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

Anonymous submission

Abstract

Modality gap between images and text has become a bottleneck of Vision Language Model (VLM) in medical image segmentation. Such modality gap hinders multi-modal fusion for segmentation task. To bridge this modality gap, we propose Evidence-driven Vision Language Model (EviVLM), a novel paradigm that pioneeringly injects Evidential Learning (EL) into VLM, which aggregates evidence-transformed image-text opinions to estimate this modality gap for multi-modal fusion. To drive this paradigm by collecting reliable and consistent cross-modal evidences, we further propose an Evidence Affinity Map Generator (EAMG) to learn a global cross-modal affinity map for modality-specific evidence embedding refinement, and an Evidence Differential Similarity Learning (EDSL) to boost strengthful alignment between image and text evidence embeddings by measuring variation inconsistency of bidirectional similarity matrices. Finally, the subjective logic is used for mapping evidences to opinions, and a Dempster-Shafer's evidence theory based combination rule is introduced for opinion aggregation, thus measuring the modality gap for sufficient modality fusion. Our code will be available.

Introduction

Vision Language Model (VLM) aims to align image-text pairs to facilitate multi-modal understanding (Radford et al. 2021). 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 (Huang et al. 2021; Wang et al. 2022a; Liu et al. 2023).

Although the progresses in VLM driven medical image segmentation, the modality gap between vision and text is still significant (Fig. 1(a)) (Liang et al. 2022). 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.

To bridge the modality gap, most previous methods (Huang et al. 2021; Wang et al. 2022a; Müller et al. 2022),

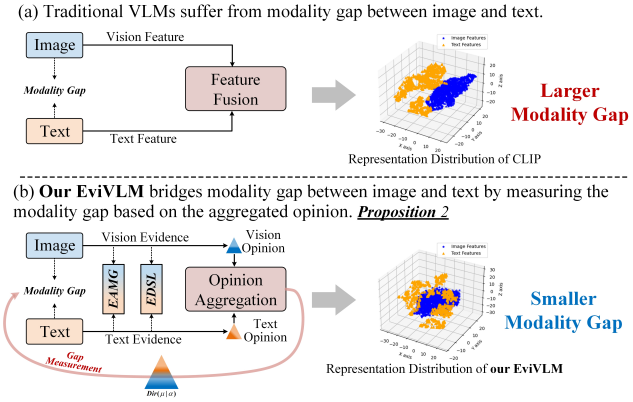

Figure 1: Our EviVLM bridges the large modality gap between visual and textual representations by injecting EL into VLM for the estimation of such modality gap.

generally utilize two separate encoders to extract cross-modal features. 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 (Bao et al. 2023), since the narrow cone of embedding space restricts the expression of complex semantic relationships between images and texts.

Therefore, we argue that narrowing the modality gap is necessary for cross-modal fusion since the closer the modality gap between images and texts, the more effective visual and textual features can be fused in multi-modal feature space (Guo et al. 2023). For each image-text pair, the goal is to learn 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. With the rich vision-language correlations in off-the-shelf foundation models such as CLIP (Radford et al. 2021), a straight solution is to collect image-text vision-language representations using the pretrained CLIP. 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.

Motivated by the above observations, we aim to reduce the modality gap to achieve superior segmentation performance. Recently, Evidential Learning (EL) (Malinin and Gales 2018; Sensoy, Kaplan, and Kandemir 2018a), which can quantify uncertainty in model outputs trustfully by collecting subjective evidence, has attracted increasing attention and been successfully used in a variety of tasks (Amini et al. 2020; Bao, Yu, and Kong 2021, 2022). In this study, 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 reliability 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:

- 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.
- To collect reliable cross-modal evidences, EAMG is proposed to produce a global cross-modal affinity map, refining both modality-specific evidence embeddings.
- To ensure the consistency for cross-modal evidences, EDSL is proposed to measure variation inconsistency of bidirectional similarity matrices, boosting vigorous alignment between cross-modal evidence embeddings.
- 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.

