EviVLM: When Evidential Learning Meets Vision Language Model for Medical Image Segmentation

Qingtao Pan, Zhengrong Li, Guang Yang, Qing Yang and Bing Ji

***Abstract*— The disparity between image and text representations, often referred to as the modality gap, remains a significant obstacle for Vision Language Models (VLMs) in medical image segmentation. This gap complicates multi-modal fusion, thereby restricting segmentation performance. To address this challenge, we propose Evidence-driven Vision Language Model (EviVLM)—a novel paradigm that integrates Evidential Learning (EL) into VLMs to systematically measure and mitigate the modality gap for enhanced multi-modal fusion. To drive this paradigm, an Evidence Affinity Map Generator (EAMG) is proposed to collect complementary cross-modal evidences by learning a global cross-modal affinity map, thus refining modality-specific evidence embedding. An Evidence Differential Similarity Learning (EDSL) is further proposed to collect consistent cross-modal evidences by performing Bias-Variance Decomposition on differential matrix derived from bidirectional similarity matrices between image and text evidence embeddings. Finally, the subjective logic is used for mapping the collected evidences to opinions, and the Dempster-Shafer's theory based combination rule is introduced for opinion aggregation, thereby quantifying the modality gap and facilitating effective multi-modal integration. Experimental results on three public medical image segmentation datasets validate that the proposed EviVLM can achieve state-of-the-art performance. Code is available at: https://github.com/QingtaoPan/EviVLM.**

***Index Terms*— Evidential learning, medical image segmentation, vision-language model, modality gap**

I. Introduction

VISION Language Model (VLM) aims to align image-text pairs to facilitate cross-modal learning [1]. This technique is also being extended to medical images, aiming to learn general medical vision representations from medical text descriptions and then transfer the learned representations to downstream tasks. In the previous studies, medical image based VLMs have achieved huge success for medical image segmentation [2,3,4].

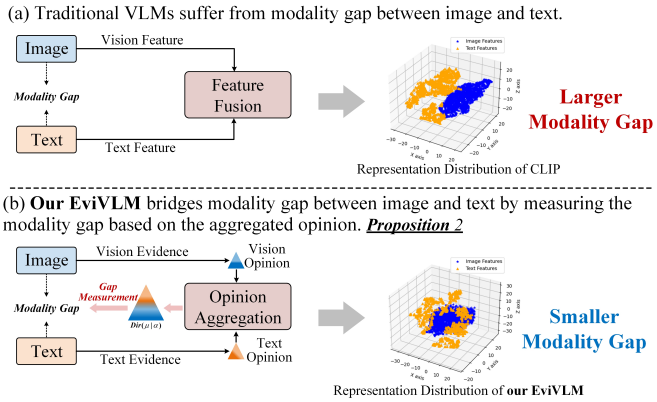

Fig. 1. Our EviVLM bridges the large modality gap between visual and textual representations by injecting EL into VLM for the estimation of such modality gap.

Although the progresses in VLM driven medical image segmentation, the modality gap between vision and text is still significant (Fig. 1(a)) [5]. This gap may compromise information fusion performance and thus limits the textual semantic guidance in medical image segmentation [8]. This is because the visual and textual representations extracted by VLMs tend to cluster into two distinct groups and have gap vectors between them. Under the presence of clear modality gap in image-text pairs, it is hard to measure reliable image-text similarity for modality fusion, leading to inferior performance on VLM-based medical image segmentation. Specifically, for each image-text pair, the goal is to learn complementary and consistent cross-modal representations for reliable modality fusion. That is, the image-modality itself should provide ample vision representation, while the text-modality is encouraged to afford the language representation as consistent as possible. However, such collected representations suffer from semantic bias between matched image and text. The root cause lies in the fundamental differences between image and text data, leading to unreliable image-text representations.

To bridge the modality gap between medical images and

This work was partly supported by the National Natural Science Foundation of China (Grant No.62173212), Taishan Scholars Program of Shandong Province (Grant No.tsqn202306017), Shandong Province"Double-Hundred Talent Plan"on 100 Foreign Experts and 100 Foreign Expert Teams (Grant No.WSR2023049), and the ERC IMI (101005122). (Corresponding author: Bing Ji, e-mail: b.ji@sdu.edu.cn)

Qingtao Pan, Zhengrong Li, and Bing Ji are with the School of Control Science and Engineering, Shandong University, Jinan, China.

Qing Yang is with Department of Breast and Thyroid Surgery, Shandong Provincial Hospital Affiliated to Shandong First Medical University, Jinan, China

Guang Yang is with the Department of Bioengineering, Faculty of Engineering, Imperial College London, SW7 2AZ London, U.K., and also with the Imperial-X and the School of Biomedical Engineering and Imaging Sciences, King's College London, WC2R 2LS London, U.K. (e-mail: g.yang@imperial.ac.uk).

texts, most previous medical VLM methods [2,3,6], generally utilize two separate encoders to extract cross-modal features. The cross attention mechanism is used to fuse image and text features early in the transformer encoder (i.e., cross-modal shallow interaction) [59]. Subsequently, such features are embedded into a latent common space through some elaborately designed objective functions to achieve modality invariance. Nonetheless, these methods built upon shallow interaction between both modalities are insufficient to eliminate the modality gap [7], since the narrow cone of embedding space restricts the expression of complex semantic relationships between images and texts.

In this article, we aim to address the modality gap via the evidential learning theory [9,10]. It can quantify cross-modal evidence aggregation uncertainty trustfully by collecting cross-modal subjective evidences and attracts increasing attention and been successfully used in a variety of tasks [11,12,13]. As our main motivation, we argue that addressing the problem of modality gap lies in estimating the reliability of the matching between modalities through the prediction confidences, rather than the widely used cosine similarity because it only provides a static similarity value that cannot indicate whether this value is reliable. The evidential learning theory, as an advanced uncertainty theory, supports an efficient and stable framework for uncertainty quantification of prediction confidences [14]. It can be regarded as an alternative measure of the modality distance [15], enabling the estimation of the modality gap.

Despite these potential advantages, applying existing evidential learning methods to VLM is infeasible. On the one hand, evidence embeddings of image and text extracted by deep evidential networks is difficult to form a complementary relationship. This is because the image evidence embedding is local representations based on pixel level, while the text evidence embedding is global representations based on contextual relationships. This poses a challenge when expecting to learn complementary cross-modal evidence embeddings. On the other hand, deep evidence learning may lead to overconfident predictions when cross-modal evidences are inconsistent [10]. This is because when two modalities provide conflicting evidences, deep evidence learning may mistakenly amplify the confidence of one modality, leading to model overconfidence. This poses another challenge when expecting to learn consistent cross-modal evidence embeddings. Hence, it is imperative to develop more appropriate evidential learning networks to alleviate these two challenges.

Inspired by the above observations, in this paper, we propose an Evidential Vision Language Model (EviVLM) to bridge modality gap by aggregating cross-modal opinions for modality gap estimation (Fig. 1(b)). To ensure the complementarity of the collected image-text evidences, we incorporate a global cross-modal affinity map to evidence embeddings through the proposed Evidence Affinity Map Generator (EAMG), and a Evidence Differential Similarity Learning (EDSL) is proposed to boost the consistency between image and text evidence embeddings by measuring variation inconsistency of bidirectional similarity matrices. Finally, we map the collected image-text evidences into corresponding image-text opinions through the subjective logic, and aggregate such image-text opinions based on empster-Shafer's evidence theory, achieving more effective modality fusion for medical image segmentation. Our contributions are summarized as follows:

(1) To the best of our knowledge, we are the first to introduce evidential learning into visual-language models, which bridges the modality gap by aggregating cross-modal opinions for modality gap estimation, thus improving modality fusion and subsequent medical image segmentation performance.

(2) To collect complementary cross-modal evidences, EAMG is proposed to learn a global cross-modal affinity map, refining both modality-specific evidence embeddings.

(3) To ensure the consistency for cross-modal evidences, EDSL is proposed to measure variation inconsistency of bidirectional similarity matrices by performing Bias-Variance Decomposition based on the differential matrix, boosting vigorous alignment between cross-modal evidence embeddings.

(4) We conduct extensive and in-depth experiments on three public medical image segmentation datasets. The encouraging results compared with state-of-the-arts demonstrate the effectiveness of our method.

II. Related works

In this section, we provide an overview of the related works in three aspects: (1) Vision Language Model, (2) Evidential Learning in Image Segmentation, and (3) Text-guided Medical Image Segmentation.

A. Vision Language Model

With the widespread success of large models in the field of language processing, vision language models (VLMs) adopt BERT-like architecture [16] contrastive learning paradigm [1,17] to learn vision-language representations. Although most VLMs focus on pre-trained vision language models, another line of works focus on integrating VLMs into vision recognition scenario. Such studies are capable of semantic segmentation [2], image-text retrieval [18], image captioning [19] and so on. Qin *et al.* [49] suggests that the language model can be used for detection task in the medical image domain through designed text prompts. In the field of medical image analysis, Liu *et al.* [51] designed a simple and effective medical image-text pretraining method to better perform contrastive learning. Specifically, by simplifying the calculation of similarity between medical image-report pairs into the calculation of similarity between reports, image report tuples are divided into positive, negative, and additional neutral groups. Wang *et al.* [52] proposed a multimodal collaborative prompt learning (MCPL) pipeline for adjusting frozen VLM to align medical image-text representations, thereby achieving downstream medical tasks. However, current VLMs suffer from modality gap, leading to suboptimal performance for downstream tasks. Motivated by the above findings, we try to introduce text embedding to pixel-level semantic understanding, i.e., medical image segmentation while bridging such modality gap.

B. Evidential Learning in Image Segmentation

Evidential Learning (EL), a method for reliable model inference and uncertainty estimation, has been gained increasing attention [20]. The core of EL is the notion of

evidence, which reflects the level of belief and uncertainty for different hypotheses within the model's predictions through Dempster-Shafer evidence theory [21] and Subjective Logic theory [22]. Benefiting from its advantages, many researchers have applied EL in medical image segmentation tasks [23,24]. For the multi-modality segmentation task, Huang *et al.* [23] proposed a multi-modality evidence fusion method for medical image segmentation, which computes a belief function at each voxel for each modality and combines evidences using Dempster's rule. Diao *et al.* [50] generated the final segmentation results by fusing uncertain evidence from PET and CT. Li *et al.* [24] focused on evidence-based cross-entropy loss function for trusted medical image segmentation and proposed an evidential soft Dice loss under the Dirichlet prior distribution. In this paper, we pioneeringly inject EL into VLM for medical image segmentation. Further, we propose a EAMG for complementary cross-modal evidences collection, and a EDSL for consistent cross-modal evidence embedding learning, improving the quality of cross-modal evidences for more effective image-text fusion.

