25.34	99.36	74.36	82.06	19.12	98.42	78.92	74.51	33.28	97.08
LViT	×	71.36	65.48	26.33	99.41	81.21	85.56	12.55	99.21	80.29	76.17	32.53	97.16
	*P*1	71.41	65.47	26.35	99.40	81.82	85.91	12.06	99.29	80.09	75.82	32.91	97.11
	*P*2	71.82	65.79	26.07	99.43	82.69	86.21	11.76	99.34	80.36	76.21	32.02	97.20
	*P*3	70.58	64.62	26.83	99.37	81.08	85.43	12.62	99.20	79.16	75.62	33.18	97.08
Ours	—	**73.61**	**66.31**	**24.98**	**99.61**	**84.39**	**87.88**	**10.26**	**99.43**	**81.75**	**77.61**	**30.81**	**97.55**

Table 5
Performance comparison of the proposed and existing methods on two multiclass segmentation datasets.

Model	Text	CAMUS						ACDC					
		Dice	IoU-1	IoU-2	IoU-3	HF95	PA	Dice	IoU-1	IoU-2	IoU-3	HF95	PA
ACC-Unet	×	88.48	87.15	73.80	**84.70**	13.42	95.75	82.40	79.93	77.15	90.39	19.85	95.07
CFATrans	×	89.22	88.66	76.38	83.72	12.57	96.04	84.35	82.71	79.47	90.92	17.20	96.23
TFCNs	×	86.29	84.66	68.74	84.10	14.84	94.07	82.82	80.47	78.51	87.85	19.66	95.28
nnSAM	×	89.05	88.35	76.18	83.29	12.58	95.45	84.76	82.85	79.66	90.83	17.47	96.51
CLIPSeg	×	88.46	88.12	72.83	83.84	13.47	95.74	83.98	80.74	77.92	88.78	18.36	96.03
	*P*1	88.54	88.25	73.37	82.28	13.37	95.83	84.76	82.37	78.78	89.32	17.32	96.49
	*P*2	88.86	88.34	74.12	82.86	12.91	95.97	84.96	83.24	79.16	89.62	17.11	96.59
	*P*3	88.67	88.13	75.23	82.15	13.03	95.88	85.37	83.87	78.92	89.48	16.02	96.78
CRIS	×	87.35	87.14	74.84	82.15	14.85	95.10	83.68	80.28	76.79	88.71	18.74	95.94
	*P*1	87.85	87.36	73.13	82.46	14.36	95.34	84.23	81.72	76.85	88.42	17.67	96.18
	*P*2	87.72	87.43	73.14	80.51	14.59	95.22	84.36	81.46	77.33	89.04	17.31	96.25
	*P*3	86.26	86.34	72.24	78.18	15.36	94.02	82.57	80.17	76.55	88.12	19.46	95.17
LViT	×	89.08	88.57	76.10	80.49	12.54	95.93	85.06	83.24	79.54	90.95	16.70	96.68
	*P*1	88.86	88.41	72.97	84.52	13.13	95.81	85.27	83.53	78.66	90.42	16.31	96.71
	*P*2	88.57	87.19	72.63	82.46	13.39	94.96	84.87	83.82	79.05	89.06	17.04	96.53
	*P*3	87.56	86.76	72.67	81.81	14.21	95.07	84.49	82.56	78.82	89.21	17.37	96.34
Ours	–	**90.37**	**89.11**	**76.76**	83.96	**11.36**	**96.34**	**86.21**	**84.36**	**80.47**	**91.54**	**15.24**	**97.15**

location of ROIs, applying the VQA method directly to original images fails to generate distinctive textual prompts on these datasets, highlighting the limitations of traditional VQA methods. In the first row of samples from ACDC, the category of the right ventricular endocardium does not exist. However, this category information in the text prompt misleads some visual models, so that the CLIPSeg and LViT segment this category region. In contrast, our proposed method of automatically learning conceptual text prompts from ROIs is not constrained by predefined categories and can adaptively perceive the differences between ROIs across images, providing robustness and task adaptability.

5.3. Modulewise ablation study

Modulewise ablation studies were conducted on the proposed CoTexNet model, focusing on the coarse real-text prompt, fine-grained pseudo-text prompt, text encoder, and text module. In experiments without the text encoder, the fine-grained pseudo-text prompts were used as text embedding, without converting into the text space and discarding the coarse real-text prompt. In experiments without the text module, the CLS token was directly fed into the FiLM module. The ablation experiments on the single-category segmentation datasets are presented in Tables 6–8. The results indicate that omitting the coarse real-text and fine-grained pseudo-text prompts, the text encoder or text module results in varying degrees of performance degradation across different datasets.

Notably, datasets with regular segmentation targets and clear textures such as the polyp datasets Kvasir, BKAI, ClinicDB, and the ISIC dataset, exhibit the second-best performance in experiments where the fine-grained pseudo-text prompt is omitted. In contrast, datasets with irregular segmentation targets and blurred textures including BUSI, DDTI, DFU, and GlaS datasets exhibit the second-best performance when the coarse real-text prompt is absent. These findings indicate that the fine-grained pseudo-text and coarse real-text prompts complement each other, and their combined use enhances the model's robustness and adaptability in medical image segmentation tasks. Moreover, the segmentation performance in ablation experiments without the text encoder is generally inferior, which is reasonable as this approach cannot leverage VLSM's visual–language matching capability by not converting the prompt information into the text latent space.

Table 9 presents the ablation study results for multicategory segmentation, demonstrating that each module improves the segmentation performance. For each category, the real-text prompts contain approximately 15 words, resulting in about 45 tokens in total, whereas the pseudo-text prompts require only six tokens. Table 9 shows that the second-best results consistently appear in experiments without real-text prompts, indicating that real-text prompts struggle to effectively provide comprehensive information for all categories in multicategory segmentation tasks. In contrast, the performance of the model is more degraded when pseudo-text prompts are absent than when real-text prompts are missing. This suggests that as the number of segmentation categories increases, pseudo-text prompts become more critical than real-text prompts for maintaining segmentation performance. A discussion on additional comparative and ablation experiments on the proposed CoTexNet is presented in Appendix Sections S4–S9.

