When Vision-Language Models Meet Fetal Cardiac Ultrasound: Dual-Level Contrastive Learning for Out-of-Distribution Detection

Ziyi Liu[1], **Weihu Song**[1], **Yupeng Ma**[1], **Tianrui Liu**[2], **Yifan Gu**[3] and **Haogang Zhu**[1,2*]
[1]School of Computer Science and Engineering, Beihang University, China
[2]Hangzhou International Innovation Institute, Beihang University, China
[3]School of Instrumentation and Optoelectronic Engineering, Beihang University, China
liu_zi_yi1999@163.com, {weihusong, liu_tianrui, haogangzhu}@buaa.edu.cn, {yupengma0301, guyifan1227}@gmail.com

Abstract

Recent advances in vision-language models (VLMs) have shown remarkable performance in medical image classification tasks. However, applying VLMs to fetal cardiac ultrasound (FCU) remains challenging due to compound distribution shifts, including covariate shifts caused by cross-center heterogeneity and semantic shifts arising from clinically non-standard views. To address this issue, we propose Dual-Level Contrastive Learning (DLCL), the first prompt-based VLM framework for out-of-distribution (OOD) detection in FCU, to the best of our knowledge. DLCL explicitly shapes the vision-language representation space through complementary local-level and global-level contrastive objectives. Specifically, local contrastive learning aligns instance-level features to mitigate covariate shifts, whereas global contrastive learning regularizes global prototypes to address semantic shifts. We conduct extensive experiments on a private multi-center FCU dataset and the public ISIC-OOD dataset to validate the proposed approach. On the challenging FCU task, DLCL achieves an AUROC of 89.61% and a harmonic mean of 80.35%, significantly outperforming state-of-the-art methods.

1 Introduction

Recent advances in vision-language models (VLMs) [Radford *et al.*, 2021] have demonstrated substantial improvements in image recognition performance. This success has driven increasing research interest in medical VLMs, ranging from general-purpose architectures [Zhang *et al.*, 2023; Khattak *et al.*, 2024] to domain-specialized expert models [Silva-Rodriguez *et al.*, 2025; Christensen *et al.*, 2024]. However, the violation of the closed-world assumption [Fei and Liu, 2016] limits the deployment of these models in healthcare, resulting in a reduced ability to reliably handle out-of-distribution (OOD) samples. This challenge is particularly acute in the context of fetal cardiac ultrasound (FCU),

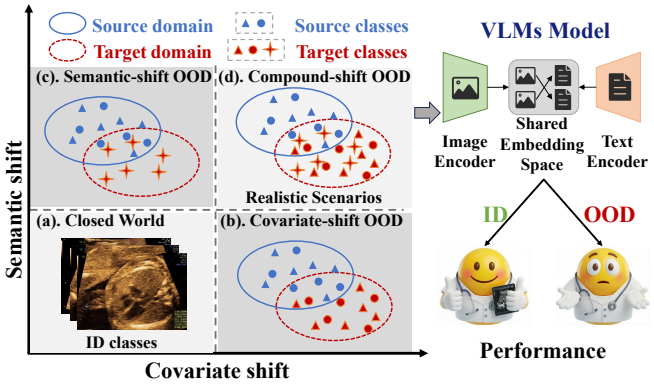


Figure 1: Overview of VLM-based OOD detection in FCU. We categorize scenarios by covariate and semantic shifts. Unlike the closed-world setting (a), realistic scenarios (d) exhibit compound shifts. VLMs align visual and textual features in a shared space and aim to distinguish ID from OOD samples.

where high operator dependency and inherent acoustic artifacts (e.g., speckle noise) [Noble and Boukerroui, 2006] lead to substantial data heterogeneity compared to standardized modalities such as MRI. As a consequence, models are prone to overconfident predictions on unseen inputs, leading to serious risks to patient safety [Hong *et al.*, 2024].

Although OOD detection has achieved remarkable progress, existing methods [Liang *et al.*, 2018; Hendrycks and Gimpel, 2017] primarily rely on post-hoc strategies and typically employ conventional vision encoder-only models, thereby leaving the rich semantic information in textual descriptions unexploited. To leverage the inherent strengths of VLMs and address the relevant challenges, several VLM-based methods [Miyai *et al.*, 2023; Ming *et al.*, 2022] have been proposed for OOD detection. However, previous studies primarily focus on standardized medical imaging modalities or natural images, leaving highly operator-dependent and heterogeneous settings underexplored, such as FCU.

