Evi-Steer: Learning to Steer Biomedical Vision-Language Models through Efficient and Generalizable Evidential Tuning

Taha Koleilat✉, Hassan Rivaz, and Yiming Xiao

Concordia University, Montreal, Canada

Abstract. Parameter-efficient adaptation of vision–language foundation models is crucial for precise multimodal understanding of biomedical images, yet existing methods remain deterministic and often struggle under domain shift or ambiguous image–text alignment. This limitation is particularly critical in the clinic, where models should remain robust in low-data regimes and domain shifts. We present `Evi-Steer`, an *evidential cross-modal low-dimensional steering* framework for BiomedCLIP that enables uncertainty-aware parameter-efficient fine-tuning while updating only **0.11% of total model parameters**. Our approach performs lightweight low-dimensional token updates in both vision and text encoders while simultaneously estimating epistemic uncertainty. These uncertainty estimates update gate residuals, allowing the model to adapt conservatively when evidence is weak. Furthermore, we introduce cross-modal confidence fusion based on *Dempster–Shafer theory*, enabling visual adaptation to be conditioned on textual confidence and suppressing conflicting or uncertain cross-modal updates. We conduct a comprehensive evaluation on 15 biomedical imaging datasets spanning 8 organs and 8 imaging modalities under few-shot learning and domain generalization settings. `Evi-Steer` consistently outperforms state-of-the-art methods under few-shot learning and domain shift settings, demonstrating a practical and robust pathway for deploying vision-language models in real-world clinical settings. Code is available at https://github.com/HealthX-Lab/Evi-Steer.

Keywords: Vision–Language Models · Fine-Tuning · Few-shot Learning

1 Introduction

Vision–language foundation models (VLMs), such as OpenAI's CLIP [30] and domain-specific variants such as BiomedCLIP [43], provide a promising route to label-efficient learning by aligning images and natural language through large-scale contrastive pretraining. This paradigm is particularly attractive in biomedical settings [19,21], where expert-labeled data are scarce and expensive, images

✉ Corresponding Author: taha.koleilat@mail.concordia.ca

exhibit subtle inter-class differences and high intra-class variability across scanners and patient populations, and domain shifts often degrade out-of-distribution (OOD) generalization. In addition to that, practical clinical deployment demands resource- and parameter-efficient adaptation for real-time use. However, effective biomedical adaptation requires more than independent tuning of vision and text encoders. Clinical semantics are inherently cross-modal, and robust transfer demands coordinated co-adaptation of visual and textual representations. Simply updating one modality or injecting static prompts may fail to resolve cross-modal inconsistencies [35], especially when prompts are ambiguous or visual features are subtle.

Full fine-tuning of large VLMs is computationally expensive and can degrade pretrained knowledge [34], especially in few-shot settings. This has motivated parameter-efficient adaptation. Low-rank adaptation (LoRA) [11,41] and related methods [36,22] operate in the model weight space by learning low-rank updates to internal projection matrices. Prompt-based approaches [45,44,17] modify input representations by introducing trainable tokens, while adapter-based methods [42,9] insert lightweight bottleneck modules that transform intermediate activations. Although these approaches reduce the number of trainable parameters, they remain deterministic, applying residual updates without accounting for uncertainty in the adaptation process. So far, most uncertainty-aware deep learning methods rely on sampling-based Bayesian inference [8,15], which requires multiple forward passes and incurs substantial overhead. Evidential inference instead estimates evidence and uncertainty in a single pass [2,31], making it attractive for efficient VLM adaptation. Recent works have explored related directions: CILMP applies ReFT-style low-rank interventions for disease-specific prompting [6], while EviVLM and Bayesian-PEFT incorporate evidential uncertainty into segmentation and PEFT, respectively [25,26]. However, they do not jointly perform cross-modal evidential steering in activation space to enable efficient adaptation while preserving generalizable pretrained representations.