5.4. Concept-wise ablation study

Based on the study by Poudel et al. [36] on the impact of textual concepts on VLSM segmentation performance, we selected the six most commonly used conceptual attributes (number, shape, size, location, color, and class keywords) for this study. We removed fine-grained pseudo-text prompts and used only coarse real-text prompts to accurately investigate the role of these six conceptual attributes. We performed ablation experiments, and the results are presented in Fig. 6. These results indicate that the segmentation performance of the model generally declines when most concepts are ablated.

Notably, excluding the "number" information declines the performance in seven datasets (ClinicDB, ISIC, BUSI, DDTI, MosMed, CHAOS, and ACDC datasets), while the performance improves in the remaining six datasets. This is because the seven datasets predominantly contain images with a single ROI for each category, making the "number" concept uninformative. When the "shape" attribute is excluded, only the GlaS dataset shows a slight improvement in accuracy, which might be owing to the substantial variability in tissue shapes within the GlaS dataset, making it challenging to describe image shapes effectively using textual information. Therefore, compelling the LMM to generate shape-related descriptive terms might negatively impact the segmentation performance on the GlaS dataset.

When the "size" information is excluded, the segmentation accuracy declines to varying extents across all datasets, indicating that accurately capturing the ROI size information is crucial for

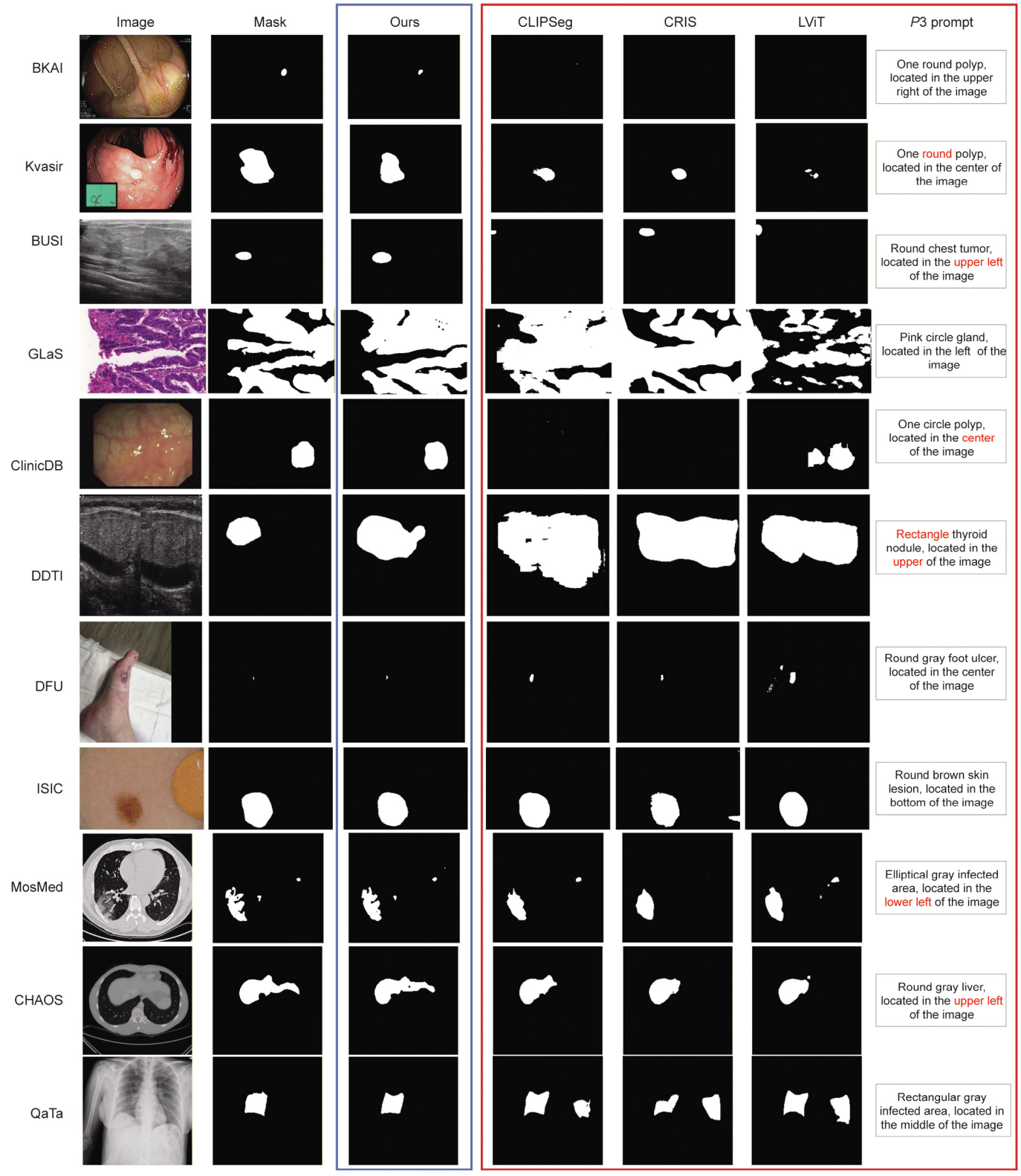

Fig. 4. Visualization of segmentation results on single-category segmentation tasks.

medical image segmentation tasks. Furthermore, for grayscale images (CAMUS, BUSI, DDTI, MosMed, CHAOS, ACDC, and QaTa) datasets, the "color" attribute appears to be less influential, which is also reflected in the experimental results (Fig. 6). When the "location" attribute is ablated, only the ACDC dataset exhibits a slight improvement in accuracy. This is mainly because the ACDC dataset comprises MRI slices where the three segmentation targets (left ventricle, right ventricle, and myocardium) are closely connected, with relatively consistent ROI positions, offering limited discriminatory information. Excluding the class keywords substantially