In addition to these challenges, previous OOD detection benchmarks primarily focus on semantic shifts, characterized by disjoint label spaces between in-distribution (ID) and OOD samples. For example, a fetal heart view classification model may encounter novel classes, such as head view im-

*Corresponding author.

ages. However, these studies overlook the fact that covariate shifts can lead to significant discrepancies even within the same class, affecting model performance. As shown in Figure 1, we focus on a realistic setting involving concurrent semantic shifts and covariate shifts [Yang *et al.*, 2024].

In this work, we present a Dual-Level Contrastive Learning (DLCL) framework to develop a clinically reliable model capable of accurate ID classification and OOD detection. To the best of our knowledge, this work represents the first rigorous attempt to adapt VLMs for OOD detection in FCU. Rather than relying on full-parameter fine-tuning, we adopt a parameter-efficient deep prompting strategy [Khattak *et al.*, 2023]. To tackle compound shifts, we introduce two complementary contrastive learning objectives. Local Contrastive Learning (LoCL) employs pairing based on the Bhattacharyya coefficient to align fine-grained feature representations at the batch level, effectively reducing acquisition noise and enhancing robustness to covariate shifts. Global Contrastive Learning (GlCL) leverages class prototypes coupled with an ambiguity-aware margin loss to regularize the feature space. This mechanism enforces intra-class compactness for known anatomies while ensuring the sufficient separability required to reject semantic-shift OOD samples. Experiments on both private FCU and public ISIC-OOD datasets confirm that DLCL surpasses state-of-the-art methods, demonstrating its effectiveness in handling compound shifts. Our contributions can be summarized as follows:

- We conduct a systematic study of compound shifts in FCU and identify a critical gap in existing benchmarks, which often fail to reflect realistic scenarios involving both semantic and covariate shifts.
- We propose a Dual-Level Contrastive Learning (DLCL) framework that integrates local instance-level contrastive learning and global prototype-level contrastive learning, effectively mitigating compound shifts in OOD detection.
- We validate our method via extensive experimental analysis on both a private multi-center FCU dataset and a public dataset. DLCL establishes a new state-of-the-art, demonstrating its effectiveness in handling realistic and complex distribution shifts.

2 Related Work

2.1 Pretrained Vision-Language Models

VLMs align images and text within a shared embedding space to facilitate cross-modal learning. By harnessing rich supervision signals from vast web-scale datasets [Li *et al.*, 2022], VLMs have achieved remarkable representational robustness. A prime example is CLIP [Radford *et al.*, 2021], which establishes state-of-the-art performance by training on 400 million natural image-text pairs. However, the lack of public medical data hinders the transfer of general models, prompting the development of domain-specific VLMs. BiomedCLIP [Zhang *et al.*, 2023] significantly scales up pretraining by leveraging 15 million scientific image-text pairs extracted from biomedical literature. UniMed-CLIP [Khattak *et al.*, 2024] introduces a unified pretraining paradigm designed for diverse medical imaging modalities. EchoCLIP [Christensen *et al.*, 2024] leverages contrastive learning on a large-scale dataset of over one million echocardiogram video-text pairs to establish a robust VLM for cardiac imaging. While these models may exhibit strong performance on ID samples, their reliability in detecting OOD samples remains largely unverified.

2.2 Prompt Tuning in Vision-Language Models

Full-parameter fine-tuning suffers from high computational demands and catastrophic forgetting. To address this, prompt tuning [Gu *et al.*, 2023; Cui *et al.*, 2025] has emerged as a parameter-efficient alternative that introduces a small set of learnable tokens while keeping the backbone frozen. Early approaches primarily focused on optimizing textual prompts within the language branch. For instance, CoOp [Zhou *et al.*, 2022b] replaces hand-crafted templates with learnable context vectors. Building on this, CoCoOp [Zhou *et al.*, 2022a] incorporates image semantic information to generate instance-level text prompts. To foster deeper cross-modal alignment, subsequent studies [Khattak *et al.*, 2023; Xing *et al.*, 2023] have advanced to multi-modal prompting, injecting learnable tokens into both vision and language encoders. In the medical domain, specific adaptations have also been explored. For example, FVP [Wang *et al.*, 2023] introduces Fourier visual prompting to enable source-free domain adaptive segmentation. MCPL [Wang *et al.*, 2024] jointly learns anatomy-pathology, text, and image prompts via a graph-guided collaboration module. However, despite the success of prompt learning in various downstream tasks, its potential for improving OOD detection in medical imaging remains underexplored.