C. Text-guided Medical Image Segmentation

Recently, with the rise of large language models, utilizing the information of medical text descriptions is helpful for medical image analysis. Tomar *et al.* [46] extracted text semantics using 'byte-pair' encoding and employed it as a soft channel attention to reinforce the representative features and suppress the less-important features. In this way, the semantic information encoded in the text was used to guide the image segmentation. Li *et al.* [36] proposed LViT, a dual-U structure consisting of a U-shaped CNN branch alongside a U-shaped Visual Transformer (ViT) branch. This model extracted the fused image and text features within the U-shaped ViT branch and then further integrated them via a Pixel-Level Attention Module (PLAM). The text-enhanced image features were then fed into the CNN decoder. However, LViT initially extracted the text features via a simple vectorization operation, which cannot capture rich semantics. Inspired by Segment Anything Model (SAM) [25], Hu *et al.* [48] proposed extracting image features from the pretrained Segment Anything Model (SAM) and simultaneously extracting text features from the pretrained BERT model [26]. These features were then combined using multi-layer cross-attention modules and inputted into the segmentation decoder. These methods have outperformed imageonly segmentation methods, highlighting the potential of language-guided medical image segmentation. Nonetheless, these methods primarily focused on the feature feature extraction and fusion strategies, neglecting the inherent large pattern gap between the two distinct image-text modalities. This gap may compromise information fusion performance and thus limits the textual semantic guidance in medical image segmentation.

III. Methods

A. Preliminary

1) Evidential Learning: The evidential learning is a method for reliable model inference and uncertainty estimation. It reflects the belief and uncertainty for different hypotheses within the model's predictions through Dempster-Shafer evidence theory and Subjective Logic theory.

For the binary medical image segmentation task, evidential learning models the uncertainty of the class prediction $\mathbf{p} = [p_1, p_2]$ on each pixel position x. Then, it constructs a Dirichlet distribution of probabilities $\mathbf{p}$:

$$\mathbf{p} \sim \mathrm{Dir}(\boldsymbol{\alpha}), \quad \boldsymbol{\alpha} = [\alpha_1, \alpha_2] \tag{1}$$

where $\boldsymbol{\alpha}$ is the concentration parameter of the Dirichlet distribution. The relationship between it and the evidence is as follows:

$$\alpha_k = e_k + 1, \quad e_k \geq 0, \quad k \in \{1, 2\} \tag{2}$$

where e_k is the evidence vector provided by the model for category k. The total amount of evidence is:

$$S = \alpha_1 + \alpha_2 = e_1 + e_2 + 2 \tag{3}$$

The expected value of the predicted probability obtained from the Dirichlet distribution is:

$$\mathbb{E}[p_k] = \frac{\alpha_k}{\sum_j \alpha_j} = \frac{e_k + 1}{S}, \quad k \in \{1, 2\} \tag{4}$$

This form is more interpretable than softmax and allows us to further derive uncertainty indicators from the parameters. The uncertainty is defined as:

$$u = \frac{2}{S} = \frac{2}{e_1 + e_2 + 2} \tag{5}$$

It can be seen that the more evidence (i.e., the larger the sum of $e_1 + e_2$), the smaller the uncertainty u. When the model lacks evidence ($e_1, e_2 \to 0$), the uncertainty tends to reach its maximum value ($u \to 1$).

In this work, Dirichlet distribution is used to model uncertainty. That is because Dirichlet distribution can be well integrated with Subjective Logic, interpreting the output of neural networks as evidence for each class and constructing the parameters of the Dirichlet distribution. It not only predicts the expected probability of the class, but also estimates its uncertainty. The more evidence, the more concentrated the distribution, and the more confident the model is; The less evidence, the flatter the distribution, and the greater the uncertainty. For other distributions, Beta distribution is a special case of the Dirichlet distribution. The difference is that the Beta distribution can only model binary classification or segmentation tasks, while the Dirichlet distribution can be extended to multi classification or segmentation tasks. Gaussian distribution is not suitable for directly modeling the distribution and uncertainty of class probabilities in classification problems. It is a continuous distribution defined on real numbers. However, we need to model a probability vector in classification or segmentation problems. This probability vector is a special geometric space (i.e., probability simplex), and Gaussian distribution cannot generate suitable probability vectors in this space, making it difficult to directly model uncertainty.

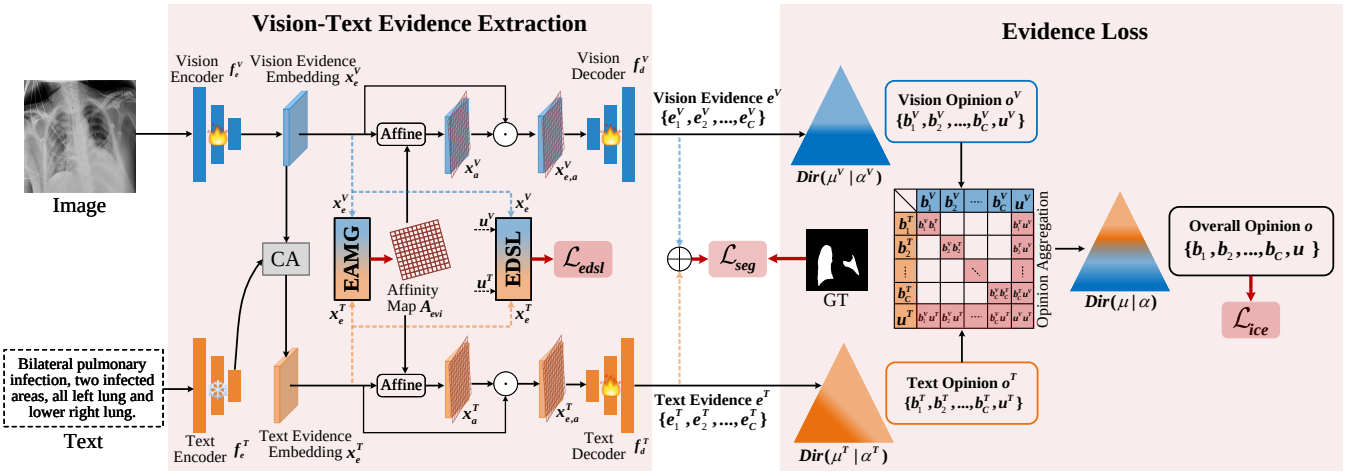

Fig. 2. Illustration of EviVLM. Our EviVLM brides the modality gap by measuring the uncertainty of the aggregated opinion. First, we extract the vision-text evidences. Specifically, the image is input to the vision encoder to obtain vision evidence embedding $\boldsymbol{x_e^V}$, and then sent to the vision decoder to obtain vision evidence $\boldsymbol{e^V}$. The text and image are fed into the cross attention module (CA) to obtain text evidence embedding $\boldsymbol{x_e^T}$, and then input into the text decoder to obtain text evidence $\boldsymbol{e^T}$. It also contains two new designs, i.e., EAMG and EDSL. The EAMG learns a affinity map to refine both modality-specific evidence embeddings. The EDSL measures the inconsistency between similarity matrices to balanced modality representations. Second, we construct vision-text Dirichlet distribution based on vision-text evidence. Then, subjective logic is adopted to optimize the parameters of Dirichlet distribution, which provides a mass belief and uncertainty for the segmentation results, thus obtaining vision-text opinion. The Dempster-Shafer's theory based combination rule used for vision-text opinion aggregation, thereby calculating the aggregated evidence loss for modality gap measurement.

2) Modality Gap: In the EviVLM setting, suppose that we are given an image-text dataset $D = \{V, T\}$ with the images $V = \{v_i\}_{i=1}^N$ and texts $T = \{t_i\}_{i=1}^N$, where N denotes the total number of image-text pairs. We use x_e^V and x_e^T to denote the encoded vision and text evidence embeddings respectively, while e^V and e^T denote the decoded vision and text evidences respectively. The goal of EviVLM is to bridge the modality gap between x_e^V and x_e^T for more efficient cross-modal consistency learning, thus boosting modality fusion between e^V and e^T. The attached opinion aggregation uncertainty u between vision and text opinions is used for modality gap measurement.

B. Overview

The proposed EviVLM (Fig. 2) maps the learned evidences to opinions using the subjective logic and aggregates opinions via the Dempster-Shafer's theory, thus measuring the modality gap for efficient modality fusion. To learn complementary cross-modal evidences, EAMG is proposed. It generates a global cross-modal affinity map to refine evidence embeddings of both images and texts. To driven consistent cross-modal evidences, EDSL is designed. It measures inconsistency of cross-modal evidence embeddings by conducting Bias-Variance Decomposition based on the differential matrix calculated by bidirectional similarity matrices. Finally, we can obtain the decoded vision-text opinions for modality fusion while simultaneously measure the decision reliability via the aggregated overall opinion. The detailed description of EviVLM is as follows.

C. Evidence Embedding Extraction

Given an image-text pair $\{V, T\}$, the encoding path of U-Net [27] and BioClinicalBERT [28] are used as vision encoder f_e^V and text encoder f_e^T to extract vision evidence embedding x_e^V and text evidence embedding x_e^T respectively, where $x_e^V \in \mathbb{R}^d$, $x_e^T \in \mathbb{R}^d$. For x_e^V, it is extracted directly by f_e^V, i.e., $x_e^V = f_e^V(V)$. For x_e^T, it is obtained by a Cross-Attention Module (CA) [29]. Formally, for the vision evidence embedding x_e^V, we let x_e^V attend to text token embeddings x^T encoded by f_e^T, and then calculate its corresponded cross-modal text evidence embedding x_e^T,

$$x_e^T = \alpha^{V2T}(\tilde{V}x^T), \tag{6}$$

$$\alpha^{V2T} = \text{softmax}(\frac{(\tilde{Q}x_e^V)^T(\tilde{K}x^T)}{\sqrt{d}}), \tag{7}$$

where $\tilde{Q} \in \mathbb{R}^d$, $\tilde{K} \in \mathbb{R}^d$, and $\tilde{V} \in \mathbb{R}^d$ are learnable matrices.

D. Evidence Affinity Map Generator (EAMG)

Motivation: In cross-modal tasks, complementarity means that images and text can provide features that are difficult for each other to express. For example, in medical imaging analysis tasks, certain lesions may be difficult to distinguish visually, but a textual description can provide additional information such as location. Traditional deep evidence networks [23,53] independently extract different evidence embeddings using different deep evidence networks. However, these methods ignore cross-modal interaction during the evidence extraction stage, so that the extracted evidence embeddings remain independent. Cross-modal affinity, is able to learn global correspondences as well as locally-varying between modalities [54]. Therefore, we argue that cross-modal affinity has the potential to enhance the mutual guidance between text evidence and visual evidence, thereby enhancing complementarity.