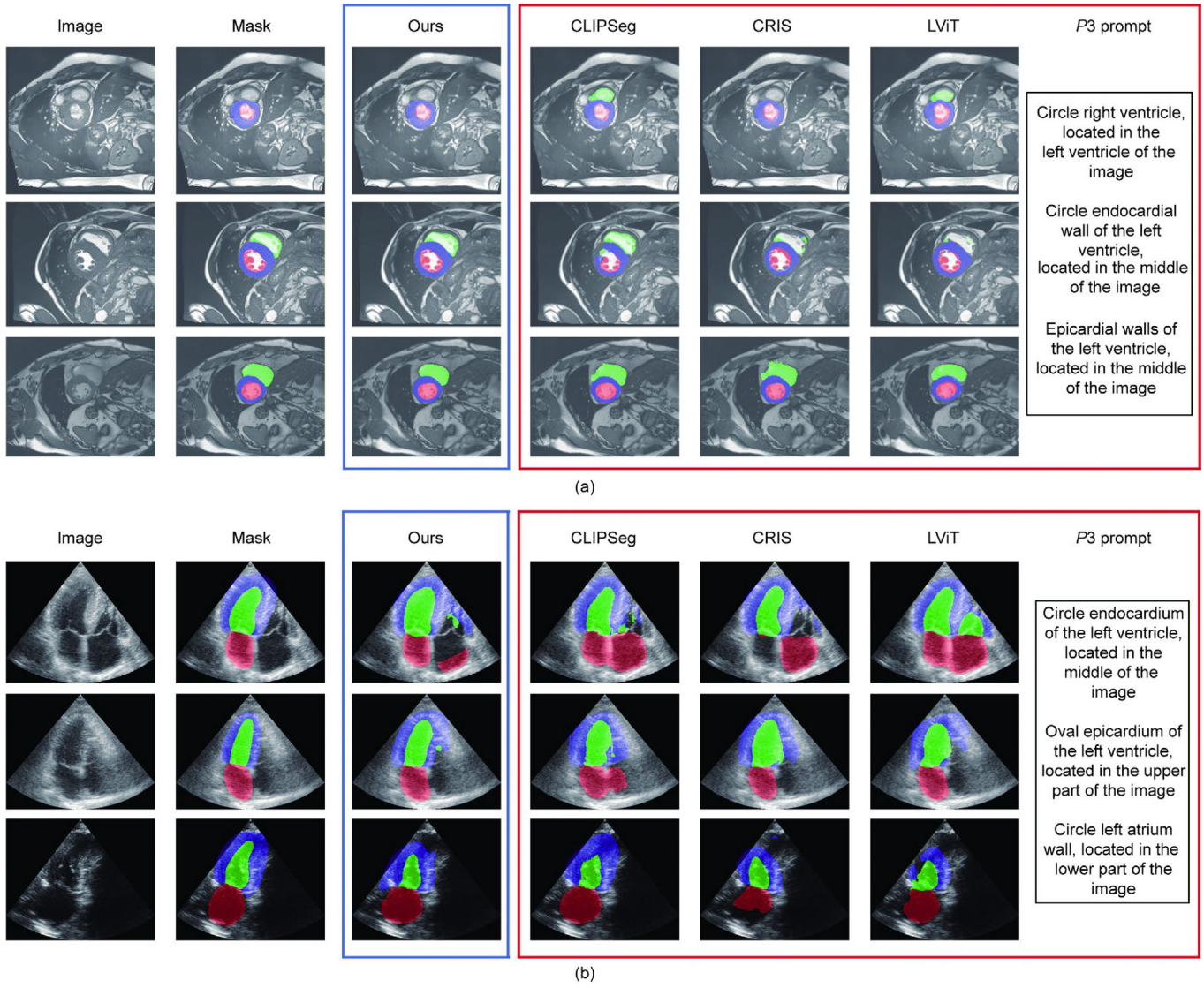

Fig. 5. Visualization of segmentation results on multicategory segmentation tasks: (a) ACDC and (b) CAMUS.

Table 6
Modulewise ablation studies on single-class segmentation datasets: Kvasir, BKAI, ClinicDB, and BUSI datasets.

Real text	Pseudo text	Text encoder	Text module	Kvasir		BKAI		ClinicDB		BUSI	
				Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
√	√	√	√	**91.32**	**84.93**	**90.66**	**84.02**	**92.64**	**87.62**	**76.07**	**69.13**
×	√	√	√	90.46	84.01	90.23	83.76	92.01	87.18	75.13	68.08
√	×	√	√	90.86	84.37	90.38	83.82	92.42	87.51	73.68	67.42
×	√	×	√	89.62	82.76	89.06	82.65	91.56	85.69	73.85	67.63
×	×	×	×	89.24	82.45	88.44	71.32	90.25	84.18	70.25	64.82

√ indicates that the corresponding information or module is used in the model, while × indicates that it is not used.

Table 7
Modulewise ablation studies on single-class segmentation datasets: DDTI, DFU, ISIC, and GlaS datasets.

Real text	Pseudo text	Text encoder	Text module	DDTI		DFU		ISIC		GLaS	
				Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
√	√	√	√	**77.23**	**70.14**	**81.61**	**70.71**	**92.48**	**85.97**	**88.36**	**78.82**
×	√	√	√	74.38	68.27	80.98	69.65	91.11	84.68	88.04	78.47
√	×	√	√	72.33	64.59	80.21	69.08	91.36	84.82	87.36	77.87
×	√	×	√	70.88	62.57	78.67	68.82	90.15	84.27	82.81	71.05
×	×	×	×	64.89	55.24	78.42	68.15	90.21	84.35	81.08	69.51

Table 8
Modulewise ablation studies on single-class segmentation datasets: MosMed, CHAOS, and QaTa datasets.

Real text	Pseudo text	Text encoder	Text module	MosMed		CHAOS		QaTa	
				Dice	IoU	Dice	Dice	IoU	Dice
√	√	√	√	**73.61**	**66.31**	**84.39**	**87.88**	**81.75**	**77.61**
×	√	√	√	72.06	65.80	82.15	85.61	80.34	76.24
√	×	√	√	72.26	65.91	82.27	85.78	80.14	76.04
×	√	×	√	72.14	65.83	82.03	85.53	79.63	75.49
×	×	×	×	71.96	65.78	81.87	85.32	78.15	73.72

Table 9
Modulewise ablation studies on multiclass segmentation datasets.