Motivated by this gap, we present `Evi-Steer`, an *evidential cross-modal low-dimensional steering* module for parameter-efficient tuning of BiomedCLIP. In contrast to encoder weight-space adaptation methods, we follow a *representation steering* paradigm [39], which performs adaptation directly in the activation space while keeping the backbone parameters frozen. Specifically, we inject lightweight r-dimensional cross-modal modules into intermediate layers of both the vision and text encoders to modulate token representations rather than altering pretrained weights. To achieve uncertainty-aware adaptation, we estimate intermediate confidence through evidential modeling and use these estimates to gate residual updates instead of applying uniform deterministic injections. We also fuse vision and text evidence via *Dempster–Shafer belief combination* [32] to suppress adaptation from conflicting or uncertain text-image samples. By jointly co-adapting both modalities under uncertainty gating, `Evi-Steer` preserves the pretrained knowledge of the foundation model while improving few-shot accuracy and robustness under domain shift. Our contributions are three-fold: **First**, we introduce an **evidential representation steering** mechanism that estimates

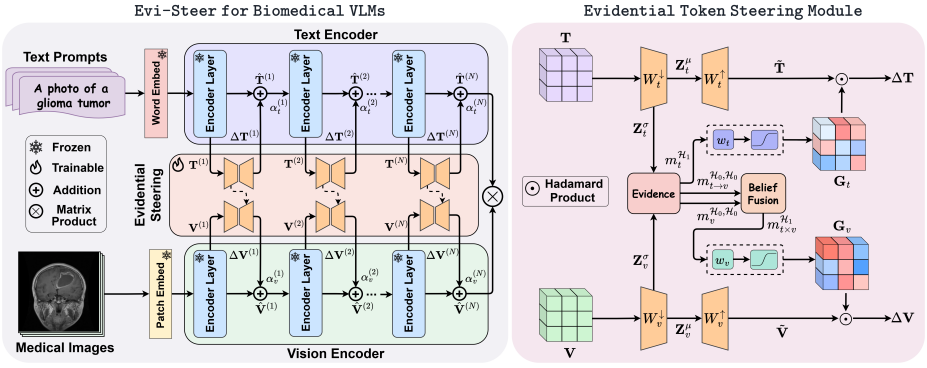

Fig. 1: Overview of our framework where evidential cross-modal low-dimensional adapters produce confidence-weighted representation updates.

latent dimension (LD) epistemic uncertainty and applies confidence-weighted updates directly in the activation space rather than the weight space. **Second**, we propose **cross-modal belief fusion** that conditions visual adaptation on textual confidence, improving robustness when prompts or features are ambiguous. **Third**, we provide **comprehensive evaluation** on 15 biomedical datasets under *few-shot learning* and *domain generalization*, demonstrating consistent gains over SOTA baselines.

2 Methods and Materials

2.1 Model architecture of `Evi-Steer`

Figure 1 illustrates the overall pipeline of `Evi-Steer`. Given an input image and a set of class prompts, the vision and text encoders first extract pretrained features. We then inject cross-modal low-dimensional representation adapters into the first d layers of both encoders. At each adapted layer, the module (i) computes a low-dimensional residual update in the activation space, (ii) estimates *latent dimension* (LD) epistemic uncertainty through evidential modeling, and (iii) reweights the representation token steering based on the estimated confidence at the layer. Textual uncertainties are aggregated and converted into belief masses that are fused with visual evidence via *Dempster–Shafer combination*. The fused belief modulates visual adaptation, ensuring that updates are emphasized only when both modalities provide reliable evidence. The final adapted representations are then used for classification.

Token representations. At encoder layer $n \in \{1, \ldots, N\}$, let the vision tokens be $\mathbf{V}^{(n)} \in \mathbb{R}^{B \times P_v \times D_v}$ and the text tokens be $\mathbf{T}^{(n)} \in \mathbb{R}^{C \times P_t \times D_t}$, where B is the batch size, $\{P_v, P_t\}$ are the numbers of vision and text tokens respectively, C is the number of classes, and $\{D_v, D_t\}$ are the channel dimensions of each modality.

We inject cross-modal low-dimensional adapters at each of the first d layers of both encoders, while keeping the backbone parameters frozen.

Latent Dimension Representation. Unlike LoRA [11], which approximates weight updates through low-rank decompositions of model parameters, our approach follows ReFT [39], which directly decomposes activations to model low-dimensional *steering directions* in token representations. At Layer n, text and vision representations are projected into a shared r-dimensional latent space using a single down-projection per modality. The resulting latent representation is split into mean-update and uncertainty channels:

$$[\mathbf{Z}_t^{\mu},\, \mathbf{Z}_t^{\sigma}] = \mathbf{T}^{(n)} W_t^{\downarrow}, \quad W_t^{\downarrow} \in \mathbb{R}^{D_t \times 2r} \tag{1}$$