Our EAMG (Fig. 3) integrates two learned vision-aware and text-aware evidence affinities to generate a global cross-modal

pixel-level affinity map for evidence embeddings refinement, thus strengthening the mutual complementarity between images and text. The non-local self-attention block (NonLocal) is exploited to capture the semantic correlations of spatial positions based on the similarities between the feature vectors of any two positions, to learn modality-specific affinities.

$$A^V_{evi} = \text{NonLocal}(x^V_e), \quad A^T_{evi} = \text{NonLocal}(x^T_e). \tag{8}$$

Specifically, for $x^V_e \in \mathbb{R}^{H\times W\times D}$, it is encoded into a triplet of Q, K, V through three 1×1 convolutional layers, and then such triplet is flattened to be of size $HW \times D$. The dot product between Q and K is used to produce $A^V_{evi} \in \mathbb{R}^{HW\times HW}$. Each row of A^V_{evi} represents the similarity values of a spatial position and the rest ones. The text evidence affinity A^T_{evi} is generated with the same non-local operation. To learn mutual evidence affinities, a self-attention (SA) module is learned to synthesize the two modality-specific affinity maps through two convolutional layers and a softmax layer. Then two spatial attention maps $[w^V, w^T] = \text{SA}(\text{concat}(A^V_{evi}, A^T_{evi}))$ are produced from SA module to aggregate A^V_{evi} and A^T_{evi} into a global cross-modal affinity map $A_{evi} \in \mathbb{R}^{HW\times HW}$,

$$A_{evi} = w^V * A^V_{evi} + w^T * A^T_{evi}, \tag{9}$$

where $w^V, w^T \in \mathbb{R}^{HW\times HW}$ are the learned two spatial attention maps. To refine both modality-specific evidence embeddings $[x^V_e, x^T_e]$, we affine the global cross-modal affinity map A_{evi} on x^V_e and x^T_e, respectively, thus obtaining two refined cross-modal evidence embeddings $x^V_{e,a}$ and $x^T_{e,a}$,

$$x^V_{e,a} = \text{affine}(A_{evi}, x^V_e) \odot x^V_e, \tag{10}$$

$$x^T_{e,a} = \text{affine}(A_{evi}, x^T_e) \odot x^T_e, \tag{11}$$

where $\text{affine}(.,.)$ denotes the evidence affine operator, and $\odot$ is the Hadamard product.

Definition 1 (Evidence Affine Operator). The affined vision evidence embedding $x^V_{e,a}$, calculated from vision evidence embedding x^V_e and affinity map A_{evi} using evidence affine operator, is derived as follows,

$$\text{affine}(A_{evi}, x^V_e) = \sum_h^H \sum_w^W A_{evi}(h,w) \cdot x^V_e(h,w), \tag{12}$$

where h and w are the width and height of the affinity map.

Summarized advantages: EAMG refines the cross-modal evidence embedding through a learned global pixel-level affinity map, enhancing the complementarity of cross-modal evidence learning.

E. Evidence Differential Similarity Learning (EDSL)

Motivation: VLMs widely use two similarity matrices (image-to-text, text-to-image) to pull similar image text pairs closer. However, if there is a bias in the learning of two similarity matrices during the training process of the model, for example, the model tends to overestimate the similarity from image to text and underestimate the similarity from text to image, then directly using the similarity matrix may lead to a certain modality dominating the matching decision. Calculating the difference matrix between these two similarity matrices can explicitly weaken the deviation information between modalities, enabling the loss function to optimize the balance of cross-modal similarity learning for better cross-modal evidence consistency.

For the image-to-text similarity matrix S_{V2T} and the text-to-image similarity matrix S_{T2V}, assuming the differential loss between these two similarity matrices is:

$$\mathcal{L}_{\text{d}} = \frac{1}{2}\|D\|^2 = \frac{1}{2}\|S_{V2T} - S_{T2V}\|^2 = \frac{1}{2}\sum_{i,j}(S^{(i,j)}_{V2T} - S^{(i,j)}_{T2V})^2 \tag{13}$$

where i and j represent the ith image and jth text respectively. Then we calculate gradients for two similarity matrices separately:

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{diff}}}{\partial S^{(i,j)}_{V2T}} &= \frac{\partial}{\partial S^{(i,j)}_{V2T}}\left(\frac{1}{2}(S^{(i,j)}_{V2T} - S^{(i,j)}_{T2V})^2\right) \\ &= S^{(i,j)}_{V2T} - S^{(i,j)}_{T2V} = D^{(i,j)}\end{aligned} \tag{14}$$

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{diff}}}{\partial S^{(i,j)}_{T2V}} &= \frac{\partial}{\partial S^{(i,j)}_{T2V}}\left(\frac{1}{2}(S^{(i,j)}_{V2T} - S^{(i,j)}_{T2V})^2\right) \\ &= -(S^{(i,j)}_{V2T} - S^{(i,j)}_{T2V}) = -D^{(i,j)}\end{aligned} \tag{15}$$

If the model overestimates the image-to-text similarity S_{V2T}, that is:

$$S^{(i,j)}_{V2T} > S^{(i,j)}_{T2V} \Rightarrow D^{(i,j)} > 0 \tag{16}$$

Then in backpropagation, $S^{(i,j)}_{V2T}$ will receive a positive gradient $D^{(i,j)}$ and will be reduced. $S^{(i,j)}_{T2V}$ will receive a negative gradient $-D^{(i,j)}$ and will be increased. This indicates that optimizing the differential loss will lower the high similarity matrix and improve the low similarity matrix.

To this end, our EDSL (Fig. 4) performs the Bias-Variance Decomposition based on the differential matrix to learn the variation inconsistency between two bidirectional cross-modal similarity matrices, thus boosting strengthful alignment between image and text evidence embeddings. To further ensure the reliability of the similarity matrix, the vision opinion uncertainty u^V and text opinion uncertainty u^T (will be explained in the next section) are applied to vision evidence embedding x^V_e and text evidence embedding x^T_e , respectively. Therefore, for x^V_e and x^T_e, their evidence similarity matrices s_{ij} and $\tilde{s}_{ji}$ can be computed by

$$s_{ij} = cos(u^V \odot x^V_{e,i}, u^T \odot x^T_{e,j}), \tag{17}$$

$$\tilde{s}_{ji} = cos(u^T \odot x^T_{e,j}, u^V \odot x^V_{e,i}), \tag{18}$$

where $cos(\cdot)$ measures the cosine similarity. $S_{V2T} = [s_{ij}] \in \mathbb{R}^{B\times B}$ and $S_{T2V} = [\tilde{s}_{ji}] \in \mathbb{R}^{B\times B}$ denote two bidirectional cross-modal similarity matrices respectively, where B is the batch size.

To measure the variation inconsistency between two similarity matrices $[S_{V2T}, S_{T2V}]$, a straightforward idea is to subtract S_{V2T} from S_{T2V} to obtain their differential matrix S_{diff}, and then perform the Bias-Variance Decomposition [30] based on S_{diff},

$$S_{diff} = f_{BVD}(S_{V2T}, S_{T2V}) \tag{19}$$

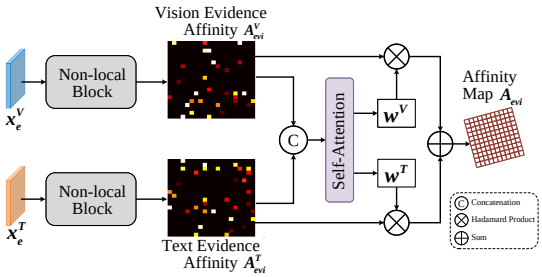

Fig. 3. EAMG learns cross-modal evidence affinities via two non-local blocks, and integrates such two modality-specific affinity maps with a self-attention to produce a global cross-modal affinity map.

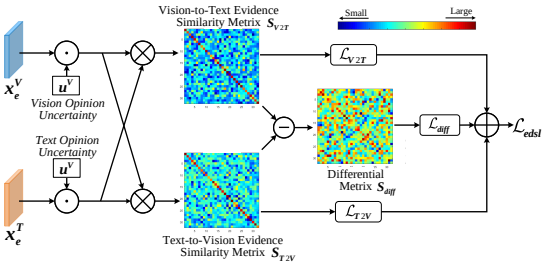

Fig. 4. EDSL performs the Bias-Variance Decomposition based on the differential matrix to learn the variation inconsistency between two cross-modal similarity matrices, thus boosting the alignment between image and text evidence embeddings.

Definition 2 (Bias-Variance Decomposition). Let S_{V2T} and S_{T2V} be the bidirectional similarity matrices, respectively. The f_{BVD} is derived as follows,

$$\begin{aligned}\mathcal{L}_{diff} &= \|D\|_{BVD}^2 \approx B^2\mathbb{E}[(s-\tilde{s})^2] = B^2(\mathbb{E}[s^2] - 2\mathbb{E}[s\tilde{s}] + \mathbb{E}[\tilde{s}^2]) \\ &= B^2(\mathbb{E}^2[s] + Var(s) - 2\mathbb{E}[s]\mathbb{E}[\tilde{s}] - 2Cov(s,\tilde{s}) + \mathbb{E}^2[\tilde{s}] + Var(\tilde{s})) \\ &= B^2(\mathbb{E}^2[s] + \mathbb{E}^2[\tilde{s}] - 2\mathbb{E}[s]\mathbb{E}[\tilde{s}]) + B^2(Var(s) + Var(\tilde{s}) - 2Cov(s,\tilde{s})) \\ &= \underbrace{B^2(\mathbb{E}[s] - \mathbb{E}[\tilde{s}])^2}_{Bias} + \underbrace{B^2 Var(s-\tilde{s})}_{Variance},\end{aligned} \tag{20}$$

where $\mathbb{E}[\cdot]$, $Var(\cdot)$, and $Cov(\cdot)$ represent the expectation, variance, and covariance operations, respectively. s and $\tilde{s}$ denote image-to-text similarity matrix and text-to-image similarity matrix, respectively. B is a scaling constant. The computed Bias and Variance measure the variation inconsistency between two cross-modal similarity matrices. Therefore, we regard $\|D\|_{BVD}^2$ as the the inconsistency differential loss $\mathcal{L}_{diff}$ of similarity matrices. Specifically, for formula 23, Our goal is to calculate and optimize the difference between two similarity matrices. A straightforward idea is to subtract one similarity matrix from another. Therefore, we obtain $\mathcal{L}_{diff} = B^2\mathbb{E}[(s-\tilde{s})^2]$ and unfold it as $\mathcal{L}_{diff} = B^2(\mathbb{E}[s^2] - 2\mathbb{E}[s\tilde{s}] + \mathbb{E}[\tilde{s}^2])$. According to $\mathbb{E}[x^2] = \mathbb{E}[x]^2 + Var(x)$ and $\mathbb{E}[xy] = \mathbb{E}[x]\mathbb{E}[y] + Cov(x,y)$, we obtain $\mathcal{L}_{diff} = B^2(\mathbb{E}[s]^2 + Var(s) - 2(\mathbb{E}[s]\mathbb{E}[\tilde{s}] + Cov(s,\tilde{s})) + \mathbb{E}[\tilde{s}]^2 + Var(\tilde{s}))$, which can be merged as $L_{diff} = B^2(\mathbb{E}[s] - \mathbb{E}[\tilde{s}])^2 + B^2 Var(s-\tilde{s})$, where $B^2(\mathbb{E}[s] - \mathbb{E}[\tilde{s}])^2$ represents the expectation difference between similarity matrices. $B^2 Var(s-\tilde{s})$ denotes the fluctuation degree between similarity matrices.