Real text	Pseudo text	Text encoder	Text module	CAMUS				ACDC			
				Dice	IoU-1	IoU-2	IoU-3	Dice	IoU-1	IoU-2	IoU-3
√	√	√	√	**90.37**	**89.11**	**76.76**	**83.96**	**86.21**	**84.36**	**80.47**	**91.54**
×	√	√	√	89.06	88.72	75.51	83.28	85.12	83.46	79.73	90.62
√	×	√	√	88.63	88.31	73.88	82.62	84.68	82.01	78.10	88.73
×	√	×	√	88.47	87.34	73.32	82.08	84.43	81.59	78.33	88.21
×	×	×	×	88.07	87.03	73.02	81.75	84.17	81.82	77.89	88.83

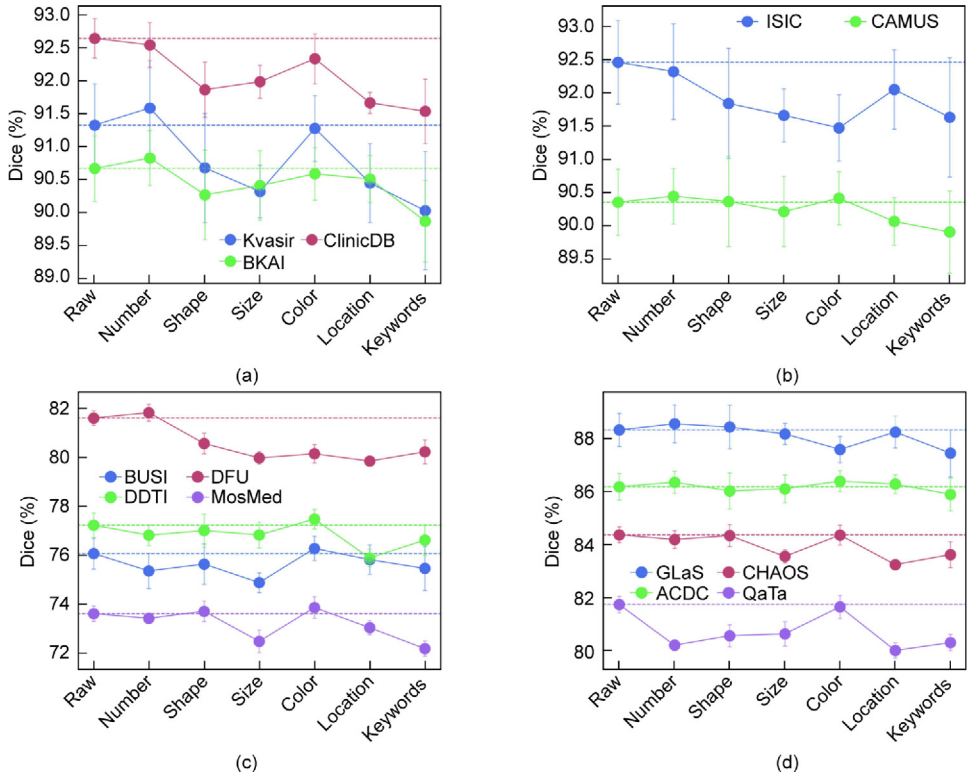

Fig. 6. Concept-wise ablation studies on different datasets: (a) ablation results on the Kvasir, ClinicDB, and BKAI datasets; (b) ablation results on the ISIC and CAMUS datasets; (c) ablation results on the BUSI, DFU, DDTI, and MosMed datasets; and (d) ablation results on the GLaS, CHAOS, ACDC, and QaTa datasets.

decreases segmentation accuracy across all 13 datasets. This finding highlights the critical role of class keywords in the VLSM, as they encapsulate important prior knowledge about segmentation targets (e.g., polyps are typically red and round). These ablation experiments yield two key insights: ① appropriate textual concepts vary significantly across different datasets, and ② additional comprehensive textual information does not always improve performance. Moreover, removing textual concepts with minimal contribution to ROI recognition can actually enhance the segmentation performance of the model.

Medical image segmentation tasks present a wide variety of usable text prompts. Poudel et al. [36] extracted text prompts with 14 attributes for medical image segmentation and identified nine distinct prompts. Besides the six attributes used in our study, several of the remaining eight attributes were specific to the dataset such as View, Pathology, Cardiac Cycle, and Tumor Type in their study, while others could not be directly obtained from ROIs such as General Class Info, Gender, Age, and Image Quality. In line with the principles of generality and effectiveness, our proposed CoTexNet model adopts class keywords with five conceptual

attributes—number, shape, size, location, and color—as coarse real-text prompts. The results of comparison and ablation experiments reveal that our designed scheme is effective for 13 medical image segmentation tasks.

5.5. Ablation study on SAN-KD

Unlike traditional KD methods that typically transfer knowledge between different models (usually from a larger model to a smaller model) and focus on adjusting the temperature coefficient to modify the soft label distribution [64,65,70], the proposed SAN-KD emphasizes distilling knowledge across different modalities within the same model. It specifically focuses on two key aspects: crossmodal knowledge alignment and reducing the impact of distillation errors during testing.

For the first aspect, we performed crossmodal data normalization using the Sigmoid function, followed by an adaptive adjustment of data distribution through the inverse *Z*-score operation. Fig. 7 presents a comparison of segmentation results obtained using SAN-KD, SAN-KD without noise addition operation (NN-SAN-KD), SAN-KD without data alignment operations (ND-SAN-KD), and numerical fitting (NF) based on the mean squared error (MSE) loss. The segmentation performance obtained using these four methods demonstrates that SAN-KD consistently exhibits the best performance.

For the second aspect, we introduced a learnable noise parameter α in SAN-KD to mitigate the gap between training and testing caused by distillation errors. Fig. 8 shows the mean and standard deviation of the Dice scores of the segmentation results for different initial values of α. When the initial value of α is zero, no noise from T_c is initially added to T_e, whereas when the initial value of $\alpha$is one, the model is initially trained using T_c exclusively. Fig. 8 shows that as the initial value of α increases, the Dice value initially increases and then decreases across all datasets, while the standard deviation gradually decreases. This trend suggests that adding noise enhances model stability on the test set and segmentation performance with small α ($\alpha < 0.4$). However, with a large α, although the model exhibits high stability on the test set, the reduced influence of T_e hinders its ability to effectively support the downstream segmentation task, lowering its performance. In summary, we set the initial value of α to 0.2 and obtain its final learned values within the range of [0.15, 0.28] across all 13 datasets. This allows the model to adaptively adjust α within a relatively narrow range for each dataset, preventing the learning of excessively large values that could degrade model performance.

6. Discussion

6.1. Interpretable structured prompts

In our framework, clinical interpretability is preserved through the integration of two complementary prompts: the real-text prompts extracted by the LMM provide coarse, human-readable descriptions corresponding to clinically interpretable features such as lesion location, shape, and echogenicity. In contrast, the pseudo-text prompts generated via the text latent space transformation module capture fine-grained visual cues—such as texture and boundary clarity—that are difficult to describe explicitly in natural language.