$$[\mathbf{Z}_v^{\mu},\, \mathbf{Z}_v^{\sigma}] = \mathbf{V}^{(n)} W_v^{\downarrow}, \quad W_v^{\downarrow} \in \mathbb{R}^{D_v \times 2r}. \tag{2}$$

where $\mathbf{Z}_t^{\mu}, \mathbf{Z}_t^{\sigma} \in \mathbb{R}^{C \times P_t \times r}$ and $\mathbf{Z}_v^{\mu}, \mathbf{Z}_v^{\sigma} \in \mathbb{R}^{B \times P_v \times r}$ denote the LD mean update and standard deviation representations for text and vision, respectively. The low-dimensional token updates are then obtained via their up-projections:

$$\tilde{\mathbf{T}}^{(n)} = \mathbf{Z}_t^{\mu} W_t^{\uparrow}, \qquad W_t^{\uparrow} \in \mathbb{R}^{r \times D_t}, \tag{3}$$

$$\tilde{\mathbf{V}}^{(n)} = \mathbf{Z}_v^{\mu} W_v^{\uparrow}, \qquad W_v^{\uparrow} \in \mathbb{R}^{r \times D_v}. \tag{4}$$

As in standard adapter design, the up-projection matrices $W_t^{\uparrow}$ (text) and $W_v^{\uparrow}$ (vision) are initialized to zero, ensuring stable warm-start training and preserving the pretrained backbone at initialization.

Evidential uncertainty in latent space. We model LD epistemic uncertainty using an evidential construction. For each token, the uncertainty channel $\mathbf{Z}^{\sigma}$ is mapped to evidence $\mathbf{e}$, Dirichlet concentration parameters $\boldsymbol{\beta}$, and the corresponding LD epistemic uncertainty $\mathbf{u}$ as

$$\mathbf{e} = \texttt{softplus}\big((\mathbf{Z}^{\sigma})^{\circ 2}\big) + \epsilon, \qquad \boldsymbol{\beta} = \mathbf{e} + 1, \qquad \mathbf{u} = \boldsymbol{\beta}^{-1}. \tag{5}$$

Here, $(\cdot)^{\circ 2}$ denotes element-wise square and ϵ is a small constant. Since latent features are not simplex-normalized class probabilities, we use $\boldsymbol{\beta}^{-1}$ as an element-wise inverse-evidence proxy for latent uncertainty in comparison to [31].

Confidence modeling. Each modality produces LD uncertainty estimates that quantify the confidence of applying low-dimensional updates. We interpret uncertainty through two complementary hypotheses: *apply update*, denoted by $\mathcal{H}_1$, and *ignorance*, denoted by $\mathcal{H}_0$. For a given uncertainty tensor $\mathbf{u}_v/\mathbf{u}_t$, the corresponding belief masses for vision and text are defined as

$$m_v^{\mathcal{H}_1} = 1 - \mathbf{u}_v, \qquad m_v^{\mathcal{H}_0} = \mathbf{u}_v, \tag{6}$$

$$m_t^{\mathcal{H}_1} = 1 - \mathbf{u}_t, \qquad m_t^{\mathcal{H}_0} = \mathbf{u}_t, \tag{7}$$

such that confident tokens contribute more strongly to adaptation, while uncertain components are conservatively suppressed.

Cross-modal belief fusion. *Dempster–Shafer theory* (DST) provides a principled framework for combining evidence from multiple sources while explicitly modeling uncertainty as belief and ignorance. Unlike standard probabilistic fusion, DST separates confidence from uncertainty and employs a combination rule that naturally downweights conflicting evidence. This property makes it particularly suitable for cross-modal scenarios, where visual and textual signals may be complementary, unreliable, or ambiguous. Here, we use DST to condition visual adaptation on textual confidence. Specifically, we first aggregate textual uncertainty across class prompts using the `[EOS]` token to obtain an LD-based uncertainty summary. This pooled uncertainty is then converted into belief masses that modulate visual updates to encourage vision encoder adaptation when textual evidence is confident and suppress it otherwise.

$$\bar{\mathbf{u}}_t = \frac{1}{C}\sum_{c=1}^{C} \mathbf{u}_{t,c}^{\texttt{[EOS]}} \in \mathbb{R}^{1\times 1\times r}, \qquad m_{t\to v}^{\mathcal{H}_1} = 1 - \bar{\mathbf{u}}_t, \qquad m_{t\to v}^{\mathcal{H}_0} = \bar{\mathbf{u}}_t. \tag{8}$$