Related Work

Vision Language Model. With the success of large language models, vision language models (VLMs) adopt BERT-like architecture (Conneau and Lample 2019) contrastive learning paradigm (Radford et al. 2021; Wang et al. 2022d) 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 (Huang et al. 2021), image-text retrieval (Zhang et al. 2022a), image captioning (Hu et al. 2022) and so on. 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.

Evidential Learning in Image Segmentation. EL, a method for reliable model inference and uncertainty estimation, has been gained increasing attention (Li et al. 2022). The core of EL is the notion of evidence, which reflects the level of belief or support for different hypotheses within the model's predictions through Dempster-Shafer evidence theory (Dempster 1968) and Subjective Logic theory (Jsang 2018). For the multi-modality segmentation task, Huang et al. (2022) 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. Li et al. (2023a) 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 study, we pioneeringly inject EL into VLM to reduce the modality gap.

Method

Problem Definition

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 prediction uncertainties u of vision and text opinions is aggregated for modality gap measurement.

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 modality fusion. EAMG and EDSL are proposed to drive reliable and consistent evidence learning.

Evidence Embedding Extraction

Given an image-text pair $\{V, T\}$, the encoding path of U-Net (Ronneberger, Fischer, and Brox 2015) and BioClinicalBERT (Alsentzer et al. 2019) 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) (Chen et al. 2019). 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,

$$x_e^T = \alpha^{V2T}(\tilde{V}x^T), \alpha^{V2T} = \text{softmax}(\frac{(\tilde{Q}x_e^V)^T(\tilde{K}x^T)}{\sqrt{d}}), \tag{1}$$

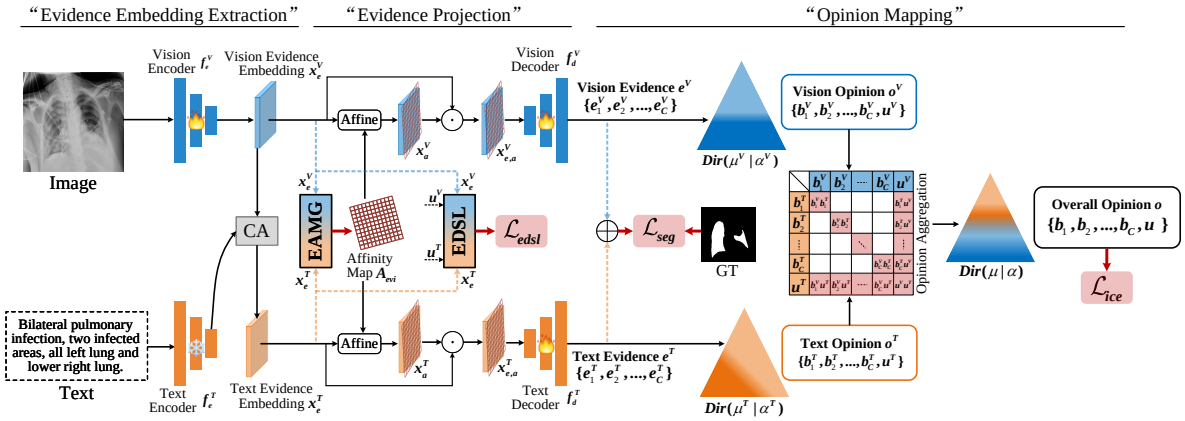

Figure 2: Illustration of EviVLM. Our EviVLM brides the modality gap by measuring the uncertainty of the aggregated opinion. The EAMG produces a affinity map to refine both modality-specific evidence embeddings. The EDSL learns the magnitude difference and variation inconsistency between similarity matrices to balanced modality representations.

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

Evidence Affinity Map Generator (EAMG)

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.

$$A^V_{evi} = \text{NonLocal}(x^V_e), \quad A^T_{evi} = \text{NonLocal}(x^T_e). \tag{2}$$

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 $A_{evi} \in \mathbb{R}^{HW\times HW}$,

$$A_{evi} = w^V * A^V_{evi} + w^T * A^T_{evi}, \tag{3}$$

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 $x^T_{e,a}$,

$$x^V_{e,a} = \text{affine}(A_{evi}, x^V_e) \odot x^V_e, \tag{4}$$

$$x^T_{e,a} = \text{affine}(A_{evi}, x^T_e) \odot x^T_e, \tag{5}$$

where $\text{affine}(\cdot)$ 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,

$$\text{affine}(A_{evi}, x^V_e) = \sum_h^H \sum_w^W A_{evi}(h,w) \cdot x^V_e(h,w), \tag{6}$$

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 reliability of cross-modal evidence learning.