2.3 Out-of-Distribution Detection

OOD detection is a fundamental safeguard in image analysis, ensuring that models can reliably identify inputs falling outside the training distribution. For instance, LoCoOp [Miyai *et al.*, 2023] introduces a few-shot OOD detection approach that enhances CLIP by learning localized, class-aware and class-agnostic prompts from limited ID data. In the medical domain, existing approaches often leverage statistical analysis of feature representations. For instance, FRODO [Calli *et al.*, 2022] proposed a training-free approach utilizing global normalized Mahalanobis distances on intermediate representations for versatile OOD detection in medical imaging. OpenMIBOOD [Gutbrod *et al.*, 2025] provides a comprehensive benchmark for medical OOD detection, systematically evaluating post-hoc methods across diverse imaging modalities. While significant advancements have been made in modalities such as X-ray, fundus imaging, dermoscopy, and CT [Hong *et al.*, 2024], the exploration of OOD detection in echocardiography remains nascent. Notably, current medical OOD detection research mainly focuses on post-hoc methods, often overlooking potential distributional discrepancies between training and test data. Consequently, research on generalized OOD detection tailored for FCU remains significantly limited.

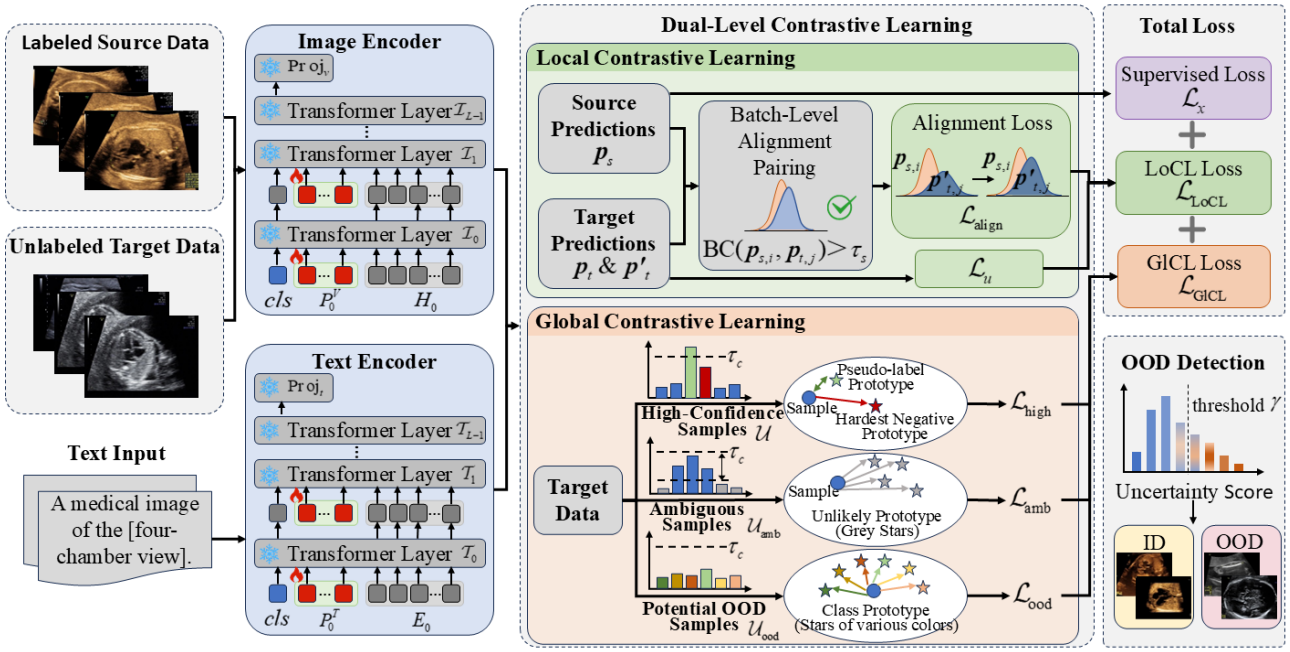


Figure 2: Overview of the DLCL framework for OOD detection in FCU. The model employs learnable text and visual prompts across various transformer layers. Local Contrastive Learning aligns predictions between source samples and the corresponding strongly augmented target samples, while Global Contrastive Learning regularizes representations via class prototypes.