Additionally, we also calculate InfoNCE [31] losses $\mathcal{L}_{V2T}^i$, $\mathcal{L}_{T2V}^j$ of two cross-modal similarity matrices above to maximally preserve the mutual information between the true image-text pairs,

$$\begin{aligned}\mathcal{L}_{V2T}^i &= -\log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B}\exp(s_{ij}/\tau)}, \\ \mathcal{L}_{T2V}^j &= -\log\frac{\exp(\tilde{s}_{jj}/\tau)}{\sum_{i=1}^{B}\exp(\tilde{s}_{ji}/\tau)},\end{aligned} \tag{21}$$

where τ is the temperature hyperparameter. The overall objective of EDSL is the sum of both InfoNCE losses and inconsistency loss,

$$\mathcal{L}_{edsl} = \lambda_1\mathcal{L}_{diff} + \lambda_2\frac{1}{2B}(\sum_{i=1}^{B}\mathcal{L}_{V2T}^i + \sum_{j=1}^{B}\mathcal{L}_{T2V}^j), \tag{22}$$

where λ_1 and λ_2 are hyperparameters to balance such two losses.

Summarized advantages: EDSL performs the Bias-Variance Decomposition based on the differential matrix for consistency learning of cross-modal similarity matrices, enhancing the consistency of cross-modal evidences.

F. Modality Gap Measurement

To measure the modality gap based on the aggregated opinion, we represent the decoded vision evidence $e^V = f_d^V(x_{e,a}^V)$ and text evidence $e^T = f_d^T(x_{e,a}^T)$ as vision opinion o^V and text opinion o^T to quantify the distributional uncertainty in $Dir(\mu^V|\alpha^V)$ and $Dir(\mu^T|\alpha^T)$.

According to subjective logic, a principled method of probabilistic reasoning under uncertainty [32], the vision Dirichlet distribution of class probabilities $Dir(\mu^V|\alpha^V)$ is determined by the evidence $e^V = \{e_1^V, e_2^V, ..., e_C^V\}$, where C is the number of classes. Following the work [33], we derive the parameter of the distribution through $\alpha^V = e^V + 1$. Then the vision Dirichlet distribution is mapped to the vision opinion $o^V = \{b_1^V, b_2^V, ..., b_C^V, u^V\}$, satisfying

$$u^V + \sum_{c=1}^{C} b_c^V = 1, \tag{23}$$

where $b_c^V = \frac{\alpha_c^V - 1}{S^V}$ is the belief mass for class c, $S^V = \sum_{c=1}^{C}\alpha_c^V$ is the Dirichlet, and $u^V = \frac{C}{S^V}$ measures the uncertainty of the Dirichlet distribution.

The final predicted pixel-level probabilities $\tilde{p}^V \in \mathbb{R}^C$ of all classes are the expectation of Dirichlet distribution, i.e., $\mathbb{E}[Dir(\mu^V|\alpha^V)[\mu^V]$, where μ^V is the original predicted probabilities. We assume that pixel-level samples can not provide any evidence for decision making, i.e., $e^V = 0$. According to the definitions of a^V, S^V, and u^V, the uncertainty u^V has a negative correlation with the sum of evidences. Therefore, such pixel-level samples would yield high uncertainty. The reasoning process of text opinion $o^T = \{b_1^T, b_2^T, ..., b_C^T, u^T\}$ is the same as that of vision.

To obtain the aggregated opinion based on both vision and text opinion, following the Dempster-Shafer's evidence theory [34], we use belief fusion operators to aggregate vision opinion o^V and text opinion o^T. Specifically, for $o^V = \{b_1^V, b_2^V, ..., b_C^V, u^V\}$ and $o^T = \{b_1^T, b_2^T, ..., b_C^T, u^T\}$,

the aggregated opinion $o = \{b_1, b_2, ..., b_C, u\} = o^V \oplus o^T$ is derived by

$$b_c = \frac{1}{M}(b_c^V b_c^T + b_c^V u^T + b_c^T u^V), u = \frac{1}{M} u^V u^T, \tag{24}$$

where $M = 1 - \sum_{i \neq j} b_i^V b_j^T$ is the normalization factor.

To compute the losses for vision, text and aggregated opinion, following the work [33], we use the integrated cross-entropy loss. The loss for vision opinion is given by

$$\begin{aligned}\mathcal{L}_{ice}^V &= \mathbb{E}_{\mu^V \sim Dir(\mu^V|\alpha^V)}[\mathcal{L}_{CE}(\mu^V, y)] \\ &= \int \left[\sum_{c=1}^{C} -y_c \log(\mu_c^V)\right] \frac{1}{B(\alpha^V)} \prod_{c=1}^{C} (\mu_c^V)^{\alpha_c^V - 1} \mathrm{d}\mu^V \\ &= \sum_{c=1}^{C} y_c^V (\psi(S^V) - \psi(\alpha_c^V))\end{aligned} \tag{25}$$

where y is the one-hot label and ψ is the digamma function. The overall loss of vision, text and aggregated opinion is given by

$$\mathcal{L}_{evi} = \mathcal{L}^V + \mathcal{L}^T + \mathcal{L}^{aggregated} \tag{26}$$

where $\mathcal{L}^T$ and $\mathcal{L}^{aggregated}$ are implemented as same as $\mathcal{L}^V$.

G. Theoretical Analysis of EviVLM

Proposition 1. Aggregating image-text opinions can improve the segmentation performance, that is, aggregating an additional opinion into the original opinion can potentially improve the segmentation belief. Under the condition $b_g^T > b_m^V$, where g is the index of ground-truth class and b_m^V is the largest belief mass in o^V, aggregating another opinion o^T makes the new opinion o satisfy $b_g \geq b_g^V$.

Proof.

$$\begin{aligned} b_g &= \frac{b_g^V b_g^T + b_g^V u^T + b_g^T u^V}{1 - \sum_{i \neq j} b_i^V b_j^T} \\ &= \frac{b_g^V b_g^T + b_g^V u^T + b_g^T u^V}{\sum_{c=1}^{C} b_c^V b_c^T + u^T + u^V - u^V u^T} \\ &\geq \frac{b_g^V (b_g^T + u^T + u^V)}{b_m^V (1 - u^T) + u^T + u^V - u^V u^T} \\ &\geq \frac{b_g^V (b_g^T + u^T + u^V)}{b_m^V + u^T + u^V} \geq b_g^V. \end{aligned} \tag{27}$$

Proposition 2. For the conflictive opinion aggregation with modality gap, aggregating image-text opinions can reduce the modality gap: the aggregated uncertainty mass u will be reduced after integrating text opinion o^T into the vision opinion o^V, i.e., $u \leq u^V$.

Proof.

$$\begin{aligned} u &= \frac{u^V u^T}{1 - \sum_{i \neq j} b_i^V b_j^T} \\ &= \frac{u^V u^T}{\sum_{c=1}^{C} b_c^V b_c^T + u^T + u^V - u^V u^T} \\ &\leq \frac{u^V u^T}{u^T + u^V - u^V u^T} \leq u^V. \end{aligned} \tag{28}$$

Most existing VLM-based segmentation methods rely on fusion strategies (e.g., cross-attention, feature concatenation, contrastive learning, or similarity-based matching), which do not explicitly model the uncertainty and conflict between modalities. In contrast, our method models each modality's prediction as a subjective opinion and utilizes Dempster's rule of combination to perform uncertainty-aware opinion fusion. According to formula 31, if b^V and b^T are in conflict, the denominator becomes smaller, and u increases. If the beliefs are in agreement, the denominator is close to 1, and the fused uncertainty u becomes $u \leq \min(u^V, u^T)$. Thus, agreement between modalities directly leads to uncertainty reduction, which corresponds to modality gap mitigation. This ability is unique to evidential fusion. Existing attention-based or similarity-based methods lack this ability. This is because existing methods output deterministic prediction probabilities and cannot measure the level of trust in prediction probabilities.

H. Theoretical Analysis against alternative methods

To demonstrate the unique advantages of evidence learning over alternative methods in addressing modal gap, we conducted a theoretical comparison and analysis between EviVLM and classical cross attention and contrastive learning methods.

1) The advantage of Evidential Learning: For vision opinion $o^V = \{b_1^V, b_2^V, ..., b_C^V, u^V\}$ and text opinion $o^T = \{b_1^T, b_2^T, ..., b_C^T, u^T\}$, the modality consistency between them is:

$$s = \sum_{c=1}^{C} b_c^V b_c^T \quad 0 \leq s \tag{29}$$

The modality conflict κ and s monotonic reverse, defined as $\kappa = \sum_{i \neq j} b_i^V b_j^T$. Below, we define and demonstrate two advantages of evidence learning in addressing modal gap.

1. Monotonicity: $\frac{\partial u}{\partial \kappa} \geq 0$, i.e., The increase of conflict κ, the fusion uncertainty u increases.

Proof.

Following formula (28), the denominator $\sum_{c=1}^{C} b_c^V b_c^T + u^T + u^V - u^V u^T = s + u^T + u^V - u^V u^T = D > 0$. We can rewrite Following formula (28) as:

$$u = \frac{u^V u^T}{s + u^T + u^V - u^V u^T} \tag{30}$$

Taking the derivative of s, we can obtain,

$$\begin{aligned} \frac{\partial u}{\partial s} &= -\frac{u^V u^T}{(s + u^T + u^V - u^V u^T)^2} \\ &= -\frac{u^V u^T}{D^2} < 0 \end{aligned} \tag{31}$$

Due to $\kappa = \text{const} - s$,

$$\frac{\partial u}{\partial \kappa} = \frac{\partial u}{\partial s} \frac{\partial s}{\partial \kappa} = -\frac{u^V u^T}{D^2} \cdot (-1) = \frac{u^V u^T}{D^2} \geq 0 \tag{32}$$

The monotonicity is proved. Therefore, by optimizing the model to reduce the uncertainty u, the conflict κ can be minimized, thereby reducing the modality gap.