Evidence Differential Similarity Learning (EDSL)

Our EDSL (Fig. 4) performs the Bias-Variance Decomposition based on the differential matrix to learn the magnitude difference and 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

$$s_{ij} = cos(u^V \odot x^V_{e,i}, u^T \odot x^T_{e,j}), \tag{7}$$

$$\tilde{s}_{ji} = cos(u^T \odot x^T_{e,j}, u^V \odot x^V_{e,i}), \tag{8}$$

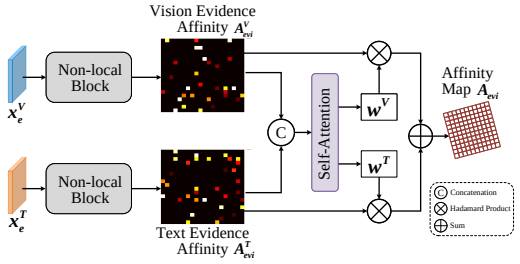

Figure 3: EAMG learns cross-modal evidence affinities via two non-local block, and integrates such two modality-specific affinity maps with a self-attention to produce a global cross-modal affinity map.

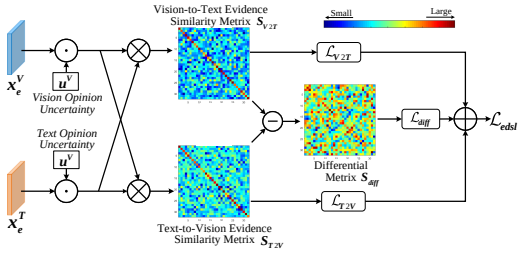

Figure 4: EDSL performs the Bias-Variance Decomposition based on the differential matrix to learn the magnitude difference and variation inconsistency between two cross-modal similarity matrices, thus boosting the alignment between image and text evidence embeddings.

where $cos(\cdot)$ measures the cosine similarity, i and j represent the ith image and jth text respectively. $S_{V2T} = [s_{ij}] \in \mathbb{R}^{B\times B}$ 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 magnitude difference and 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 (Domingos 2000) based on S_{diff},

$$S_{diff} = f_{BVD}(S_{V2T}, S_{T2V}) \tag{9}$$

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

$$\begin{aligned}
&\mathcal{L}_{diff} = \|D\|^2_{BVD} \approx B^2\mathbb{E}[(s-\tilde{s})^2] = B^2(\mathbb{E}[s^2] - 2\mathbb{E}[s\tilde{s}] + \mathbb{E}[\tilde{s}^2]) \\
&= B^2(\mathbb{E}^2[s] + Var(s) - 2\mathbb{E}[s]\mathbb{E}[\tilde{s}] - 2Cov(s,\tilde{s}) + \mathbb{E}^2[\tilde{s}] + Var(\tilde{s})) \\
&= B^2(\mathbb{E}^2[s] + \mathbb{E}^2[\tilde{s}] - 2\mathbb{E}[s]\mathbb{E}[\tilde{s}]) + B^2(Var(s) + Var(\tilde{s}) - 2Cov(s,\tilde{s})) \\
&= \underbrace{B^2(\mathbb{E}[s] - \mathbb{E}[\tilde{s}])^2}_{Bias} + \underbrace{B^2 Var(s-\tilde{s})}_{Variance},
\end{aligned} \tag{10}$$