The two belief sources are fused using the *Dempster–Shafer combination*:

$$m_{t\times v}^{\mathcal{H}_1} = \frac{m_{t\to v}^{\mathcal{H}_1} m_v^{\mathcal{H}_1}}{1 - (m_{t\to v}^{\mathcal{H}_1} m_v^{\mathcal{H}_0} + m_{t\to v}^{\mathcal{H}_0} m_v^{\mathcal{H}_1}) + \epsilon}. \tag{9}$$

The fused belief $m_{t\times v}^{\mathcal{H}_1} \in \mathbb{R}^{B\times P_v\times r}$ captures *consensus confidence* between modalities, emphasizing updates only when both text and vision provide evidence, and suppressing adaptation under ambiguity or conflict. ϵ is a small constant.

Text and vision confidence. Linear layers $\{w_v, w_t\}$ aggregate LD belief masses along the LD axis to produce token-wise *confidence weights* $\mathbf{G}$ and weighted updates $(\Delta\mathbf{T}, \Delta\mathbf{V})$. For text prompts, unimodal belief is used directly:

$$\Delta\mathbf{T}^{(n)} = \mathbf{G}_t^{(n)} \odot \tilde{\mathbf{T}}^{(n)}, \quad \mathbf{G}_t^{(n)} = \texttt{sigmoid}\Big(m_t^{\mathcal{H}_1}\mathbf{w}_t\Big), \quad \mathbf{w}_t \in \mathbb{R}^{r\times 1} \tag{10}$$

while for vision tokens, confidence is computed from the fused cross-modal belief:

$$\Delta\mathbf{V}^{(n)} = \mathbf{G}_v^{(n)} \odot \tilde{\mathbf{V}}^{(n)}, \quad \mathbf{G}_v^{(n)} = \texttt{sigmoid}\Big(m_{t\times v}^{\mathcal{H}_1}\mathbf{w}_v\Big), \quad \mathbf{w}_v \in \mathbb{R}^{r\times 1} \tag{11}$$

This formulation ensures consistent treatment of uncertainty across modalities, while allowing vision updates to be modulated by joint text–image confidence.

Evidential token steering. The confidence-weighted updates are applied with learnable scalings $\{\alpha_v^{(n)}, \alpha_t^{(n)}\} \in [0, 1]$ to control overall steering magnitude:

$$\hat{\mathbf{V}}^{(n)} = \mathbf{V}^{(n)} + \alpha_v^{(n)} * \Delta\mathbf{V}^{(n)}, \tag{12}$$

$$\hat{\mathbf{T}}^{(n)} = \mathbf{T}^{(n)} + \alpha_t^{(n)} * \Delta\mathbf{T}^{(n)}. \tag{13}$$

CLIP uses the `[EOS]` token as the `[CLS]` token

Evidential regularization. To prevent degenerate uncertainty estimates and enforce statistically meaningful evidence, we regularize the concentration parameters $\boldsymbol{\beta}$ produced by the latent uncertainty channel by penalizing their deviation from a neutral prior Gamma$(1, 1)$ via a Kullback–Leibler divergence:

$$\mathcal{L}_{\mathrm{KL}} = KL(\mathrm{Gamma}(\boldsymbol{\beta}, 1) \,\|\, \mathrm{Gamma}(1, 1)) = (\boldsymbol{\beta} - 1)\psi(\boldsymbol{\beta}) - \log \Gamma(\boldsymbol{\beta}), \quad (14)$$

where $\psi(\cdot)$ denotes the digamma function and $\Gamma(\cdot)$ the Gamma function. The final objective combines the classification loss with this regularization:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}}, \quad \lambda_{\mathrm{KL}} = 10^{-4} \quad (15)$$

2.2 Experimental Setup

We evaluate the proposed method on a broad collection of biomedical imaging benchmarks under protocols designed to measure both data efficiency and robustness across few-shot classification scenarios.

Few-Shot Learning. To examine performance in low-data regimes, we conduct experiments with $K = \{4, 8, 16\}$ labeled samples per class. This setting reflects realistic clinical scenarios where expert annotations are scarce.

Domain Generalization. To study robustness to distribution shift, models are trained using the 16-shot setting on a source dataset and evaluated on both the source and several unseen target datasets without any additional fine-tuning. This protocol measures the ability to transfer knowledge across domains.