2. Boundedness: $u \leq \min(u^V, u^T)$ **if** b^V **is close to** b^T

Proof.
Divide both sides of the formula (30) by u^V,

$$\frac{u}{u^V} = \frac{u^T}{s + u^T + u^V - u^V u^T} = \frac{u^T}{D} \tag{33}$$

Assuming $\frac{u^T}{D} <= 1$ holds true,

$$u^T \leq D = s + u^V + u^T - u^V u^T \tag{34}$$

equivalent to,

$$0 \leq s + u^V - u^V u^T \tag{35}$$

Due to $s >= 0$ and $u^V - u^V u^T = u^V(1 - u^T) >= 0$, $\frac{u^T}{D} <= 1$ is established. Therefore, $\frac{u}{u^V} <= 1$, hence, $u <= u^T$. Similarly, $u <= u^V$. Finally, $u \leq \min(u^V, u^T)$ is proved. This indicates that when two modalities are close, the uncertainty is limited to not be too large, thereby constraining the modality gap.

2) The limitation of Cross Attention: **Degradation to mean pooling.** The cross attention conducts weighted aggregation through query-key similarity.

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \tag{36}$$

Assuming Q comes from the text, and K, V comes from an image. When the modality gap is significant, the similarity between the Q and K approaches zero, i.e., $QK^\top \approx 0$. Thus,

$$\text{softmax}(QK^\top) \approx \frac{1}{n_v}1 \tag{37}$$

where n_v is the number of patch tokens of an image. As a result, the attention output degenerates as,

$$\text{Attn}(Q, K, V) \approx \frac{1}{n_v}\sum_{i=1}^{n_v} V_i \tag{38}$$

It is simply the average pooling of all image regions. This indicates that cross attention cannot reliably model modal relationships when the modality gap is too large.

3) The limitation of Contrastive Learning: **Modal ambiguity.** Comparative learning alleviates modality gap by maximizing the similarity of matching image-text pairs and minimizing the similarity of non-matching image-text pairs. The objective function is:

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(sim(h_i^I, h_i^T)/\tau)}{\sum_{j=1}^{B}\exp(sim(h_i^I, h_j^T)/\tau)} \tag{39}$$

where h_i^I, h_i^T represent the embeddings of the image and text. However, this loss only learns point-based similarity, i.e., one-to-one matching, which may lead to modal ambiguity. For example, when the similarities between the image patch and multiple text tokens are close,

$$sim(h^I, h_1^T) \approx sim(h^I, h_2^T), \quad T_1 \neq T_2, \tag{40}$$

where h_1^T and h_2^T are two similar texts but in the same batch, this loss rigidly executes single image-text pair (i.e., h_1^T and h_2^T) optimal alignment and cannot express ambiguity in multiple pairs. This means that the originally similar h^I and h_2^T will be pushed away because they are considered as the negative pair in contrastive learning.

Summarized advantages: EviVLM maps cross-modal evidences to opinions through subjective logic and aggregates such opinions using the Dempster-Shafer's theory, measuring the modality gap based on the uncertainty of the aggregated opinion.

I. Training Paradigm

The proposed EviVLM can be trained in an end-to-end manner. $\mathcal{L}_{edsl}$ is used to boost strong matching between image and text evidence embeddings. $\mathcal{L}_{evi}$ is utilized to measure modality gap for cross-modal jointly reasoning. Additionally, we introduce the segmentation loss $\mathcal{L}_{seg}$ between the fused vision-text evidence and true mask,

$$\mathcal{L}_{seg} = -\frac{1}{HW}\sum_{m=1}^{HW}\mathcal{L}_{CE}(\sigma(e_m^V + e_m^T), y_m), \tag{41}$$

where m represents the mth pixel in mask. W and H are the width and height of an image. Therefore, the overall loss of our EviVLM is denoted as:

$$\mathcal{L}_{overall} = \omega_1\mathcal{L}_{edsl} + \omega_2\mathcal{L}_{evi} + \mathcal{L}_{seg}, \tag{42}$$

where ω_1 and ω_2 are hyperparameters to balance three objective functions.

IV. Experiments

A. Datasets

Three public datasets are adopted for evaluation, including

1) **QaTa-COV19** dataset [35] consists of 9258 COVID-19 chest X-ray radiographs, which is compiled by researchers from Qatar University and Tampere University. Additionally, text annotations are given by [36]. The text annotations focus on whether both lungs are infected, the number of lesion regions, and the approximate location of the infected areas. For example, "Unilateral pulmonary infection, one infected area, lower right lung." denotes unilateral lung infection, and there is one infection area located in the lower right lung. Following [36], we divide this dataset into 7:1:2 for training, validation, and testing, respectively.
2) **MosMedData+** dataset [37,38] contains 2729 CT scan slices of lung infections, which merges the COVID-19 lesion masks and their corresponding frames of these 3 public datasets. All different types of lesions are mapped to white color for consistency across datasets. The text annotations are also given by [36]. The text annotation is similar to QaTa-COV19. For example, "Bilateral pulmonary infection, two infected areas, middle left lung and upper right lung." Following [36], we divide this dataset into 7:1:2 for training, validation, and testing, respectively.
3) **Duke-Breast-Cancer-MRI** dataset [39] contains 922 MRI scans from 922 breast cancer patients. The text annotations are provided and verified by two professionals. These annotations contain information on the location, shape, size, and the number of lesions. For example, "location left, shape round, size small, numbe one." We divide this dataset into 7:1:2 for training, validation, and testing respectively.

TABLE I

THE QUANTITATIVE COMPARISON BETWEEN OUR METHOD AND OTHER COMPARISON METHODS ON THE QATA-COV19, MOSMEDDATA+, AND DUKE-BREAST-CANCER-MRI TESTING DEMONSTRATES THE SUPERIORITY OF OUR METHOD. METHODS MARKED IN GREY REPRESENT TRADITIONAL SEGMENTATION APPROACHES THAT DO NOT USE TEXT INFORMATION, WHILE THOSE MARKED IN ORANGE REPRESENT ADVANCED VLM METHODS (BEING CAPABLE OF ADDRESSING MODALITY GAP) UTILIZING TEXT INFORMATION. THE BEST VALUES ARE IN BOLD.

Method	Text	Param (M)	Flops (G)	QaTa-COV19		MosMedData+		Duke-Breast-Cancer-MRI	
				Dice (%)	mIoU (%)	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)
U-Net [27]	×	14.8	25.2	79.02	69.46	64.60	50.73	83.23	74.41
UNet++ [40]	×	74.5	94.6	79.62	70.25	71.75	58.39	82.52	74.96
AttUNet [41]	×	34.9	101.9	79.31	70.04	66.34	52.82	79.01	70.43
nnUNet [42]	×	19.1	412.7	80.42	70.81	72.59	60.36	81.11	72.02
TransUNet [43]	×	105.0	56.7	78.63	69.13	71.24	58.44	84.16	77.18
SwinUnet [44]	×	82.3	67.3	78.07	68.34	63.29	50.19	82.79	71.75
UCTransNet [45]	×	65.6	63.2	79.15	69.60	65.90	52.69	80.21	70.80
STPNet [62]	×	31.2	55.9	80.63	71.42	76.18	63.41	82.30	73.44
ConVIRT [18]	✓	35.2	44.6	79.72	70.58	72.06	59.73	82.89	73.60
TGANet [46]	✓	19.8	41.9	79.87	70.75	71.81	59.28	83.69	72.93
CLIP [1]	✓	87.0	105.3	79.81	70.66	71.97	59.64	84.50	73.19
GLoRIA [2]	✓	45.6	60.8	79.94	70.68	72.42	60.18	82.84	71.52
ViLT [47]	✓	87.4	55.9	79.63	70.12	72.36	60.15	84.18	73.40
LAVT [58]	✓	118.6	83.8	79.28	69.89	73.29	60.41	85.84	76.31
MGCA [3]	✓	135.6	18.1	80.92	71.04	73.22	60.53	85.08	**76.86**
LViT [36]	✓	29.7	54.1	83.66	75.11	74.57	61.33	83.42	73.55
CARZero [55]	✓	142.16	30.75	81.41	73.66	72.96	61.08	82.15	73.84
MLIP [56]	✓	150.78	45.63	83.10	74.28	73.84	60.77	83.16	73.44
SAT [57]	✓	50.63	64.11	82.32	73.69	74.24	61.47	84.61	73.37
Ours	✓	63.6	42.1	**85.79**	**77.34**	**77.64**	**65.81**	**86.96**	76.29

B. Implementation Details

Our method is implemented within PyTorch on the Ubuntu 20.04.4 LTS with 24GB V100 GPU. BioClinicalBERT [28] is used as the text encoder, and the encoding part of U-Net [27] is used as the vision encoder. The cross attention module does not use the multi-head mechanism, but directly performs feature matrix multiplication. The decoding part of U-Net is used as the vision decoder and text decoder. The parameters of these two decoders are shared. The initial learning rate is set to 3e-5, and the batch size is set to 32. The Adam optimizer is utilized for training with a weight decay of 1e-4. We unify all images resolution to 224 × 224, and set the hyperparameters (λ_1, λ_2) as $(0.1, 1.0)$, and (ω_1, ω_2) as $(0.2, 0.5)$. In addition, the early stop mechanism is used until the performance of model does not improve for 50 epochs.

The Dice $Dice = \sum_{m=1}^{M}\sum_{c=1}^{C}\frac{1}{MC}\cdot\frac{2|p_{mc}\cap y_{mc}|}{(|p_{mc}|+|y_{mc}|)}$ and mIoU $mIoU = \sum_{m=1}^{M}\sum_{c=1}^{C}\frac{1}{MC}\cdot\frac{|p_{mc}\cap y_{mc}|}{|p_{mc}\cup y_{mc}|}$ are used to evaluate our method and other compared methods, where C is the number of categories and M is the number of pixels.

C. Comparison with State-of-the-Arts

To verify the superiority of our method for medical image segmentation, we compare our EviVLM with widely-used state-of-the-arts approaches, including traditional segmentation methods (U-Net [27], UNet++ [40], AttUNet [41], nnUNet [42], TransUNet [43], SwinUnet [44], UCTransNet [45]) and VLM methods (ConVIRT [18], TGANet [46], CLIP [1], GLoRIA [2], ViLT [47], LAVT [58], MGCA [3], LViT [36], CARZero [55], MLIP [56], SAT [57]), as illustrated in Table I. These VLM methods use elaborately designed objective functions to address the modality gap between images and texts.

Experimental results on the QaTa-COV19 dataset show that our EviVLM achieves the best performance. In detail, EviVLM improves the Dice score by 5.37% and the mIoU score by 6.53% with lower computational cost, compared to the suboptimal nnUNet without text prompt. And it also has a 2.13% higher Dice score, a 2.23% better mIoU score than the suboptimal LViT with text prompt. This indicates that introducing EL to VLM can effectively combine image and text information to improve the segmentation performance. A similar trend is observed for the MosMedData+ dataset. On the MosMedData+ dataset, compared to the second best method, i.e., LViT, our EviVLM improves the Dice value by 3.07% and the mIoU value by 4.48%. Similarly, for Duke-Breast-Cancer-MRI dataset, it can be seen that EviVLM has a 1.12% higher Dice score than the second best method (LAVT). Overall, experimental results above demonstrate that our method outperforms both traditional segmentation methods and advanced VLM methods being capable of addressing modality gap issue.