By combining these two prompts, the model exhibits a structured representation that maintains semantic transparency while preserving high visual precision. This dual-prompt design bridges textual semantics and visual perception, facilitating multimodal reasoning aligned with clinical understanding. Visualization analyses further confirm that the pseudo-text prompts consistently focus on pathologically relevant regions, supporting their clinical interpretability. These results suggest that the proposed structured prompt framework provides an interpretable and clinically meaningful link between language, vision, and disease-related image features.

6.2. Innovative insights

Although the proposed CoTexNet framework is designed based on established components such as the CLIPSeg backbone, prompt learning, and KD, its importance lies in the synergistic integration of these modules to address a previously underexplored challenge in medical vision, namely VLSMs, bridging explicit text semantics and implicit visual context without manual prompt engineering.

The proposed dual-prompt formulation introduces a new mechanism to couple coarse real-text prompts that capture conceptual semantics with fine-grained pseudo-text prompts that encode visual-specific nuances using a text latent space transformation module. This structured design allows the model to achieve semantic–visual alignment at multiple granularity levels. Furthermore, the SAN-KD strategy provides a principled and stable route to transfer multimodal linguistic knowledge into the CLS token for direct text-guided inference, effectively mitigating the error accumulation commonly observed in VQA-based pseudo-text generation methods.

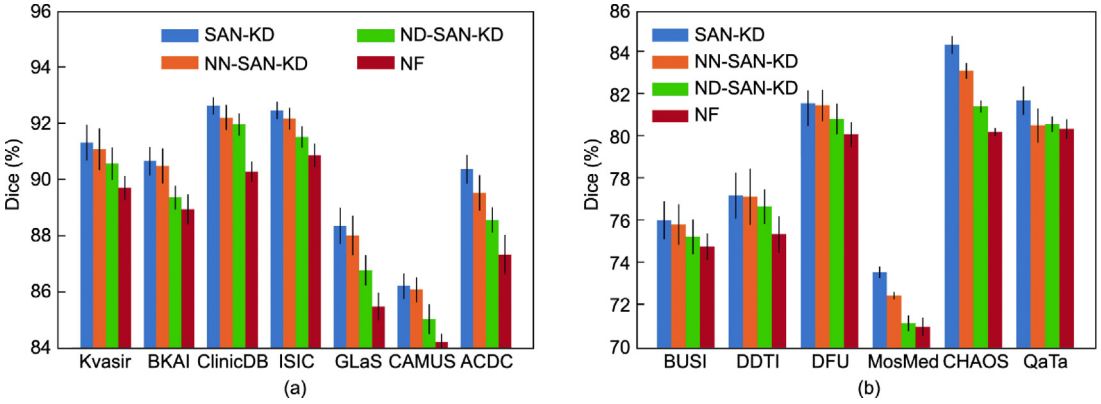

Fig. 7. Comparison of segmentation performance using SAN-KD, SAN-KD without noise addition operation (NN-SAN-KD), SAN-KD without data alignment operations (ND-SAN-KD), and numerical fitting (NF) based on the MSE loss: (a) results on the Kvasir, BKAI, ClinicDB, ISIC, GLaS, CAMUS, and ACDC datasets; and (b) results on the BUSI, DDTI, DFU, MosMed, CHAOS, and QaTa datasets.

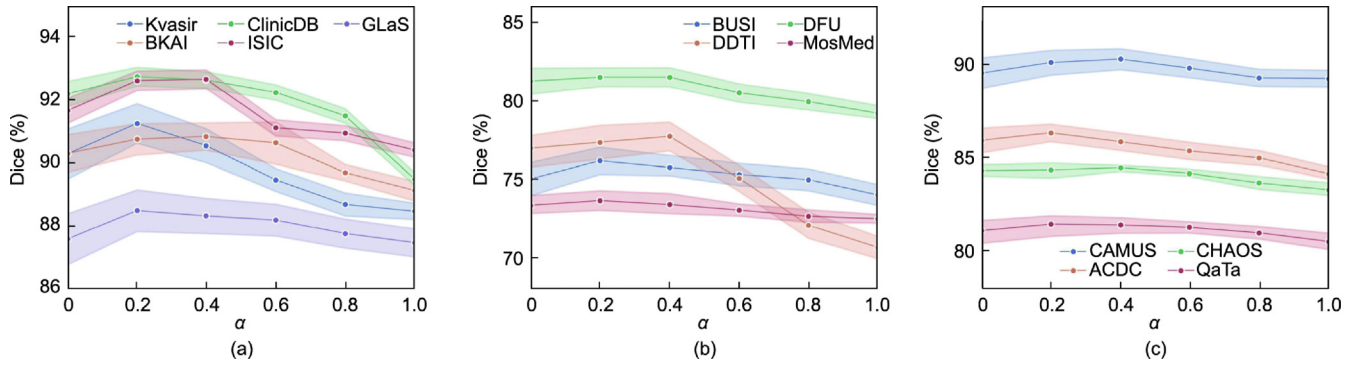

Fig. 8. Ablation studies on the hyperparameter α: (a) ablation results on the Kvasir, CLinicDB, GLaS, BKAI, and ISIC datasets; (b) ablation results on the BUSI, DFU, DDTI, and MosMed datasets; and (c) ablation results on the CAMUS, CHAOS, ACDC, and QaTa datasets. The shaded area along the line represents the magnitude of the standard deviation.

Beyond its empirical improvements, CoTexNet contributes a generalizable perspective on visual–linguistic corepresentation learning, demonstrating that conceptual text semantics can be derived from visual data itself rather than externally supplied annotations. This direction—learning implicit language grounding directly from images—opens new opportunities for autonomous prompt generation and offers valuable insights for future studies on multimodal interpretability and self-supervised text–vision alignment in medical imaging.

6.3. Limitations and future directions

In this study, we designed a dual-prompt learning framework to enhance medical image segmentation by integrating real- and pseudo-text information within a unified vision–language model. While the proposed framework demonstrates strong segmentation performance across multiple datasets, certain limitations exist. A key consideration is clinical generalizability.Although the model performs well under current experimental settings, its robustness across medical images obtained from different imaging devices, modalities, or lesion types requires further validation.Variations in image quality, resolution, and acquisition protocols might affect real- and pseudo-text prompt generation, potentially influencing segmentation outcomes.