Datasets. Experiments are performed on 15 medical imaging datasets spanning 8 organs and 8 imaging modalities. For *few-shot learning*, we use computerized tomography (CTKidney [13]), endoscopy (Kvasir [27]), fundus photography (RETINA [28,23]), histopathology (LC25000 [4], CHMNIST [14]), magnetic resonance imaging (BTMRI [24]), optical coherence tomography (OCTMNIST [16]), ultrasound (BUSI [1]), and X-ray (COVID-QU-Ex [37]). For *domain generalization*, we test on the breast ultrasound domain (BUID [3], BUSBRA [10], UDIAT [5]) and the brain MRI domain (BTMRI-P [29], BTMRI-S [33], BRISC [7]). This diverse suite includes challenging modalities, enabling a thorough evaluation across heterogeneous biomedical settings.

Implementation Details. We adopt BiomedCLIP with a ViT-B/16 backbone and report averages over three independent runs. For all experiments, training is performed for 100 epochs using the prompt template "`a photo of a [class]`", along with an Adam optimizer [18] with a learning rate of 7.5×10^{-4} and a batch size of 16. All experiments are conducted on one NVIDIA A100 GPU (40GB).

3 Results

3.1 Few-shot Adaptation & Domain Generalization Evaluation

Few-shot adaptation. As shown in Table 1, `Evi-Steer` consistently achieves the best overall performance across all K-shot regimes, improving the average

Table 1: Accuracy (%) benchmarking for the 9 datasets with BiomedCLIP as the backbone for all methods. **bold**=best results & underlined=second best results.

Shots	Method	BTMRI	BUSI	COVID	CTKIDNEY	Kvasir	CHMNIST	LC25000	RETINA	OCTMNIST	Average
0	**BiomedCLIP** [43]	56.79	59.75	43.80	42.43	54.58	30.65	50.03	26.26	30.00	43.81
4	CoOp [45]	74.68	60.17	67.03	68.12	70.78	68.66	84.66	42.22	53.37	65.52
	CoCoOp [44]	67.83	59.75	63.70	61.07	68.94	58.58	77.44	39.75	48.57	60.63
	KgCoOp [40]	75.40	62.01	65.91	68.68	68.28	68.77	82.10	42.61	52.97	65.19
	ProGrad [46]	76.24	62.29	68.56	67.90	70.00	69.13	84.72	43.09	55.07	66.33
	BiomedCoOp [20]	77.23	59.32	**73.28**	66.50	74.08	71.19	85.60	45.58	54.73	67.50
	LP++ [12]	75.48	60.31	62.32	65.73	69.36	67.79	82.61	46.95	59.02	65.51
	CLIP-Adapter [9]	56.80	61.72	46.28	42.19	54.83	33.26	52.91	26.07	49.96	47.11
	TIP-Adapter-F [42]	77.90	64.54	69.57	60.18	69.94	71.74	79.57	47.37	55.20	66.22
	GDA [38]	77.56	64.41	66.77	63.41	72.19	70.92	85.00	48.42	57.42	67.34
	CLIP-LoRA [41]	77.31	64.91	69.85	66.47	66.11	64.34	84.93	47.38	52.10	65.93
	`Evi-Steer` (Ours)	**78.28**	**65.54**	69.93	**71.98**	**74.22**	**75.29**	**86.12**	**49.19**	**72.33**	**71.43**
8	CoOp [45]	79.27	64.69	74.66	77.40	77.14	75.00	87.50	51.87	63.67	72.36
	CoCoOp [44]	71.69	65.82	69.36	73.93	72.92	66.58	85.57	48.45	55.40	67.75
	KgCoOp [40]	79.79	67.37	74.86	77.43	72.05	69.50	84.63	49.97	61.03	70.74
	ProGrad [46]	78.82	64.83	74.65	78.23	76.03	70.99	87.86	52.26	62.17	71.76
	BiomedCoOp [20]	78.55	63.27	**76.26**	77.16	77.72	74.78	88.77	56.47	58.87	72.43
	LP++ [12]	77.11	66.10	66.19	77.06	72.52	72.40	89.14	53.44	63.69	70.85
	CLIP-Adapter [9]	57.15	61.86	48.68	44.64	56.08	36.48	56.33	25.84	49.50	48.51
	TIP-Adapter-F [42]	79.18	68.50	69.89	75.24	75.86	74.51	90.31	56.07	65.00	72.73
	GDA [38]	81.93	70.20	74.96	**82.33**	74.47	82.03	89.17	52.58	66.57	74.92
	CLIP-LoRA [41]	78.43	69.13	75.81	77.95	70.72	82.61	86.21	50.08	61.33	72.47
	`Evi-Steer` (Ours)	**83.42**	**72.74**	74.69	77.71	**79.64**	**82.78**	**90.39**	**62.22**	**72.40**	**77.33**
16	CoOp [45]	82.37	69.49	76.37	83.52	77.88	79.63	92.19	59.38	65.47	76.26
	CoCoOp [44]	78.45	70.20	74.52	77.70	75.22	72.16	87.38	53.91	60.67	72.25
	KgCoOp [40]	81.07	70.62	75.65	77.67	72.95	73.58	86.79	51.18	62.80	72.48
	ProGrad [46]	82.84	71.47	74.93	81.13	75.88	75.11	90.70	50.47	63.33	73.98
	BiomedCoOp [20]	83.30	70.34	**78.72**	83.20	78.89	79.05	92.68	61.28	66.93	77.15
	LP++ [12]	81.61	70.05	72.79	79.07	75.41	78.32	92.58	60.62	68.35	75.42
	CLIP-Adapter [9]	60.16	63.55	49.55	47.28	56.50	42.06	57.56	26.05	52.73	50.60
	TIP-Adapter-F [42]	82.27	71.89	76.07	82.07	78.00	80.43	92.35	62.85	72.50	77.60
	GDA [38]	81.91	66.81	75.65	81.54	79.53	83.66	**93.55**	57.02	75.43	77.23
	CLIP-LoRA [41]	80.19	71.42	76.71	81.85	76.81	84.14	87.98	52.19	61.50	74.75
	`Evi-Steer` (Ours)	**86.39**	**73.16**	73.44	**90.11**	**82.33**	**84.22**	92.05	**68.53**	**80.37**	**81.18**