3 Preliminaries

3.1 Problem Formulation

We investigate the challenge of adapting VLMs to FCU for compound-shift OOD detection. Formally, let $\mathcal{D}_s = \{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{N_s}$ denote a labeled source dataset drawn from the distribution $P_s(\boldsymbol{x}, y)$, where $\boldsymbol{x} \in \mathcal{X}$ and $y \in \mathcal{Y}_s$. Additionally, we assume access to an unlabeled target dataset $\mathcal{D}_t = \{\boldsymbol{x}_j^t\}_{j=1}^{N_t}$ drawn from $P_t(\boldsymbol{x})$. The target domain may diverge from the source due to two distinct types of distribution shift. Covariate shifts refer to a divergence in the marginal input distribution (i.e., $P_s(\boldsymbol{x}) \neq P_t(\boldsymbol{x})$) while the conditional distribution $P(y|\boldsymbol{x})$ remains invariant. Semantic shifts arise when the target label space expands beyond the source (i.e., $\mathcal{Y}_s \subset \mathcal{Y}_t$). We focus on a more realistic generalized OOD detection setting, where both covariate and semantic shifts occur simultaneously.

3.2 Contrastive Vision-Language Models

Our framework leverages CLIP [Radford *et al.*, 2021] as the foundational backbone. CLIP is designed to align visual and textual representations in a shared latent space through contrastive learning. It comprises two parallel encoders: an image encoder $\mathcal{I}(\cdot)$ and a text encoder $\mathcal{T}(\cdot)$. Following established conventions [Khattak *et al.*, 2023], we adopt a Transformer [Vaswani *et al.*, 2017] as the text encoder, and a Vision Transformer (ViT) [Dosovitskiy *et al.*, 2021] as the image encoder.

Image Encoder. Given a medical image $\boldsymbol{x}$, the image encoder first divides it into M fixed-size patches. These patches are projected into a sequence of patch embeddings $\boldsymbol{H}_0$, which acts as the input to the ViT layers. The encoder then processes $\boldsymbol{H}_0$ to extract the visual representation $\boldsymbol{f} = \mathcal{I}(\boldsymbol{x})$, where the class token ([cls]) is typically used as the aggregate feature.

Text Encoder. To encode semantic category information, CLIP utilizes a prompt template (e.g., "a medical image of the $[y_k]$"). For a specific category $y_k \in \{y_1, \ldots, y_K\}$, the template is instantiated and tokenized into a sequence of word embeddings $\boldsymbol{E}_0$. The text encoder processes $\boldsymbol{E}_0$ to generate the class-specific textual embedding $\boldsymbol{w}_k = \mathcal{T}(y_k)$.

The predicted probability that an image $\boldsymbol{x}$ belongs to category y_k is computed as

$$p(y = k \mid \boldsymbol{x}) = \frac{\exp\left(\cos(\boldsymbol{w}_k, \boldsymbol{f})/\tau\right)}{\sum_{j=1}^{K} \exp\left(\cos(\boldsymbol{w}_j, \boldsymbol{f})/\tau\right)}, \tag{1}$$

where τ is a learnable temperature parameter and $\cos(\cdot, \cdot)$ denotes the cosine similarity.

3.3 OOD Detection

OOD detection aims to construct a decision function $G(\cdot)$ that determines whether a test sample $\boldsymbol{x}$ originates from the ID or OOD distribution. This is realized via an uncertainty scoring function $\mathcal{S}(\boldsymbol{x})$ and a threshold γ [Hong *et al.*, 2024]:

$$G(\boldsymbol{x}) = \begin{cases} \text{ID}, & \text{if } \mathcal{S}(\boldsymbol{x}) < \gamma \\ \text{OOD}, & \text{if } \mathcal{S}(\boldsymbol{x}) \geq \gamma \end{cases}. \tag{2}$$

In this work, the uncertainty score $S(\boldsymbol{x})$ is defined as:

$$\mathcal{S}(\boldsymbol{x}) = \frac{-\sum_{k=1}^{K} p_k \log\left(p_k\right)}{\max_{1 \leq k \leq K} p_k}, \tag{3}$$

where $p_k = p(y = k \mid \boldsymbol{x})$ denotes the predicted probability for class k. The numerator measures predictive uncertainty via entropy, whereas the denominator normalizes by the highest class confidence. This metric amplifies the uncertainty signal for ambiguous samples. Higher scores reflect increased uncertainty, indicating a higher likelihood of OOD.