where $\mathbb{E}[\cdot]$, $Var(\cdot)$, and $Cov(\cdot)$ represent the expectation, variance, and covariance operations, respectively. The computed Bias and Variance measure the magnitude difference and variation inconsistency between two cross-modal similarity matrices. Therefore, we regard $\|D\|^2_{BVD}$ as the the inconsistency differential loss $\mathcal{L}_{diff}$ of similarity matrices. Additionally, we also calculate InfoNCE (Oord, Li, and Vinyals 2018) losses $\mathcal{L}^i_{V2T}$, $\mathcal{L}^j_{T2V}$ of two cross-modal similarity matrices above to maximally preserve the mutual information between the true image-text pairs,

$$\begin{aligned}
\mathcal{L}^i_{V2T} &= -\log\frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B}\exp(s_{ij}/\tau)}, \\
\mathcal{L}^j_{T2V} &= -\log\frac{\exp(\tilde{s}_{jj}/\tau)}{\sum_{i=1}^{B}\exp(\tilde{s}_{ji}/\tau)},
\end{aligned} \tag{11}$$

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

$$\mathcal{L}_{edsl} = \lambda_1\mathcal{L}_{diff} + \lambda_2\frac{1}{2B}(\sum_{i=1}^{B}\mathcal{L}^i_{V2T} + \sum_{j=1}^{B}\mathcal{L}^j_{T2V}), \tag{12}$$

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 inconsistency learning of cross-modal similarity matrices, enhancing the consistency of cross-modal evidences.

Modality Gap Measurement

To measure the modality gap based on the aggregated opinion, we represent the decoded vision evidence $e^V = f^V_d(x^V_{e,a})$ and text evidence $e^T = f^T_d(x^T_{e,a})$ 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 (Jøsang 2016), the vision Dirichlet distribution of class probabilities $Dir(\mu^V|\alpha^V)$ is determined by the evidence $e^V = \{e^V_1, e^V_2, ..., e^V_C\}$, where C is the number of classes. Following the work (Sensoy, Kaplan, and Kandemir 2018b), we derive the parameter of the distribution through $\alpha^V = e^V + 1$. Then the vision Dirichlet distribution is mapped to the vision opinion $o^V = \{b^V_1, b^V_2, ..., b^V_C, u^V\}$, satisfying

$$u^V + \sum_{c=1}^{C} b^V_c = 1, \tag{13}$$

where $b^V_c = \frac{\alpha^V_c - 1}{S^V}$ is the belief mass for class c, $S^V = \sum_{c=1}^{C}\alpha^V_c$ 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 (Jøsang and Hankin 2012), 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

$$b_c = \frac{1}{M}(b_c^V b_c^T + b_c^V u^T + b_c^T u^V), u = \frac{1}{M} u^V u^T, \quad (14)$$

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 (Sensoy, Kaplan, and Kandemir 2018b), we use the integrated cross-entropy loss. The loss for vision opinion is given by

$$\begin{aligned}\mathcal{L}_{ice}^V &= \mathbb{E}_{\mu^V \sim Dir(\mu^V|\alpha^V)}[\mathcal{L}_{CE}(\mu^V, y)] \\ &= \int \left[\sum_{c=1}^{C} -y_c \log(\mu_c^V)\right] \frac{1}{B(\alpha^V)} \prod_{c=1}^{C} (\mu_c^V)^{\alpha_c^V - 1} \mathrm{d}\mu^V \\ &= \sum_{c=1}^{C} y_c^V (\psi(S^V) - \psi(\alpha_c^V))\end{aligned} \quad (15)$$

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

$$\mathcal{L}_{evi} = \mathcal{L}^V + \mathcal{L}^T + \mathcal{L}^{aggregated} \quad (16)$$

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

Theoretical Analysis

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.

$$\begin{aligned} b_g &= \frac{b_g^V b_g^T + b_g^V u^T + b_g^T u^V}{1 - \sum_{i \neq j} b_i^V b_j^T} \\ &= \frac{b_g^V b_g^T + b_g^V u^T + b_g^T u^V}{\sum_{c=1}^{C} b_c^V b_c^T + u^T + u^V - u^V u^T} \\ &\geq \frac{b_g^V (b_g^T + u^T + u^V)}{b_m^V (1 - u^T) + u^T + u^V - u^V u^T} \\ &\geq \frac{b_g^V (b_g^T + u^T + u^V)}{b_m^V + u^T + u^V} \geq b_g^V. \end{aligned} \quad (17)$$

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.

$$\begin{aligned} u &= \frac{u^V u^T}{1 - \sum_{i \neq j} b_i^V b_j^T} \\ &= \frac{u^V u^T}{\sum_{c=1}^{C} b_c^V b_c^T + u^T + u^V - u^V u^T} \\ &\leq \frac{u^V u^T}{u^T + u^V - u^V u^T} \leq u^V. \end{aligned} \quad (18)$$

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.