In our future study, we shall conduct segmentation experiments on additional medical image datasets to further evaluate and improve the generalizability of the proposed framework. Our ultimate goal is to develop a broadly applicable medical VLSM and integrate it into a large-scale intelligent medical diagnosis system, facilitating highly accurate lesion delineation and clinical decision-making.

7. Conclusions

In this study, we developed a method for learning conceptual text prompts from ROIs for medical image segmentation. The learned conceptual text prompts comprised coarse real-text and fine-grained pseudo-text prompts, which complemented each other and enhanced the proposed model's adaptability to various medical image segmentation tasks. Furthermore, ablation experiments demonstrated that the proposed SAN-KD method effectively improved the segmentation performance of the model by enhancing its robustness to distillation errors during the testing phase. Comparative experiments and visualization results indicated that the proposed CoTexNet model exhibited superior segmentation performance across multiple medical image segmentation tasks with minimal human intervention.

CRediT authorship contribution statement

Zhu He: Writing – original draft, Visualization, Software, Methodology, Investigation. **Haoran Zhang:** Validation, Investigation. **Wentao Zhang:** Resources, Data curation. **Shen Zhao:** Resources, Investigation. **Qiqi Liu:** Validation, Formal analysis. **Xiaohu Wu:** Validation, Supervision, Project administration, Conceptualization. **Qicheng Lao:** Writing – review & editing, Supervision, Project administration, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eng.2026.04.006.

The source code associated with this work can be accessed at: https://github.com/Rango-bit/CoTexNet.

References

[1] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18–24; online. Cambridge: PMLR; 2021. p. 8748–63.

[2] Saito K, Sohn K, Zhang X, Li CL, Lee CY, Saenko K, et al. Pic2Word: mapping pictures to words for zero-shot composed image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17–24; Vancouver, BC, Canada. New York City: IEEE; 2023. p. 19305–14.

[3] Jia C, Yang Y, Xia Y, Chen YT, Parekh Z, Pham H, et al. Scaling up visual and vision–language representation learning with noisy text supervision. In: Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18–24; online. Cambridge: PMLR; 2021. p. 4904–16.

[4] Yang Y, Panagopoulou A, Zhou S, Jin D, Callison-Burch C, Yatskar M. Language in a bottle: language model guided concept bottlenecks for interpretable image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17–24; Vancouver, BC, Canada. New York City: IEEE; 2023. p. 19187–97.

[5] Li LH, Zhang P, Zhang H, Yang J, Li C, Zhong Y, et al. Grounded language–image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. New York City: IEEE; 2022. p. 10965–75.

[6] Lüddecke T, Ecker A. Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. New York City: IEEE; 2022. p. 7086–96.

[7] Chng SY, Tern PJW, Kan MRX, Cheng LTE. Automated labelling of radiology reports using natural language processing: comparison of traditional and newer methods. Health Care Sci 2023;2(2):120–8.