4 Method

In this section, we propose a Dual-Level Contrastive Learning (DLCL) framework for OOD detection in FCU. As shown in Figure 2, our method freezes the pre-trained VLM backbone and optimizes learnable prompts. The core of DLCL lies in combining two complementary objectives: Local Contrastive Learning and Global Contrastive Learning. Code is available at https://github.com/Liuziyi1999/DLCL.

4.1 Dual-Modal Deep Prompting

Standard prompt tuning methods typically insert prompts only at the input layer (shallow prompting). Recent studies [Khattak *et al.*, 2023] suggest that inserting prompts into deeper encoder blocks (deep prompting) facilitates the progressive learning of stage-wise features. We introduce learnable prompts into each transformer block of both the text and image encoders. The text encoder $\{\mathcal{T}_l\}_{l=0}^{L-1}$ comprises L transformer layers. For the l-th layer $\mathcal{T}_l$, we insert d learnable text prompt tokens $\boldsymbol{P}_l^T = \{\boldsymbol{t}_l^i\}_{i=1}^d$. The input to the l-th layer is constructed by concatenating the class token $\boldsymbol{s}_l$, the layer-specific prompts $\boldsymbol{P}_l^T$, and the word embeddings $\boldsymbol{E}_l$ from the previous layer. The forward process is formulated as:

$$[\boldsymbol{s}_{l+1}, \_, \boldsymbol{E}_{l+1}] = \mathcal{T}_l\left([\boldsymbol{s}_l, \boldsymbol{P}_l^T, \boldsymbol{E}_l]\right), \tag{4}$$

where $[\cdot, \cdot]$ denotes the concatenation operation. Then, the classification weight $\boldsymbol{w}_k$ for category k is obtained via the text projection layer:

$$\boldsymbol{w}_k = \text{Proj}_t(\boldsymbol{s}_L). \tag{5}$$

Similarly, the image encoder $\{\mathcal{I}_l\}_{l=0}^{L-1}$ consists of L transformer layers. We insert d learnable visual prompt tokens $\boldsymbol{P}_l^V = \{\boldsymbol{v}_l^i\}_{i=1}^d$ into each layer. Let $\boldsymbol{g}_l$ and $\boldsymbol{H}_l$ be the visual class token and patch embeddings from the previous layer, respectively. The processing at layer l is defined as:

$$[\boldsymbol{g}_{l+1}, \_, \boldsymbol{H}_{l+1}] = \mathcal{I}_l\left([\boldsymbol{g}_l, \boldsymbol{P}_l^V, \boldsymbol{H}_l]\right), \tag{6}$$

The final visual representation $\boldsymbol{f}$ is derived by projecting the class token from the last layer:

$$\boldsymbol{f} = \text{Proj}_v(\boldsymbol{g}_L). \tag{7}$$

4.2 Local Contrastive Learning (LoCL)

To mitigate covariate shifts arising from heterogeneous acquisition protocols, we introduce LoCL to enforce fine-grained feature alignment within each mini-batch. The Bhattacharyya Coefficient (BC) [Bhattacharyya, 1946] measures the distributional overlap between two probability vectors $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$. Unlike KL divergence, which is asymmetric and sensitive to low-probability mismatches, the BC offers a symmetric, bounded measure that emphasizes overlap in high-confidence regions of the probability simplex. This makes BC well suited for mini-batch level alignment, where robust and interpretable similarity thresholding is critical for filtering reliable source–target pairs. The BC and its corresponding distance metric f_{BC} are defined as:

$$\text{BC}(\boldsymbol{p}_i, \boldsymbol{p}_j) = \sqrt{\boldsymbol{p}_i}^\top \sqrt{\boldsymbol{p}_j}, f_{BC}(\boldsymbol{p}_i, \boldsymbol{p}_j) = 1 - \text{BC}(\boldsymbol{p}_i, \boldsymbol{p}_j). \tag{8}$$

A higher BC value indicates a stronger similarity.