accuracy from 43.81% (zero-shot) to 71.45%, 77.34%, and 81.21% at $K = 4/8/16$ shots, with strong gains on challenging datasets (e.g. BTMRI, OCTMNIST), updating only 0.11% of the total parameters (221K parameters), on par with CLIP-LoRA in parameter budget, while consistently surpassing it in performance.

Domain generalization. Table 2 shows that `Evi-Steer` generalizes better under domain shift, yielding the strongest OOD transfer in both breast Ultrasound and brain MRI domains (e.g., BUSI→BUID: 78.70%, BTMRI→BTMRI-P: 80.23%). Note the higher in-distribution (ID) performance compared to OOD in the breast ultrasound domain is expected, as the OOD dataset lacks the *normal scan* class, whereas BUSI contains three classes (malignant/benign/normal).

3.2 Ablation Experiments

Effectiveness of components. Table 3 shows that all components contribute to domain generalization, with visual adaptation being most critical. Removing it causes the largest degradation (−5.72% OOD; −4.68% harmonic mean (HM)), highlighting the need to adapt visual representations. Textual adaptation also

Table 2: Domain generalization accuracy (%): models are trained on a source dataset with 16-shots (except BiomedCLIP) and then directly evaluated on OOD target datasets without adaptation.

Method	Breast Ultrasound				Brain MRI			
	Source	Target			Source	Target		
	BUSI	BUID	BUSBRA	UDIAT	BTMRI	BTMRI-P	BTMRI-S	BRISC
BiomedCLIP [43]	59.75	75.00	66.78	61.54	56.79	52.80	55.20	52.60
CoOp [45]	69.49	72.22	68.28	70.49	82.37	76.77	78.13	74.87
CoCoOp [44]	70.20	67.59	66.55	71.54	78.45	75.43	76.62	70.20
ProGrad [46]	71.47	69.44	67.25	69.23	82.63	78.00	79.96	76.27
KgCoOp [40]	70.62	77.78	68.79	75.64	81.07	75.60	79.02	73.37
GDA [38]	66.81	63.89	63.02	67.95	81.91	77.47	76.75	75.23
CLIP-LoRA [41]	71.42	65.85	65.37	69.82	80.19	77.90	77.68	73.50
BiomedCoOp [20]	70.34	67.59	62.90	71.67	83.30	77.60	79.52	74.50
`Evi-Steer` (Ours)	**73.16**	**78.70**	**71.61**	**78.21**	**86.39**	**80.23**	**80.53**	**78.43**

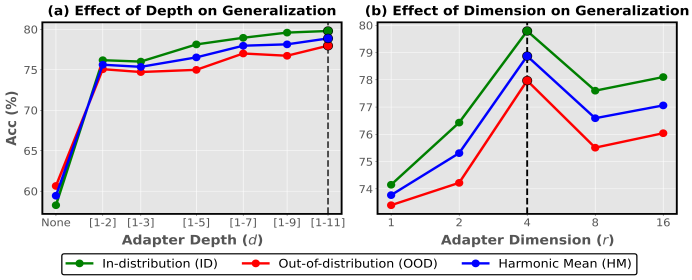

Fig. 2: Effect of layer interventions and adapter dimension on accuracy.