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,

$$\mathcal{L}_{seg} = -\frac{1}{HW} \sum_{m=1}^{HW} \mathcal{L}_{CE}(\sigma(e_m^V + e_m^T), y_m), \quad (19)$$

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:

$$\mathcal{L}_{overall} = \omega_1 \mathcal{L}_{edsl} + \omega_2 \mathcal{L}_{evi} + \omega_3 \mathcal{L}_{seg}, \quad (20)$$

where ω_1, ω_2, and ω_3 are hyperparameters to balance three objective functions.

Experiments

Datasets

Three public datasets are adopted for evaluation, including

1. **QaTa-COV19** dataset (Degerli et al. 2022) 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 (Li et al. 2023b). The text annotations focus on whether both lungs are infected, the number of lesion regions, and the approximate location of the infected areas. Following (Li et al. 2023b), we divide this dataset into 7:1:2 for training, validation, and testing respectively.
2. **MosMedData+** dataset (Morozov et al. 2020; Hofmanninger et al. 2020) contains 2729 CT scan slices of lung infections. The text annotations are also given by (Li et al. 2023b). Following (Li et al. 2023b), we divide this dataset into 7:1:2 for training, validation, and testing.
3. **Duke-Breast-Cancer-MRI** dataset (Saha et al. 2018) contains 922 MRI scans from 922 breast cancer patients. The text annotations are provided and verified by two professionals. We divide this dataset into 7:1:2 for training, validation, and testing respectively.

Table 1: 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. The best values are in bold.

Method	Text	QaTa-COV19		MosMedData+		Duke-Breast-Cancer-MRI	
		Dice (%)	mIoU (%)	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)
U-Net	×	79.02	69.46	64.60	50.73	84.23	75.41
UNet++	×	79.62	70.25	71.75	58.39	84.52	76.96
AttUNet	×	79.31	70.04	66.34	52.82	85.01	75.43
nnUNet	×	80.42	70.81	72.59	60.36	85.11	76.02
TransUNet	×	78.63	69.13	71.24	58.44	85.16	78.18
SwinUnet	×	78.07	68.34	63.29	50.19	85.79	74.75
UCTransNet	×	79.15	69.60	65.90	52.69	85.21	75.80
ConVIRT	✓	79.72	70.58	72.06	59.73	85.89	75.60
TGANet	✓	79.87	70.75	71.81	59.28	85.69	75.93
CLIP	✓	79.81	70.66	71.97	59.64	85.50	74.19
GLoRIA	✓	79.94	70.68	72.42	60.18	85.84	75.52
ViLT	✓	79.63	70.12	72.36	60.15	86.18	76.40
LAVT	✓	79.28	69.89	73.29	60.41	85.84	76.31
MGCA	✓	80.92	71.04	73.22	60.53	86.08	**76.86**
Ours	✓	**83.79**	**74.34**	**75.06**	**62.41**	**86.96**	76.29

Implementation Details

Our method is implemented within PyTorch on the Ubuntu 20.04.4 LTS with 24GB V100 GPU. BioClinicalBERT (Alsentzer et al. 2019) is used as the text encoder, and the encoding part of U-Net (Ronneberger, Fischer, and Brox 2015) is used as the vision encoder. 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 $(\omega_1, \omega_2, \omega_3)$ as $(0.2, 0.5, 0.5)$.