Reliable source–target pairs are identified within each mini-batch. We select correctly classified source samples $\boldsymbol{x}_i^s$ (where $\arg\max \boldsymbol{p}_{s,i} = y_i^s$) as high-confidence samples and pair them with weakly augmented target samples $\boldsymbol{x}_j^t$ based on high similarity, measured by $\text{BC}(\boldsymbol{p}_{s,i}, \boldsymbol{p}_{t,j}) \geq \tau_s$. We then minimize the divergence between the source prediction $\boldsymbol{p}_{s,i}$ and the corresponding strongly augmented target prediction $\boldsymbol{p}'_{t,j}$. These pairs $(\boldsymbol{p}_{s,i}, \boldsymbol{p}'_{t,j})$ are treated as positive examples to promote source–target alignment. The alignment loss is formulated as:

$$\mathcal{L}_{\text{align}} = \frac{1}{|\mathcal{H}|} \sum_{(i,j)\in\mathcal{H}} f_{\text{BC}}\left(\boldsymbol{p}_{s,i}, \boldsymbol{p}'_{t,j}\right), \tag{9}$$

where $\mathcal{H}$ is the set of index pairs satisfying both confidence and similarity constraints.

We also leverage unlabeled target data to enhance the reliability of predictions. A target sample is considered valid if the maximum probability $p_j^{\max} = \max \boldsymbol{p}_{t,j}$ of its weakly augmented view exceeds a confidence threshold τ_c. The pseudo-label is assigned as $\tilde{y}_j = \arg\max \boldsymbol{p}_{t,j}$. Subsequently, we minimize the entropy of the averaged predictions from corresponding weakly and strongly augmented target samples:

$$\mathcal{L}_u = -\frac{1}{N_t} \sum_{j=1}^{N_t} \mathbb{1}(p_j^{\max} \geq \tau_c) \log\left(\frac{\boldsymbol{p}_{t,j,\tilde{y}_j} + \boldsymbol{p}'_{t,j,\tilde{y}_j}}{2}\right), \tag{10}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function.

The overall objective for LoCL is defined as:

$$\mathcal{L}_{\text{LoCL}} = \mathcal{L}_{\text{align}} + \lambda_u \mathcal{L}_u, \tag{11}$$

where λ_u is a hyperparameter balancing the two terms. This objective effectively reduces the domain gap caused by covariate shifts, establishing a robust feature space for subsequent global contrastive learning.

4.3 Global Contrastive Learning (GlCL)

Complementing the local feature alignment of LoCL, GlCL is specifically designed to address semantic shifts by enforcing global categorical constraints. The objective of GlCL is to maintain well-separated decision boundaries for known classes while effectively identifying potential OOD samples.

Class Prototype Extraction

We represent each semantic category using a global prototype $\boldsymbol{c}_k$, which serves as a stable representative in the shared latent space. To mitigate the inherent noise in ultrasound pseudo-labels and enhance prototype reliability, we utilize both labeled source samples and high-confidence target samples. At the end of each training epoch e, a temporary prototype $\hat{\boldsymbol{c}}_k^{(e)}$

Method	3VT	3VV	4CV	AA	LVOT	RVOT	TAV	Avg.	Acc_S	Acc_U	AUROC	H
CLIP*	32.83	4.03	75.70	8.56	2.60	10.46	70.12	29.19	2.67	98.29	48.75	5.20
EchoCLIP*	3.57	55.83	13.67	0.80	5.76	2.36	21.26	14.75	1.02	**99.89**	46.25	2.02
BiomedCLIP	40.13	19.81	60.77	85.59	47.05	20.09	69.88	49.05	43.12	8.87	40.27	14.71
MedCLIP	56.48	25.45	75.49	90.59	60.15	36.45	87.26	61.70	50.15	20.78	43.82	29.38
UniMedCLIP	58.26	27.64	77.54	92.34	62.67	38.76	89.48	63.81	61.75	29.21	46.12	39.66
CoOp	86.21	50.73	90.84	91.24	81.66	46.62	96.88	77.74	40.50	93.16	73.06	56.45
LoCoOp	87.25	51.65	91.81	92.14	82.56	50.49	94.64	78.65	51.87	81.27	76.76	63.34
CoCoOp	85.35	49.82	89.92	89.90	80.39	45.79	95.28	76.64	46.15	86.11	74.65	60.08
VPT	89.37	66.88	93.62	85.17	88.49	63.44	**99.68**	83.81	57.40	81.12	79.57	67.23
MaPLe	88.63	55.36	92.01	93.37	85.08	50.68	99.12	80.61	48.82	86.17	79.41	62.33
DLCL (Ours)	**96.67**	**88.90**	**97.01**	**95.39**	**91.75**	**88.94**	98.90	**93.94**	**76.79**	84.25	**89.61**	**80.35**