Table 3: Effect of different key components in `Evi-Steer` (HM=harmonic mean).

Method	Domain Generalization		
	ID Acc. (%)	OOD Acc. (%)	HM Acc. (%)
`Evi-Steer` (Ours)	**79.79**	**77.97**	**78.87**
w/o Visual Adaptation	$76.23_{(-3.56)}\downarrow$	$72.25_{(-5.72)}\downarrow$	$74.19_{(-4.68)}\downarrow$
w/o Textual Adaptation	$78.40_{(-1.39)}\downarrow$	$76.97_{(-1.00)}\downarrow$	$77.68_{(-1.19)}\downarrow$
w/o Evidential Update	$79.25_{(-0.54)}\downarrow$	$76.34_{(-1.63)}\downarrow$	$77.77_{(-1.10)}\downarrow$
w/o Cross-modal Belief	$79.52_{(-0.27)}\downarrow$	$76.90_{(-1.07)}\downarrow$	$78.19_{(-0.68)}\downarrow$

consistently contributes, yielding −1.00% OOD and −1.19% HM drops. The evidential update and cross-modal belief modules mainly affect OOD performance, suggesting uncertainty-aware learning improves robustness under domain shift.

Effect of adapter depth. Figure 2a shows that increasing the number of adapted layers d consistently improves generalization. Performance rises from no adaptation (HM 59.44%) to shallow adaptation and continues improving with depth, reaching the best performance at depth 11 (HM 78.87%). This monotonic trend indicates that distributing lightweight adapters across the network is more effective than shallow adaptation, enabling progressive cross-modal refinement.

Effect of adapter dimension. We evaluate the adapter dimension $r \in \{1, 2, 4, 8, 16\}$, as shown in Fig. 2b. Performance improves from $r = 1$ to $r = 4$, reaching a peak at $r = 4$ with the best HM accuracy of 78.87%. Larger dimensions provide no further gains and slightly degrade performance, suggesting that moderate-dimensional adapters best balance capacity and overfitting, and that `Evi-Steer` benefits from dimension-constrained updates.

4 Conclusion

We introduced `Evi-Steer`, a novel parameter-efficient adaptation framework for biomedical VLMs, by combining low-dimensional cross-modal updates with evidential uncertainty modeling and confidence-based gating. With strong few-shot and domain-generalization results, `Evi-Steer` demonstrates the value of confidence-aware biomedical vision-language adaptation in clinical settings. Future work will extend the approach to dense prediction and interactive workflows.

Acknowledgments. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada and the Fonds de recherche du Québec – Nature et technologies (B2X-363874).

Disclosure of Interests. The authors have no competing interests.

References

1. Al-Dhabyani, W., et al.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020) 6
2. Amini, A., et al.: Deep evidential regression. Advances in neural information processing systems **33**, 14927–14937 (2020) 2
3. Ardakani, A.A., et al.: An open-access breast lesion ultrasound image database: Applicable in artificial intelligence studies. Computers in Biology and Medicine **152**, 106438 (2023) 6
4. Borkowski, A.A., et al.: Lung and colon cancer histopathological dataset (2019) 6
5. Byra, M., et al.: Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network. Biomed Signal Process Control **61** (2020) 6
6. Du, Y., et al.: Medical knowledge intervention prompt tuning for medical image classification. IEEE Transactions on Medical Imaging (2025) 2
7. Fateh, A., et al.: Brisc: Annotated dataset for brain tumor segmentation and classification. Scientific Data **13**(1), 361 (2026) 6
8. Gal, Y., et al.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016) 2
9. Gao, P., et al.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**(2), 581–595 (2024) 2, 7
10. Gómez-Flores, W., et al.: Bus-bra: a breast ultrasound dataset for assessing computer-aided diagnosis systems. Medical Physics **51**(4), 3110–3123 (2024) 6
11. Hu, E.J., et al.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022) 2, 4