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 U-Net (Ronneberger, Fischer, and Brox 2015), UNet++ (Zhou et al. 2018), AttUNet (Oktay et al. 2018), nnUNet (Isensee et al. 2021), TransUNet (Chen et al. 2021), SwinUnet (Cao et al. 2022), UCTransNet (Wang et al. 2022c), ConVIRT (Zhang et al. 2022b), TGANet (Tomar et al. 2022), CLIP (Radford et al. 2021), GLoRIA (Huang et al. 2021), ViLT (Kim, Son, and Kim 2021), LViT (Li et al. 2023b), and MGCA (Wang et al. 2022b), as illustrated in Table 1. Experimental results on the QaTa-COV19 dataset show that our EviVLM achieves the best performance. In detail, EviVLM improves the Dice score by 3.37% and the mIoU score by 3.53% compared to the suboptimal nnUNet without text prompt. And it also has a 2.87% higher Dice score and a 3.30% better mIoU score than the suboptimal MGCA 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., LAVT, our EviVLM improves the Dice value by 1.77% and the mIoU value by 2.00%. Similarly, for Duke-Breast-Cancer-MRI dataset, it can be seen that EviVLM has a 0.78% higher Dice score than the second best method (ViLT).

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 the modality gap for effective modality fusion.

Ablation Studies

Extensive ablation experiments in Table 2 are performed on three public datasets to verify the contribution of each component. U-Net is adopted as the backbone.

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 1.51%, 1.94%, and 0.52% on the QaTa-COV19, MosMedData+, and Duke-Breast-Cancer-MRI datasets, respectively.

EAMG drives reliable vision and text evidence embedding learning. Table 2 shows that Method No. 4 outperforms Method No. 3 by 1.89% in Dice and 2.15% 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 reliable vision and text evidences.

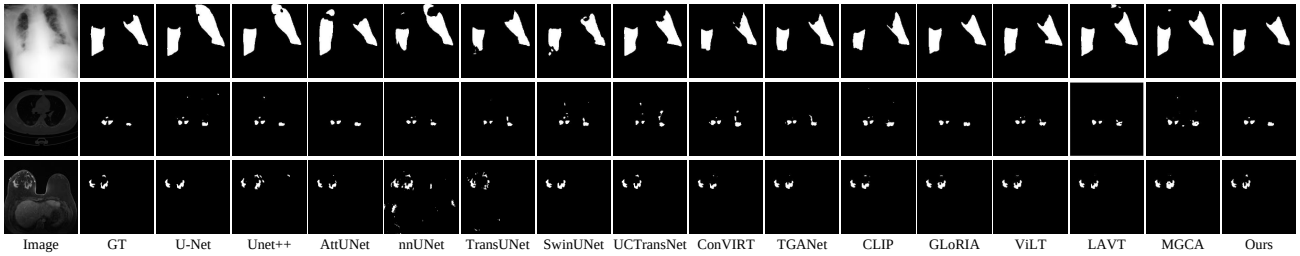

Figure 5: Our method shows superior segmentation performance on the QaTa-COV19 (first line), MosMedData+ (second line), and Duke-Breast-Cancer-MRI (third line) datasets, compared to various UNet variants and VLM methods.

Table 2: 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		BM-Seg		MoNuSeg	
	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
No. 2	✓	✓				79.68	70.55	70.89	59.45	85.22	74.06
No. 3	✓	✓	✓			81.19	71.59	72.83	60.57	85.74	75.36
No. 4	✓	✓	✓	✓		83.08	73.74	74.09	61.30	86.43	76.01
No. 5	✓	✓	✓		✓	82.75	72.69	73.87	61.14	86.11	75.92
No. 6	✓	✓	✓	✓	✓	**83.79**	**74.34**	**75.06**	**62.41**	**86.96**	**76.29**

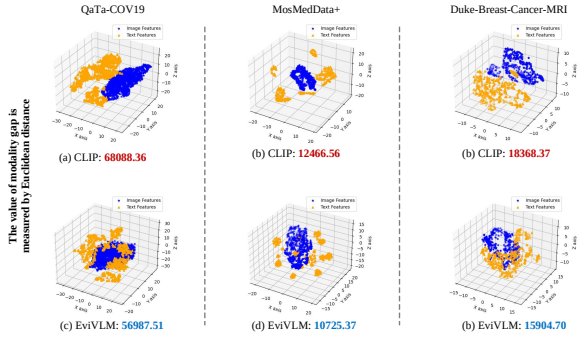

Figure 6: Our EviVLM reduces the modality gap by 11100.85, 1741.19, and 2463.67 on the QaTa-COV19, MosMedData+, and Duke-Breast-Cancer-MRI datasets, compared to CLIP.