Qualitative results show that our EviVLM has excellent segmentation capabilities compared to other state-of-the-art methods. As shown in Fig. 5, various UNet variants have more severe mis-segmentation than EviVLM, which shows that the introduction of text information can better guide the training of the model. In addition, compared with different VLM methods, EviVLM also has obvious visual advantages, which is owing to the reduction of modality gap for effective modality fusion.

To demonstrate that the Evidential Learning (EL) provides benefits over simpler fusion strategies, we compared EL with several simple fusion methods, including add, concatenate, and ensemble. For add operation, we add the predicted probabilities of the outputs from the vision decoder and text decoder,

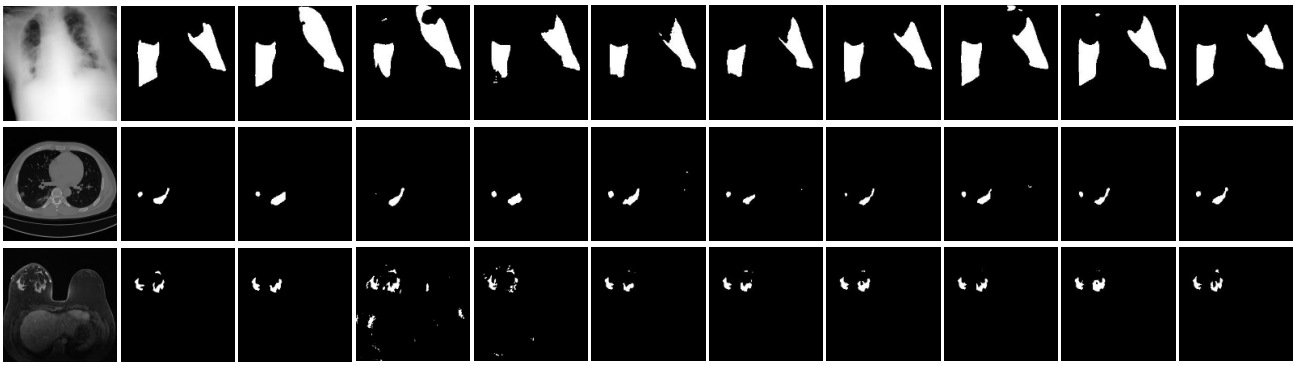

Fig. 5. Our method shows superior segmentation performance on the QaTa-COV19 (first line), MosMedData+ (second line), and Duke-Breast-Cancer-MRI (third line) datasets, compared to various UNet variants and VLM methods.

TABLE II

ABLATION STUDIES ON THE QATA-COV19, MOSMEDDATA+, AND DUKE-BREAST-CANCER-MRI TESTING DEMONSTRATE THE SIGNIFICANT IMPROVEMENTS OF THE PROPOSED INNOVATIONS. THE BEST VALUES ARE IN BOLD.

Number	Methods					QaTa-COV19		MosMedData+		Duke-Breast-Cancer-MRI		Flops (G)
	Backbone	Text	EL	EAMG	EDSL	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)	
No. 1	✓					79.02	69.46	64.60	50.73	84.23	75.41	25.2
No. 2	✓	✓				79.68	70.55	70.89	59.45	85.22	74.06	40.9
No. 3	✓	✓	✓			83.41	74.22	74.62	62.94	85.74	75.36	41.6
No. 4	✓	✓	✓	✓		85.11	76.31	75.93	63.77	86.43	76.01	42.0
No. 5	✓	✓	✓		✓	84.50	76.07	75.33	63.48	86.11	75.92	41.7
No. 6	✓	✓	✓	✓	✓	**85.79**	**77.34**	**77.64**	**65.81**	**86.96**	**76.29**	42.1

TABLE III

COMPARISON BETWEEN OUR METHOD AND OTHER VLMS ON THE QATA-COV19, MOSMEDDATA+, AND DUKE-BREAST-CANCER-MRI TESTING DEMONSTRATES THE SUPERIORITY OF OUR METHOD FOR MODALITY GAP, WHERE COSINE SIMILARITY VALUES ARE REPORTED. THE BEST VALUES ARE IN BOLD.

Method	Cosine Similarity Value (%)		
	QaTa-COV19	MosMedData+	Duke-Breast-Cancer-MRI
ConVIRT [18]	51.23	48.47	51.27
TGANet [46]	46.55	46.52	49.34
CLIP [1]	50.36	50.61	52.63
GLoRIA [2]	50.68	49.29	52.06
ViLT [47]	48.12	49.66	50.81
LAVT [58]	50.47	51.02	51.69
MGCA [3]	51.08	**52.41**	51.14
CARZero [55]	52.11	51.03	52.32
MLIP [56]	51.45	51.64	51.97
SAT [57]	52.04	51.10	52.57
Ours	**53.69**	51.12	**55.07**

TABLE IV

ABLATION STUDY OF PARAMETERS ($\lambda_1, \lambda_2, \omega_1, \omega_2$). DICE VALUES ON THE QATA-COV19, MOSMEDDATA+, AND DUKE-BREAST-CANCER-MRI DATASETS ARE REPORTED. BOLD VALUES ARE THE DICE VALUES CORRESPONDING TO THE OPTIMAL PARAMETERS.

$\lambda_1 =$	0	0.1	0.2	0.5	0.7	1.0
QaTa-COV19	85.66	**85.79**	85.11	85.71	85.33	84.55
MosMedData+	77.12	**77.64**	76.32	76.06	76.45	75.81
Duke-Breast-Cancer-MRI	86.47	**86.96**	86.41	86.64	86.27	86.30
$\lambda_2 =$	0	0.1	0.2	0.5	0.7	1.0
QaTa-COV19	85.20	85.62	85.31	85.62	85.41	**85.79**
MosMedData+	76.63	76.43	76.54	77.21	78.21	**77.64**
Duke-Breast-Cancer-MRI	86.22	86.40	85.26	86.44	86.71	**86.96**
$\omega_1 =$	0	0.1	0.2	0.5	0.7	1.0
QaTa-COV19	85.11	85.24	**85.79**	85.14	84.86	84.33
MosMedData+	75.33	76.35	**77.64**	77.32	76.05	75.62
Duke-Breast-Cancer-MRI	86.11	86.24	**86.96**	86.09	85.77	85.89
$\omega_2 =$	0	0.1	0.2	0.5	0.7	1.0
QaTa-COV19	80.31	84.52	85.64	**85.79**	84.76	83.23
MosMedData+	72.40	75.63	78.07	**77.64**	77.22	76.63
Duke-Breast-Cancer-MRI	84.06	85.27	86.53	**86.96**	86.11	85.74

and feed the added probabilities into the sigmoid activation function to obtain the final prediction result. For concatenate operation, we concatenate the output predicted probabilities of the vision decoder and text decoder according to feature dimensions and feed them into a convolutional layer to obtain the final prediction probability. For ensemble, we perform decision fusion on the output prediction probabilities of the vision decoder and the text decoder, i.e., weighted average of these two output prediction probabilities. As shown in Table VI, the EL method obtained the optimal Dice metrics on three datasets, proving the superiority of the EL method over other fusion methods.

D. Ablation Studies

Extensive ablation experiments in Table II are performed on three public datasets to verify the contribution of each component. U-Net is adopted as the backbone.

The text prompt facilitates more accurate lesion localization. The improvement of text prompt is investigated between Method No. 1 and No. 2. We can see that, by introducing text prompts, the model can improve the Dice by 0.66%, 6.29%,

and 0.99% on the QaTa-COV19, MosMedData+, and Duke-Breast-Cancer-MRI datasets, respectively. It demonstrates that the text prompt can effectively help locate lesion region.

The cross-modal evidences promote more effective modality fusion. The comparison between Method No. 2 and No. 3 can further verify the effectiveness of cross-modal evidence learning. By modeling cross-modal semantics as cross-modal evidences, we obtain better modality fusion effect for superior segmentation, which increases the Dice by 3.73%, 3.73%, and 0.52% on the QaTa-COV19, MosMedData+, and Duke-Breast-Cancer-MRI datasets, respectively.

EAMG drives complementary vision and text evidence embedding learning. Table II shows that Method No. 4 outperforms Method No. 3 by 1.70% in Dice and 2.00% in mIoU by equipping with EAMG on the QaTa-COV19 dataset. Unlike Method No. 3, which only uses the encoded evidence embedding for modality fusion, we can effectively refine both modality-specific evidence embedding by injecting a global cross-modal pixel-level affinity map into evidence embedding, learning more complementary vision and text evidences. In addition, although EAMG increased Flops by 0.4G, it improved Dice by 1.70%, 1.31%, and 0.69% on the QaTa-COV19, MosMedData+, and Duke-Breast-Cancer-MRI datasets, respectively.

EDSL boosts the alignment between image and text evidence embeddings. Method No. 5 in Table II demonstrates that adding EDSL to Method No. 3 improves the Dice by 1.09%, 0.71%, and 0.37% on the QaTa-COV19, MosMedData+, and Duke-Breast-Cancer-MRI datasets, respectively. These improvements indicate that the introduction of EDSL strengthens the consistency learning between image and text evidence embeddings by learning variation inconsistency of their similarity matrices.

E. Modality Gap Analysis

we calculate the Euclidean distance and cosine similarity between image and text features to measure the modality gap.

To substantiate that our EviVLM can effectively reduce the modality gap, we visualize image and text embeddings on the QaTa-COV19, MosMedData+, and Duke-Breast-Cancer-MRI datasets. As illustrated in Fig. 6, a substantial modality gap is observable in the data distribution using CLIP without EL. By contrast, our EviVLM evidently makes image and text embeddings closer, which reduces the Euclidean distances by 11100.85, 1741.19, and 2463.67 on the QaTa-COV19, MosMedData+, and Duke-Breast-Cancer-MRI testing samples, bridging the modality gap. It indicates that our EviVLM is able to narrow the modality gap by measuring the uncertainty of the aggregated opinion.

Table III demonstrates the effectiveness of our method in addressing the modality gap issue. The cosine similarity is obtained by calculating image-text evidence embeddings. It can be seen that EviVLM improves the cosine similarity value by 1.58% compared to the suboptimal CARZero on the QaTa-COV19 dataset. On the Duke-Breast-Cancer-MRI dataset, compared to the second best method, i.e., MGCA, our EviVLM improves the cosine similarity value by 1.93%. This shows the effectiveness of incorporating EL to VLM for consistent cross-modality learning, and further explains segmentation improvements attributing to narrowing modality gap.