12. Huang, Y., et al.: Lp++: A surprisingly strong linear probe for few-shot clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23773–23782 (2024) 7
13. Islam, M.N., et al.: Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from ct-radiography. Scientific Reports **12**(1), 1–14 (2022) 6
14. Kather, J.N., et al.: Multi-class texture analysis in colorectal cancer histology. Scientific reports **6**(27988) (2016) 6
15. Kendall, A., et al.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems **30** (2017) 2
16. Kermany, D.S., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172**(5), 1122 – 1131.e9 (2018) 6
17. Khattak, M.U., et al.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 2
18. Kingma, D.P.: Adam: A method for stochastic optimization (2014) 6
19. Koleilat, T., et al.: Medclip-sam: Bridging text and image towards universal medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 643–653. Springer (2024) 1
20. Koleilat, T., et al.: Biomedcoop: Learning to prompt for biomedical vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14766–14776 (2025) 7, 8
21. Koleilat, T., et al.: Medclip-samv2: Towards universal text-driven medical image segmentation. Medical Image Analysis p. 103749 (2025) 1
22. Koleilat, T., et al.: Clip-svd: Efficient and interpretable vision–language adaptation via singular values. Transactions on Machine Learning Research (2026) 2
23. Köhler, T., et al.: Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. In: Proceedings of CBMS. pp. 95–100 (2013) 6
24. Nickparvar, M.: Brain MRI Data. https://www.kaggle.com/dsv/2645886 (2021) 6
25. Pan, Q., et al.: Evivlm: When evidential learning meets vision language model for medical image segmentation. IEEE Transactions on Medical Imaging (2025) 2
26. Pandey, D.S., et al.: Be confident in what you know: Bayesian parameter efficient fine-tuning of vision foundation models. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024) 2
27. Pogorelov, K., et al.: Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of the 8th ACM on Multimedia Systems Conference. pp. 164–169. ACM (2017) 6
28. Porwal, P., et al.: Indian diabetic retinopathy image dataset (idrid) (2018) 6
29. Pradeep, A.: Brain mri. https://www.kaggle.com/datasets/pradeep2665/brain-mri (2022) 6
30. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR (2021) 1
31. Sensoy, M., et al.: Evidential deep learning to quantify classification uncertainty. Advances in neural information processing systems (2018) 2, 4
32. Shafer, G.: Dempster-shafer theory. Encyclopedia of artificial intelligence (1992) 2
33. Sherif, M.M.: Brain Tumor Dataset. https://www.kaggle.com/datasets/mohamedmetwalysherif/braintumordataset (2021) 6
34. Shuttleworth, R., et al.: Lora vs full fine-tuning: An illusion of equivalence. Advances in Neural Information Processing Systems **38**, 174627–174662 (2026) 2
35. Spiegler, P., et al.: Textsam-eus: Text prompt learning for sam to accurately segment pancreatic tumor in endoscopic ultrasound. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 948–957 (2025) 2

36. Sun, Y., et al.: Singular value fine-tuning: Few-shot segmentation requires few-parameters fine-tuning. Neural Information Processing Systems (2022) 2
37. Tahir, A.M., et al.: Covid-19 infection localization and severity grading from chest x-ray images. Computers in Biology and Medicine **139**, 105002 (2021) 6
38. Wang, Z., et al.: A hard-to-beat baseline for training-free CLIP-based adaptation. In: The Twelfth International Conference on Learning Representations (2024) 7, 8
39. Wu, Z., et al.: Reft: Representation finetuning for language models. Advances in Neural Information Processing Systems **37**, 63908–63962 (2024) 2, 4
40. Yao, H., et al.: Visual-language prompt tuning with knowledge-guided context optimization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6757–6767 (2023) 7, 8
41. Zanella, M., et al.: Low-Rank Few-Shot Adaptation of Vision-Language Models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 1593–1603 (2024) 2, 7, 8
42. Zhang, R., et al.: Tip-adapter: Training-free adaption of clip for few-shot classification. In: European Conference on Computer Vision. Springer (2022) 2, 7
43. Zhang, S., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs (2025) 1, 7, 8
44. Zhou, K., et al.: Conditional prompt learning for vision-language models. In: CVPR. pp. 16816–16825 (2022) 2, 7, 8
45. Zhou, K., et al.: Learning to prompt for vision-language models. IJCV **130**(9), 2337–2348 (2022) 2, 7, 8
46. Zhu, B., et al.: Prompt-aligned gradient for prompt tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023) 7, 8