EDSL boosts the alignment between image and text evidence embeddings. Method No. 5 in Table 2 demonstrates that adding EDSL to Method No. 3 improves the Dice by 1.56%, 1.04%, 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.

Modality Gap Visualization

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.

Conclusion

In this work, we propose a novel VLM paradigm, EviVLM, by pioneeringly introducing EL to VLM, which aims to bridge the modality gap problem between image and text for cross-modal fusion. To collect reliable cross-modal evidence, a EAMG is proposed to refine both modality-specific evidence embeddings through 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 the state-of-the-art performance of the proposed EviVLM.

References

Alsentzer, E.; Murphy, J. R.; Boag, W.; Weng, W.-H.; Jin, D.; Naumann, T.; and McDermott, M. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

Amini, A.; Schwarting, W.; Soleimany, A.; and Rus, D.

2020. Deep evidential regression. *Advances in neural information processing systems*, 33: 14927–14937.

Bao, L.; Wei, L.; Zhou, W.; Liu, L.; Xie, L.; Li, H.; and Tian, Q. 2023. Multi-Granularity Matching Transformer for Text-Based Person Search. *IEEE Transactions on Multimedia*.

Bao, W.; Yu, Q.; and Kong, Y. 2021. Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13349–13358.

Bao, W.; Yu, Q.; and Kong, Y. 2022. Opental: Towards open set temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2979–2989.

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, 205–218.

Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; and Zhou, Y. 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2019. Uniter: Learning universal image-text representations.(2019). *arXiv preprint arXiv:1909.11740*.

Conneau, A.; and Lample, G. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Degerli, A.; Kiranyaz, S.; Chowdhury, M. E.; and Gabbouj, M. 2022. Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In *2022 IEEE International Conference on Image Processing (ICIP)*, 2306–2310.

Dempster, A. P. 1968. A generalization of Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2): 205–232.

Domingos, P. 2000. A unified bias-variance decomposition. In *Proceedings of 17th international conference on machine learning*, 231–238.

Guo, J.; Ye, J.; Xiang, Y.; and Yu, Z. 2023. Layer-level progressive transformer with modality difference awareness for multi-modal neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Hofmanninger, J.; Prayer, F.; Pan, J.; Röhrich, S.; Prosch, H.; and Langs, G. 2020. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental*, 4: 1–13.

Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; and Wang, L. 2022. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17980–17989.

Huang, L.; Denoeux, T.; Vera, P.; and Ruan, S. 2022. Evidence fusion with contextual discounting for multi-modality medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 401–411.

Huang, S.-C.; Shen, L.; Lungren, M. P.; and Yeung, S. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3942–3951.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211.

Jøsang, A. 2016. *Subjective logic*, volume 3. Springer.

Jøsang, A.; and Hankin, R. 2012. Interpretation and fusion of hyper opinions in subjective logic. In *2012 15th International Conference on Information Fusion*, 1225–1232.

Jsang, A. 2018. *Subjective Logic: A formalism for reasoning under uncertainty*.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, 5583–5594.

Li, B.; Han, Z.; Li, H.; Fu, H.; and Zhang, C. 2022. Trustworthy long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6970–6979.

Li, H.; Nan, Y.; Del Ser, J.; and Yang, G. 2023a. Region-based evidential deep learning to quantify uncertainty and improve robustness of brain tumor segmentation. *Neural Computing and Applications*, 35(30): 22071–22085.

Li, Z.; Li, Y.; Li, Q.; Wang, P.; Guo, D.; Lu, L.; Jin, D.; Zhang, Y.; and Hong, Q. 2023b. Lvit: language meets vision transformer in medical image segmentation. *IEEE transactions on medical imaging*.