TABLE V

ABLATION STUDY OF PARAMETERS WHEN CONVERTING EVIDENCE INTO DIRICHLET DISTRIBUTION. DICE VALUES ON THE QATA-COV19, MOSMEDDATA+, AND DUKE-BREAST-CANCER-MRI DATASETS ARE REPORTED. BOLD VALUES ARE THE DICE VALUES CORRESPONDING TO THE OPTIMAL PARAMETERS.

Values	QaTa-COV19	MosMedData+	Duke-Breast-Cancer-MRI
0.1	85.12	77.24	86.31
0.5	85.50	77.01	85.74
1.0	**85.79**	**77.64**	**86.96**
2.0	85.49	76.85	86.22

TABLE VI

EXPERIMENTAL RESULTS OF DIFFERENT FUSION METHODS. DICE VALUES ON THE QATA-COV19, MOSMEDDATA+, AND DUKE-BREAST-CANCER-MRI DATASETS ARE REPORTED. BOLD VALUES ARE THE DICE VALUES CORRESPONDING TO THE OPTIMAL METHOD.

Method	QaTa-COV19	MosMedData+	Duke-Breast-Cancer-MRI
Add	79.68	70.89	85.22
Concatenate	80.14	71.53	84.74
Ensemble	82.06	72.83	85.27
EL (Ours)	**83.41**	**74.62**	**85.74**

TABLE VII

EXPERIMENTAL RESULTS IN 3D DATASET (BRATS 2020). DICE VALUES ARE REPORTED. BOLD VALUES ARE THE DICE VALUES CORRESPONDING TO THE OPTIMAL METHOD. ET, WT, AND TC REPRESENTS ENHANCING TUMOR, WHOLE TUMOR, AND TUMOR CORE.

Methods	ET (%)	WT (%)	TC (%)	Mean (%)
3D UNet [61]	79.41	89.22	85.13	84.59
3D EviVLM (Ours)	**81.46**	**90.70**	**85.67**	**85.94**

F. Parameter Selection

For four-weighted loss functions ($\mathcal{L}_{diff}, (\sum_{i=1}^{B} \mathcal{L}_{V2T}^{i} + \sum_{j=1}^{B} \mathcal{L}_{T2V}^{j}), \mathcal{L}_{edsl}, \mathcal{L}_{evi}$), we also conduct ablation experiments on the QaTa-COV19, MosMedData+, and Duke-Breast-Cancer-MRI datasets to explore the impact of four parameters $(\lambda_1, \lambda_2, \omega_1, \omega_2)$. Each parameter is varied across six different values (0, 0.1, 0.2, 0.5, 0.7, 1.0). According to the experimental results shown in Table IV, λ_1 and λ_2 show optimal performance at 0.1 and 1.0 respectively, whereas ω_1 and ω_2 peak at 0.2 and 0.5 respectively. It can be observed that the changes of three parameters $(\lambda_1, \lambda_2, \omega_1)$ do not have a significant impact on the results. On the contrary, when the parameter ω_2, which controls the loss of evidence learning, is set to 0, the Dice value drops sharply 5.48%, 5.42%, and 2.90% on the QaTa-COV19, MosMedData+, and Duke-Breast-Cancer-MRI datasets, compared to $\omega_2 = 0.5$. This indicates that introducing evidence learning has significant advantages in medical image segmentation.

TABLE VIII
COMPARISON BETWEEN OUR METHOD AND OTHER VLMS ON THE MONUSEG DATASET DEMONSTRATES THE SUPERIORITY OF OUR METHOD WHEN THE TEXT DESCRIPTION IS COMPLEX. THE BEST VALUES ARE IN BOLD.

Method	MoNuSeg	
	Dice (%)	mIoU (%)
STPNet [62]	80.22	69.43
ConVIRT [18]	78.52	68.39
TGANet [46]	79.96	68.77
CLIP [1]	78.17	68.06
GLoRIA [2]	79.29	69.50
ViLT [47]	79.33	67.93
LAVT [58]	79.06	68.65
MGCA [3]	79.42	68.20
LViT [36]	80.34	69.11
CARZero [55]	79.14.	69.14
MLIP [56]	78.62	68.57
SAT [57]	79.61	68.61
Ours	**81.47**	**69.82**

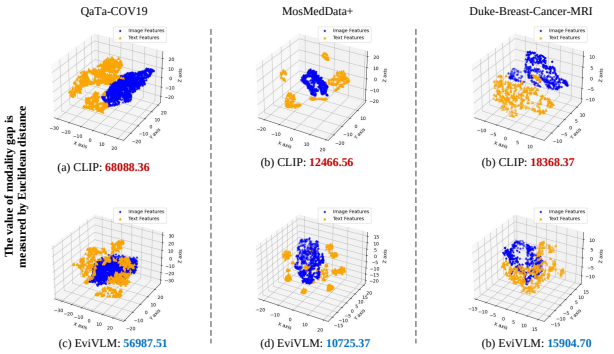

Fig. 6. Our EviVLM reduces the modality gap by 11100.85, 1741.19, and 2463.67 on the QaTa-COV19, MosMedData+, and Duke-Breast-Cancer-MRI datasets, compared to CLIP.

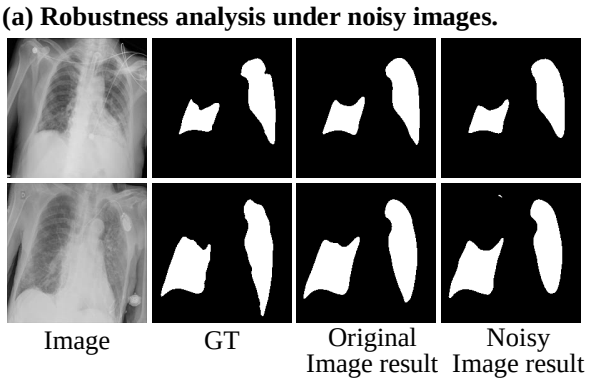

Fig. 7. The robustness analysis based on noisy images and ambiguous text.

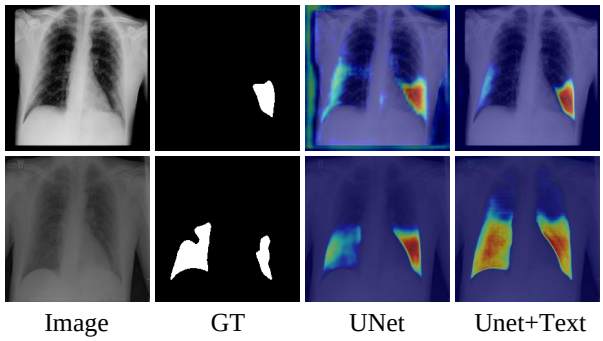

Fig. 8. Saliency map for interpretability study of text information. It is obvious that the introduction of text information can better locate the target region.

G. Robustness Analysis and Interpretability Study of Text Information

To validate the robustness of our EviVLM, we conduct two set experiments under challenging conditions: (1) We added Gaussian noise to the test image to evaluate the robustness of the model to the input image; (2) We use ambiguous text prompts to validate the robustness of the model to textual information. As shown in Fig. 7(a), compared with the prediction results of the original image, the prediction results of the image with noise are still highly similar to the ground truth, proving the robustness of our EviVLM to noisy images. In Fig. 7(b), our method can still effectively segment the target area in the presence of text ambiguity.

To demonstrate whether the introduction of text information can enhance the attention to target regions, we provide visual results of text-guided segmentation. As shown in Fig. 8, we conduct experiments on two cases. We perform activation mapping at the final layer of the segmentation network. The text information is input to the model through cross-attention module, thus the difference in activation regions between UNet and UNet+Text can be approximated as the difference brought by the text information. It can be seen that the activation effect of the region of interest of UNet+Text is similar to that of Ground Truth. It is also worth noting that the UNet fails to accurately activate the target regions. It indicates that the text information can effectively help locate target region, thus prompting the network to pay more attentions on the region indicated by the text information.

H. Ablation Study of Converting Evidence into Dirichlet Distribution

To demonstrate that adding 1 in our task is the optimal value for converting evidence into a Dirichlet distribution, we conduct relevant ablation experiments. Specifically, we selected four different values (i.e. 0.1, 0.5, 1.0, and 2.0) to train the proposed EviVLM and reported Dice values on the QaTa-COV19, MosMedData+, and Duke-Breast-Cancer-MRI datasets, respectively. As shown in Table V, when adding 1.0 to convert evidence into Dirichlet distribution, we obtain the optimal Dice on three datasets, i.e., 85.79% for QaTa-COV19, 77.64% for MosMedData+, and 86.96% for Duke-Breast-Cancer-MRI. This indicates that adding 1.0 is the optimal choice when converting evidence into Dirichlet distribution.

I. 3D Dataset Segmentation Analysis

To verify the superiority of the proposed method in 3D segmentation tasks, we conducted experiments on a 3D dataset, i.e., BraTS 2020 dataset [60]. It is divided into three cohorts: Training, Validation, and Testing. The Training dataset is composed of multi-parametric MRI (mpMRI) scans from 369 diffuse glioma patients. Each mpMRI set contains four different sequences: native T1-weighted (T1), post-contrast T1 weighted (T1ce), T2-weighted (T2), and T2 Fluid-Attenuated-Inversion-Recovery (FLAIR). In our work, we conduct the five fold cross validation experiment on the training set and report the Dice values of enhancing tumor (ET), whole tumor (WT), and tumor core (TC). As shown in Table VII, compared with the 3D UNet [61], our EviVLM consistently achieves improvement, with Dice increases of 2.05%, 1.48%, and 1.35% on ET, WT, and TC, respectively. This further proves the effectiveness of our method in 3D segmentation task.

J. The generalizability for free-form text descriptions

To demonstrate generalizability of EviVLM for diverse datasets with complex free-form medical text descriptions, we conduct experiments on the MoNuSeg [63] dataset. MoNuSeg includes 44 images with annotations provided by Li *et al.* [36], and the image size is 1000×1000. The text annotations are simple, for example, "The nuclei are evenly distributed." To make the text annotations more comprehensive and detailed, we send the text annotations to ChatGPT and output the modified text descriptions. For the text annotation "The nuclei are evenly distributed.", the modified text description is "The nuclei are uniformly distributed across the field, showing consistent size, shape, and staining intensity without evident clustering or sparse regions.".

We conduct comparative experiments with several state-of-the-art VLM methods, using the modified text descriptions. As shown in Table VIII, Our EviVLM outperforms other VLM methods, with our Dice and mIoU improving by 1.13% and 0.71%, respectively, compared to the suboptimal LViT. This indicates the effectiveness of our method on the diverse dataset with complex text descriptions, demonstrating its generalization ability.