[8] You C, Dai W, Liu F, Min Y, Dvornek NC, Li X, et al. Mine your own anatomy: revisiting medical image segmentation with extremely limited labels. IEEE Trans Pattern Anal Mach Intell 2024;46(12):11136–51.
[9] Zhong Y, Xu M, Liang K, Chen K, Wu M. Ariadne's thread: using text prompts to improve segmentation of infected areas from chest X-ray images. In: Proceedings of the 26th International Conference on Medical Image Computing and Computer Assisted Intervention; 2023 Oct 8–12; Vancouver, BC, Canada. Cham: Springer; 2023. p. 724–33.
[10] Zhang Y, Jiang H, Miura Y, Manning CD, Langlotz CP. Contrastive learning of medical visual representations from paired images and text. In: Proceedings of the 7th Machine Learning for Healthcare Conference; 2022 Aug 11–12; Boston, MA, USA. Cambridge: PMLR; 2022. p. 2–25.
[11] Wei Q, Gu Z, Tan W, Kong H, Fu H, Jiang Q, et al. Development and validation of an automatic ultrawide-field fundus imaging enhancement system for facilitating clinical diagnosis: a cross-sectional multicenter study. Engineering 2024;41:179–88.
[12] Li X, Li L, Jiang Y, Wang H, Qiao X, Feng T, et al. Vision–language models in medical image analysis: from simple fusion to general large models. Inf Fusion 2025;118:102995.
[13] Li A, Zeng X, Zeng P, Ding S, Wang P, Wang C, et al. Textmatch: using text prompts to improve semi-supervised medical image segmentation. In: Proceedings of the 27th International Conference on Medical Image Computing and Computer Assisted Intervention; 2024 Oct 6–10; Marrakech, Morocco. Cham: Springer; 2024. p. 699–709.
[14] Zhang Z, Yao L, Wang B, Jha D, Durak G, Keles E, et al. DiffBoost: enhancing medical image segmentation via text-guided diffusion model. IEEE Trans Med Imaging 2025;44(9):3670–82.
[15] Lee GE, Kim SH, Cho J, Choi ST, Choi SI. Text-guided cross-position attention for segmentation: case of medical image. In: Proceedings of the 26th International Conference on Medical Image Computing and Computer Assisted Intervention; 2023 Oct 8–12; Vancouver, BC, Canada. Cham: Springer; 2023. p. 537–46.
[16] You C, Zhou Y, Zhao R, Staib L, Duncan JS. SimCVD: simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. IEEE Trans Med Imaging 2022;41(9):2228–37.
[17] Lin W, Zhao Z, Zhang X, Wu C, Zhang Y, Wang Y, et al. PMC-CLIP: contrastive language–image pre-training using biomedical documents. In: Proceedings of the 26th International Conference on Medical Image Computing and Computer Assisted Intervention; 2023 Oct 8–12; Vancouver, BC, Canada. Cham: Springer; 2023. p. 525–36.
[18] Liu Y, Li X, Luo Y, Du J, Zhang Y, Lv T, et al. Toward a large language model-driven medical knowledge retrieval and QA system: framework design and evaluation. Engineering 2025;50:270–82.
[19] Du C, Zhang Z, Liu B, Cao Z, Jiang N, Zhang Z. Explainable machine learning model for pre-frailty risk assessment in community-dwelling older adults. Health Care Sci 2024;3(6):426–37.
[20] Bazi Y, Rahhal MMA, Bashmal L, Zuair M. Vision–language model for visual question answering in medical imagery. Bioengineering 2023;10(3):380.
[21] Wang S, Zhao Z, Ouyang X, Wang Q, Shen D. ChatCAD: interactive computer-aided diagnosis on medical image using large language models. 2023. arXiv:2302.07257.
[22] You C, Dai W, Min Y, Staib L, Duncan JS. Implicit anatomical rendering for medical image segmentation with stochastic experts. In: Proceedings of the 26th International Conference on Medical Image Computing and Computer Assisted Intervention; 2023 Oct 8–12; Vancouver, BC, Canada. Cham: Springer; 2023. p. 561–71.
[23] Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. Nat Commun 2024;15(1):654.
[24] Zhang B, Zhang P, Dong X, Zang Y, Wang J. Long-CLIP: unlocking the long-text capability of CLIP. In: Proceedings of the 18th European Conference on Computer Vision; 2024 Oct 23–28; Milan, Italy. Cham: Springer; 2024. p. 310–25.
[25] Liu H, Li C, Li Y, Li B, Zhang Y, Shen S, et al. LLaVA-NeXT: improved reasoning, OCR, and world knowledge [Internet]. San Francisco: GetHub; 2024 Jan 30 [cited 2026 Mar 10]. Available from: https://llava-vl.github.io/blog/2024-01-30-llava-next/.
[26] Wu F, Shen T, Bäck T, Chen J, Huang G, Jin Y, et al. Knowledge-empowered, collaborative, and co-evolving ai models: the post-LLM roadmap. Engineering 2025;44:87–100.
[27] You C, Dai W, Min Y, Liu F, Clifton D, Zhou SK, et al. Rethinking semi-supervised medical image segmentation: a variance-reduction perspective. In: Proceedings of the 37th Conference on Neural Information Processing Systems; 2023 Dec 10–16; New Orleans, LA, USA. Red Hook: Curran Associates Inc.; 2023,36. p. 9984–10021.
[28] Zhou Z, Lei Y, Zhang B, Liu L, Liu Y. ZegCLIP: towards adapting CLIP for zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023 Jun 17–24; Vancouver, BC, Canada. New York City: IEEE; 2023. p. 11175–85.
[29] He Z, Liu Y, Yang G, Bao X, Chai Y, Lao Q. Learning task-level pseudo-text prompt for improved medical image segmentation. In: Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine; 2024 Dec 3–6; Lisbon, Portugal. New York City: IEEE; 2024. p. 3262–7.
[30] Liu Y, Pei J, He Z, Yang G, Jiang Z, Lao Q. Medical language mixture of experts for improving medical image segmentation. In: Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine; 2024 Dec 3–6; Lisbon, Portugal. New York City: IEEE; 2024. p. 2210–6.
[31] Wang F, Zhou Y, Wang S, Vardhanabhuti V, Yu L. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In: Proceedings of the 36th Conference on Neural Information Processing Systems; 2022 Dec 6–14; New Orleans, LA, USA. Red Hook: Curran Associates Inc.; 2022. p. 33536–49.
[32] Boecking B, Usuyama N, Bannur S, Castro DC, Schwaighofer A, Hyland S, et al. Making the most of text semantics to improve biomedical vision–language processing. In: Proceedings of the 17th European Conference on Computer Vision; 2022 Oct 23–27; Tel Aviv, Israel. Cham: Springer; 2022. p. 1–21.
[33] Han X, Chen Q, Xie Z, Li X, Yang H. Multiscale progressive text prompt network for medical image segmentation. Comput Graph 2023;116:262–74.
[34] Zhan C, Zhang Y, Lin Y, Wang G, Wang H. UniDCP: unifying multiple medical vision-language tasks via dynamic cross-modal learnable prompts. IEEE Trans Multimed 2024;26:9736–48.
[35] Tomar NK, Jha D, Bagci U, Ali S. TGANet: text-guided attention for improved polyp segmentation. In: Proceedings of the 25th International Conference on Medical Image Computing and Computer Assisted Intervention; 2022 Oct 2–6; Singapore. Cham: Springer; 2022. p. 151–60.
[36] Poudel K, Dhakal M, Bhandari P, Adhikari R, Thapaliya S, Khanal B. Exploring transfer learning in medical image seg-mentation using vision–language models. In: Proceedings of the Medical Imaging with Deep Learning Conference; 2024 Jul 1–4; Singapore. Paris: MIDL Organizing Committee; 2024. p. 1–24.
[37] Qin Z, Yi HH, Lao Q, Li K. Medical image understanding with pretrained vision language models: a comprehensive study. In: Proceedings of the 11th International Conference on Learning Representations; 2023 May 1–5; Kigali, Rwanda. Kigali: ICLR Organizing Committee; 2022. p. 1–20.
[38] Dadoun H, Delingette H, Rousseau AL, de Kerviler E, Ayache N. Joint representation learning from French radiological reports and ultrasound images. In: Proceedings of the IEEE 20th International Symposium on Biomedical Imaging; 2023 Apr 18–21; Cartagena, Colombia. New York City: IEEE; 2023. p. 1–5.
[39] Li Z, Li Y, Li Q, Wang P, Guo D, Lu L, et al. LViT: language meets vision transformer in medical image segmentation. IEEE Trans Med Imaging 2024;43(1):96–107.
[40] Xu Y, Kong M, Xie W, Duan R, Fang Z, Lin Y, et al. Deep sequential feature learning in clinical image classification of infectious keratitis. Engineering 2021;7(7):1002–10.
[41] He Z, Lin M, Luo X, Xu Z. Structure-preserved self-attention for fusion image information in multiple color spaces. IEEE Trans Neural Netw Learn Syst 2025;36(7):13021–35.
[42] Liu H, Li C, Li Y, Lee YJ. Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 10–16; Seattle, WA, USA. New York City: IEEE; 2024; p. 26296–306.
[43] Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. In: Proceedings of the 37th Conference on Neural Information Processing Systems; 2023 Dec 10–16; New Orleans, LA, USA. Red Hook: Curran Associates Inc.; 2023, 36. p. 1–25.
[44] Dumoulin V, Perez E, Schucher N, Strub F, Vries H, Courville A, et al. Feature-wise transformations. Distill 2018;3(7):e11.
[45] You C, Dai W, Min Y, Staib L, Sekhon J, Duncan JS. Action++: improving semi-supervised medical image segmentation with adaptive anatomical contrast. In: Proceedings of the 26th International Conference on Medical Image Computing and Computer Assisted Intervention; 2023 Oct 8–12; Vancouver, BC, Canada. Cham: Springer; 2023. p. 194–205.
[46] Pogorelov K, Randel KR, Griwodz C, Eskeland SL, de Lange T, Johansen D, et al. KVASIR: a multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of the 8th ACM International Conference on Multimedia Systems; 2017 Jun 20–23; Taipei, China. New York City: ACM; 2017. p. 164–9.
[47] Bernal J, Sánchez FJ, Fernández-Esparrach G, Gil D, Rodríguez C, Vilariño F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Comput Med Imaging Graph 2015;43:99–111.
[48] Ngoc Lan P, An NS, Hang DV, Long DV, Trung TQ, Thuy NT, et al. NeoUNet: towards accurate colon polyp segmentation and neoplasm detection. In: Proceedings of the 16th International Symposium on Advanced Visual Computing; 2021 Dec 13–15; online. Cham: Springer; 2021. p. 15–28.
[49] Codella NC, Gutman D, Celebi ME, Helba B, Marchetti MA, Dusza SW, et al. Skin lesion analysis toward melanoma detection: a challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). In: Proceedings of the 15th IEEE International Symposium on Biomedical Imaging; 2018 Apr 4–7; Washington, DC, USA. New York City: IEEE; 2018. p. 168–72.
[50] Wang C, Anisuzzaman DM, Williamson V, Dhar MK, Rostami B, Niezgoda J, et al. Fully automatic wound segmentation with deep convolutional neural networks. Sci Rep 2020;10(1):21897.
[51] Pedraza L, Vargas C, Narváez F, Durán O, Muñoz E, Romero E. An open access thyroid ultrasound image database. In: Proceedings of the 10th International Symposium on Medical Information Processing and Analysis; 2015 Apr 22–24; Cartagena, Colombia. Bellingham: SPIE; 2015. p. 188–93.
[52] Xian M, Zhang Y, Cheng HD, Xu F, Huang K, Zhang B, et al. BUSIS: a benchmark for breast ultrasound image segmentation. 2018. arXiv:1801.03182.
[53] Sirinukunwattana K, Pluim JPW, Chen H, Qi X, Heng PA, Guo YB, et al. Gland segmentation in colon histology images: the GlaSchallenge contest. Med Image Anal 2017;35:489–502.