Liang, V. W.; Zhang, Y.; Kwon, Y.; Yeung, S.; and Zou, J. Y. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35: 17612–17625.

Liu, J.; Zhang, Y.; Chen, J.-N.; Xiao, J.; Lu, Y.; A Landman, B.; Yuan, Y.; Yuille, A.; Tang, Y.; and Zhou, Z. 2023. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21152–21164.

Malinin, A.; and Gales, M. 2018. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31.

Morozov, S. P.; Andreychenko, A. E.; Pavlov, N.; Vladzymyrskyy, A.; Ledikhova, N.; Gombolevskiy, V.; Blokhin, I. A.; Gelezhe, P.; Gonchar, A.; and Chernina, V. Y. 2020. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*.

Müller, P.; Kaissis, G.; Zou, C.; and Rueckert, D. 2022. Joint learning of localized representations from medical images and reports. In *European Conference on Computer Vision*, 685–701.

Oktay, O.; Schlemper, J.; Folgoc, L. L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N. Y.; Kainz, B.; et al. 2018. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241.

Saha, A.; Harowicz, M. R.; Grimm, L. J.; Kim, C. E.; Ghate, S. V.; Walsh, R.; and Mazurowski, M. A. 2018. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. *British journal of cancer*, 119(4): 508–516.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018a. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018b. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Tomar, N. K.; Jha, D.; Bagci, U.; and Ali, S. 2022. TGANet: Text-guided attention for improved polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 151–160.

Wang, F.; Zhou, Y.; Wang, S.; Vardhanabhuti, V.; and Yu, L. 2022a. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35: 33536–33549.

Wang, F.; Zhou, Y.; Wang, S.; Vardhanabhuti, V.; and Yu, L. 2022b. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35: 33536–33549.

Wang, H.; Cao, P.; Wang, J.; and Zaiane, O. R. 2022c. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 2441–2449.

Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022d. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*.

Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C. D.; and Langlotz, C. P. 2022a. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, 2–25.

Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C. D.; and Langlotz, C. P. 2022b. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, 2–25.

Zhou, Z.; Rahman Siddiquee, M. M.; Tajbakhsh, N.; and Liang, J. 2018. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 3–11.

Reproducibility Checklist

Unless specified otherwise, please answer "yes" to each question if the relevant information is described either in the paper itself or in a technical appendix with an explicit reference from the main paper. If you wish to explain an answer further, please do so in a section titled "Reproducibility Checklist" at the end of the technical appendix.

This paper:

1. Includes a conceptual outline and/or pseudocode description of AI methods introduced. **yes**
2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results. **yes**
3. Provides well marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper. **yes**

Does this paper make theoretical contributions? yes

If yes, please complete the list below.

1. All assumptions and restrictions are stated clearly and formally. **yes**
2. All novel claims are stated formally (e.g., in theorem statements). **yes**
3. Proofs of all novel claims are included. **yes**
4. Proof sketches or intuitions are given for complex and/or novel results. **yes**
5. Appropriate citations to theoretical tools used are given. **yes**
6. All theoretical claims are demonstrated empirically to hold. **yes**
7. All experimental code used to eliminate or disprove claims is included. **yes**

Does this paper rely on one or more datasets? more

If yes, please complete the list below.

1. A motivation is given for why the experiments are conducted on the selected datasets. **yes**
2. All novel datasets introduced in this paper are included in a data appendix. **yes**
3. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. **yes**
4. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. **yes**
5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. **yes**

6. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. **N/A**

Does this paper include computational experiments? yes

If yes, please complete the list below.

1. Any code required for pre-processing data is included in the appendix. **yes**
2. All source code required for conducting and analyzing the experiments is included in a code appendix. **yes**
3. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. **yes**
4. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from. **yes**
5. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. **yes**
6. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. **yes**
7. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. **yes**
8. This paper states the number of algorithm runs used to compute each reported result. **yes**
9. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. **yes**
10. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). **yes**
11. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. **yes**
12. This paper states the number and range of values tried per (hyper-)parameter during development of the paper, along with the criterion used for selecting the final parameter setting. **no**