V. Limitation

In this work, we employ evidential learning to estimate the modality gap between images and texts by quantifying the uncertainty of aggregated opinions. While this approach offers a novel and interpretable way to capture modality mismatch, we acknowledge its limitations. The proposed uncertainty is an indirect indicator rather than a direct measurement of the modality gap. Specifically, the evidential networks transform image and text features into evidence vectors, and their aggregation produces subjective opinions whose uncertainty reflects the degree of modality gap. This uncertainty is correlated with semantic misalignment but does not provide an explicit measurement of the gap. Therefore, our method is not able to deliver direct quantification of the modality gap.

In the future, several directions can further strengthen this line of research. One possibility is to complement uncertainty-based indicators with direct measurement of cross-modal alignment, such as embedding distance, optimal transport, or mutual information metrics, thereby combining interpretability with explicit metric. Another promising direction is to develop hierarchical evidential models to disentangle different sources of misalignment (e.g., semantic, structural, or domain-specific gaps). Finally, integrating these gap estimations into the model training may allow adaptive alignment strategies, dynamically reducing modality gap.

We believe that despite uncertainty from aggregated opinions is an indirect indicator, it provides a theoretically grounded and practically useful tool for modality gap, opening a new perspective on quantifying modality gap in medical vision-language models.

In addition, the text prompts are almost designed in a fixed format. In future studies, we plan to extend our approach to experiments on multiple datasets with more complex and diverse text descriptions, thus better evaluating its generalizability.

VI. Conclusion

In this work, we propose a novel VLM paradigm, EviVLM, by pioneeringly introducing EL to VLM, which aims to bridge the modality gap between image and text for cross-modal fusion. To collect complementary cross-modal evidence, a EAMG is proposed to refine both modality-specific evidence embeddings by learning a global cross-modal affinity map. To ensure the consistency for cross-modal evidences, a EDSL is proposed to boost vigorous alignment between cross-modal evidence embeddings by measuring variation inconsistency of similarity matrices. Finally, the collected cross-modal evidences are transformed to opinions for modality gap estimation. Extensive experiments show that our method surpasses traditional segmentation methods without text information. More importantly, our method also outperforms advanced VLM methods due to narrowing modality gap more efficiently.

References

[1] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Mach. Learn. Res. (PMLR)*, Jul. 2021, pp. 8748-8763.

[2] S. -C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3922-3931.

[3] F. Wang, Y. Zhou, S. Wang, V. Vardhanabhuti, and L. Yu, "Multi-granularity cross-modal alignment for generalized medical visual representation learning," *in Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, Nov. 2022, pp. 33536–33549.

[4] J. Liu *et al.*, "Clip-driven universal model for organ segmentation and tumor detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 21152-21164.

[5] V.W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J.Y. Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," *in Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, Nov. 2022, pp. 17612-17625.

[6] P. Müller, G. Kaissis, C. Zou, and D. Rueckert, "Joint learning of localized representations from medical images and reports," *in Proc. European Conference on Computer Vision. (ECCV)*, Aug. 2022, pp. 685-701.

[7] L. Bao *et al.*, "Multi-Granularity Matching Transformer for Text-Based Person Search," *IEEE Trans. Multimedia.*, vol. 26, pp. 4281-4293, Oct. 2023.

[8] J. Guo, J. Ye, Y. Xiang, and Z. Yu, "Layer-Level Progressive Transformer With Modality Difference Awareness for Multi-Modal Neural Machine Translation," *IEEE/ACM Trans Audio Speech Lang Process.*, pp. 3015-3026, Aug. 2023.

[9] A. Malinin, and M. Gales, "Predictive uncertainty estimation via prior networks," *in Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, Nov. 2018.

[10] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *in Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 31, Nov. 2018.

[11] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep evidential regression," *in Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, Nov. 2020, pp. 14927-14937.

[12] W. Bao, Q. Yu, and Y. Kong, "Evidential deep learning for open set action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13349-13358.

[13] W. Bao, Q. Yu, and Y. Kong, "Opental: Towards open set temporal action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2979-2989.

[14] Z. Han, C. Zhang, H. Fu, J.T. Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2551-2566, May. 2022.

[15] J. Pei, A. Men, Y. Liu, X. Zhuang, and Q. Chen, "Evidential multi-source-free unsupervised domain adaptation," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5288-5305, Feb. 2024.

[16] A. Conneau, and G. Lample, "Cross-lingual language model pretraining," *in Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, Nov. 2019.

[17] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, Dec. 2022, pp. 3876-3887.

[18] Y. Zhang, H. Jiang, Y. Miura, C.D. Manning, and C.P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Machine Learning for Healthcare Conference.*, 2022, pp. 2-25.

[19] X. Hu *et al.*, "Scaling up vision-language pre-training for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17980-17989.

[20] B. Li, Z. Han, H. Li, H. Fu, and C. Zhang, "Trustworthy long-tailed classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 6970-6979.

[21] A.P. Dempster, "A generalization of Bayesian inference," *J. R. Stat. Soc. B*, vol. 30, no. 2, pp. 205-232, 1968.

[22] A. Jsang, "Subjective Logic: A formalism for reasoning under uncertainty,", 2018.

[23] L. Huang, T. Denoeux, P. Vera, and S. Ruan, "Evidence fusion with contextual discounting for multi-modality medical image segmentation" in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 401-411.

[24] H. Li, Y. Nan, J. Del Ser, and G. Yang, "Region-based evidential deep learning to quantify uncertainty and improve robustness of brain tumor segmentation," *Neural Comput. Appl.*, vol. 35, no. 30, pp. 22071-22085, Nov. 2023.

[25] A. Kirillov *et al.*, "Segment anything," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 4015-4026.

[26] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of the 2019 conf. of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171-4186.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[28] E. Alsentzer *et al.*, "Publicly available clinical BERT embeddings," 2019, *arXiv:1904.03323.*

[29] YC. Chen *et al.*, "Uniter: Learning universal image-text representations," *in Proc. European Conference on Computer Vision. (ECCV)*, Sep. 2020, pp. 104-120.

[30] P. Domingos, "A unified bias-variance decomposition," in*Proc. Conf. International Conference on Machine Learning. (ICML)*, 2000, pp. 231-238.

[31] A. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748.*

[32] A. Jøsang, "Subjective logic," vol.3, 2016.

[33] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *in Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, Nov. 2018.

[34] A. Jøsang and R. Hankin, "Interpretation and fusion of hyper opinions in subjective logic," *in Proc. Int. Conf. Inf. Fusion (ICIF)*, 2012, pp. 1225-1232.

[35] A. Degerli, S. Kiranyaz, MEH. Chowdhury, and M. Gabbouj, "Osegnet: Operational segmentation network for Covid-19 detection using chest X-ray images," in *Proc. Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 2306-2310.

[36] Z. Li *et al.*, "LViT: Language Meets Vision Transformer in Medical Image Segmentation," *IEEE Trans. Med. Imag.*, vol. 43, no. 1, pp. 96-107, Jan. 2024.

[37] S. Morozov *et al.*, "Mosmeddata: Chest ct scans with COVID-19 related findings dataset," 2020, *arXiv:2005.06465.*

[38] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," *Eur. Radio. Exp.*, vol. 4, pp. 1-13, 2020.

[39] A. Saha *et al.*, "A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features," *Br. J. Cancer*, vol. 119, no. 4, pp. 508-516, 2018.

[40] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 1-11, 2018.

[41] O. Oktay *et al.*, "Attention u-net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999.*

[42] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen and K.H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nat. Methods.*, vol. 18, no. 2, pp. 203-211, 2021.

[43] J. Chen *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306.*

[44] H. Cao *et al.*, "Swin-unet: Unet-like pure transformer for medical image segmentation," *in Proc. European Conference on Computer Vision. (ECCV)*, Aug. 2022, pp. 205-218.

[45] H. Wang, P. Cao, J. Wang, and O.R. Zaiane, , "Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2022, vol. 36, no. 3, pp. 2441–2449.

[46] N.K. Tomar, D. Jha, U. Bagci, and S. Ali, "TGANet: Text-guided attention for improved polyp segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2022, pp. 151-160.

[47] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in*Proc. Conf. International Conference on Machine Learning. (ICML)*, 2021, pp. 5583-5594.

[48] J. Hu *et al.*, "LGA: A Language Guide Adapter for Advancing the SAM Model's Capabilities in Medical Image Segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2024, pp. 610-620.

[49] Z. Qin, H. Yi, Q. Lao, and K. Li, "Medical image understanding with pretrained vision language models: A comprehensive study," 2022, *arXiv:2209.15517.*

[50] Z. Diao, H. Jiang, X. Han, Y. Yao, and T. Shi, "EFNet: evidence fusion network for tumor segmentation from PET-CT volumes," *Phys. Med. Biol.*, vol. 66, no. 20, pp. 205005, 2021.

[51] B. Liu *et al.*, "Improving Medical Vision-Language Contrastive Pretraining With Semantics-Aware Triage," *IEEE Trans. Med. Imag.*, vol. 42, no. 12, pp. 3579-3589, Jul. 2023.
[52] P. Wang, H. Zhang, and Y. Yuan, "MCPL: Multi-Modal Collaborative Prompt Learning for Medical Vision-Language Model," *IEEE Trans. Med. Imag.*, vol. 43, no. 12, pp. 4224-4235, Jun. 2024.
[53] Y. Liu *et al.*, "MERIT: Multi-view evidential learning for reliable and interpretable liver fibrosis staging," *Med. Image Anal.*, vol. 102, May. 2025, Art. no. 103507.
[54] J. Lee, S. Chung, S. Kim, H. Kang, and K. Sohn, "Looking into Your Speech: Learning Cross-modal Affinity for Audio-visual Speech Separation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1336-1345.
[55] H. Lai *et al.*, "Carzero: Cross-attention alignment for radiology zero-shot classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 11137-11146.
[56] Z. Li *et al.*, "Mlip: Enhancing medical visual representation with divergence encoder and knowledge-guided contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 11704-11714.
[57] B. Liu *et al.*, "Improving medical vision-language contrastive pretraining with semantics-aware triage," *IEEE Trans. Med. Imag.*, vol. 42, no. 12, pp. 3579-3589, Jun. 2023.
[58] Z. Yang *et al.*, "Lavt: Language-aware vision transformer for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18155-18165.
[59] Y. Cho, H. Yu, and S. Kang, "Cross-aware early fusion with stage-divided vision and language transformer encoders for referring image segmentation," *IEEE Trans. Multimedia.*, vol. 26, pp. 5823-5833, 2023.
[60] BH. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993-2024, Jun. 2014.
[61] Ö. Çiçek, A. Abdulkadir, S. S.Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: learning dense volumetric segmentation from sparse annotation" in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016, pp. 424-432.
[62] D. Shan *et al.*, "STPNet: Scale-aware Text Prompt Network for Medical Image Segmentation," *IEEE Trans Image Process.*, vol. 34, pp. 3169-3180, May. 2025.
[63] N. Kumar *et al.*, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE Trans. Med. Imag.*, vol. 36, no. 7, pp. 1550-1560, Mar. 2017.