[54] Morozov SP, Andreychenko AE, Pavlov NA, Vladzymyrskyy AV, Ledikhova NV, Gombolevskiy VA, et al. MosMedData: chest CT scans with COVID-19 related findings dataset. 2020. arXiv:2005.06465.
[55] Hofmanninger J, Prayer F, Pan J, Röhrich S, Prosch H, Langs G. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. Eur Radiol Exp 2020;4(1):50.
[56] Kavur AE, Gezer NS, Barış M, Aslan S, Conze PH, Groza V, et al. CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation. Med Image Anal 2021;69:101950.
[57] Degerli A, Kiranyaz S, Chowdhury ME, Gabbouj M. OSegNet: operational segmentation network for COVID-19 detection using chest X-ray images. In: Proceedings of the IEEE International Conference on Image Processing; 2022 Oct 16–19; Bordeaux, France. New York City: IEEE; 2022. p. 2306–10.
[58] Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng PA, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans Med Imaging 2018;37(11):2514–25.
[59] Leclerc S, Smistad E, Pedrosa J, Østvik A, Cervenansky F, Espinosa F, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Trans Med Imaging 2019;38(9):2198–210.
[60] Ibtehaz N, Kihara D. ACC-Unet: a completely convolutional Unet model for the 2020s. In: Proceedings of the 26th International Conference on Medical Image Computing and Computer Assisted Intervention; 2023 Oct 8–12; Vancouver, BC, Canada. Cham: Springer; 2023. p. 692–702.
[61] Wang C, Wang L, Wang N, Wei X, Feng T, Wu M, et al. CFATransUnet: channel-wise cross fusion attention and transformer for 2D medical image segmentation. Comput Biol Med 2024;168:107803.
[62] Li Z, Li D, Xu C, Wang W, Hong Q, Li Q, et al. TFCNs: a CNN-transformer hybrid network for medical image segmentation. In: Proceedings of the International Conference on Artificial Neural Networks; 2022 Sep 6–9; Bristol, UK. Cham: Springer; 2022. p. 781–92.
[63] Li Y, Jing B, Li Z, Wang J, Zhang Y. Plug-and-play segment anything model improves nnUNet performance. Med Phys 2025;52(2):899–912.
[64] Wang Z, Lu Y, Li Q, Tao X, Guo Y, Gong M, et al. CRIS: CLIP-driven referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. New York City: IEEE; 2022. p. 11686–95.
[65] Wang P, Yang A, Men R, Lin J, Bai S, Li Z, et al. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: Proceedings of the International Conference on Machine Learning; 2022 Jul 17–23; Baltimore, MD, USA. Cambridge: PMLR; 2022. p. 23318–40.
[66] Nath V, Li W, Yang D, Myronenko A, Zheng M, Lu Y, et al. VILA-M3: enhancing vision–language models with medical expert knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2025 Jun 15–21; Los Angeles, CA, USA. New York City: IEEE; 2025. p. 14788–98.
[67] Li C, Wong C, Zhang S, Usuyama N, Liu H, Yang J, et al. LLaVA-Med: training a large language-and-vision assistant for biomedicine in one day. In: Proceedings of the 37th Conference on Neural Information Processing Systems; 2023 Dec 10–16; New Orleans, LA, USA. Red Hook: Curran Associates Inc.; 2023. p. 28541–64.
[68] Wu Z, Chen X, Pan Z, Liu X, Liu W, Dai D, et al. Deepseek-vl2: mixture-of-experts vision–language models for advanced multimodal understanding. 2024. arXiv:2412.10302.
[69] Wang P, Bai S, Tan S, Wang S, Fan Z, Bai J, et al. Qwen2-vl: enhancing vision–language model's perception of the world at any resolution. 2024. arXiv:2409.12191.
[70] Gou J, Chen Y, Yu B, Liu J, Du L, Wan S, et al. Reciprocal teacher–student learning via forward and feedback knowledge distillation. IEEE Trans Multimed 2024;26:7901–16.