Table 1: DLCL results on the MFCU dataset in the OOD setting. The left columns display per-class accuracy (%) under covariate shifts. The right columns report the average accuracy for base (Acc_S) and novel (Acc_U) classes, along with the harmonic mean (H) and AUROC, under combined covariate and semantic shifts. Models marked with * are directly applied for inference. The best results are highlighted in **bold**.

for each class $k \in \{1, \dots, K\}$ is computed by aggregating feature representations:

$$\hat{\boldsymbol{c}}_k^{(e)} = \frac{\sum_{i=1}^{N_s} \delta_{s,i}^{(k)} \boldsymbol{f}_{s,i} + \sum_{j=1}^{N_t} \delta_{t,j}^{(k)} \boldsymbol{f}_{t,j}}{\sum_{i=1}^{N_s} \delta_{s,i}^{(k)} + \sum_{j=1}^{N_t} \delta_{t,j}^{(k)}}, \tag{12}$$

where $\delta_{s,i}^{(k)} = \mathbb{1}(y_i^s = k)$ and $\delta_{t,j}^{(k)} = \mathbb{1}(p_j^{\max} \geq \tau_c)\mathbb{1}(\tilde{y}_j = k)$ are indicator functions for source labels and confident target pseudo-labels, respectively. We update global prototypes via momentum-based Exponential Moving Average [Tarvainen and Valpola, 2017] to stabilize the feature space:

$$\boldsymbol{c}_k^{(e)} = (1 - \alpha)\boldsymbol{c}_k^{(e-1)} + \alpha \hat{\boldsymbol{c}}_k^{(e)}, \tag{13}$$

where $\alpha \in (0, 1]$ is the momentum coefficient.

Prototypical Contrastive Learning

To facilitate discriminative representation learning across the target domain, we employ a margin-aware prototypical contrastive loss. Let $s_{j,k} = \boldsymbol{f}_j^\top \boldsymbol{c}_k / T$ denote the temperature-scaled cosine similarity between the feature representation $\boldsymbol{f}_j$ of a sample $\boldsymbol{x}_j^t$ and the class prototype $\boldsymbol{c}_k$. We partition the target samples into a high-confidence set $\mathcal{U}$ and a low-confidence set $\bar{\mathcal{U}}$ based on the threshold τ_c. For the high-confidence samples in $\mathcal{U}$, our goal is to enforce compactness with their pseudo-labeled prototypes while maximizing the margin from the prototypes of the hardest negative classes. Specifically, for each sample $\boldsymbol{x}_j^t \in \mathcal{U}$, we identify the hardest negative class $k_j^- = \arg\max_{k \neq \tilde{y}_j} \boldsymbol{p}_{j,k}$. To accommodate the varying intra-class diversity of fetal heart views, we define an adaptive margin $m^t = \mu^t + \sigma^t$, where μ^t and σ^t denote the mean and standard deviation of the similarity gaps between the hardest negative prototype and the pseudo-labeled prototype across the batch. The contrastive loss for high-confidence samples is defined as:

$$\mathcal{L}_{\text{high}} = \frac{1}{|\mathcal{U}|} \sum_{j \in \mathcal{U}} \left[s_{j,k_j^-} - s_{j,\tilde{y}_j} + m^t \right]_+, \tag{14}$$

where $[\cdot]_+ = \max(0, \cdot)$ is the hinge function. This term enforces a margin between the pseudo-labeled and hardest negative prototypes.

Unlike previous methods that discard low-confidence samples, we argue that $\bar{\mathcal{U}}$ contains valuable information regarding decision boundaries and OOD patterns. To exploit this, we further partition $\bar{\mathcal{U}}$ into two disjoint subsets based on the shape of their prediction distributions: Ambiguous Samples ($\mathcal{U}_{\text{amb}}$) and Potential OOD Samples ($\mathcal{U}_{\text{ood}}$). Formally, a sample is categorized as ambiguous if it exhibits a distinguishable but weak peak, satisfying the gap condition: $\exists\, k \neq \tilde{y}_j,\ p_j^{\max} - p_{t,j,k} > \delta$, where δ represents a fixed margin of 0.4. For each sample in $\mathcal{U}_{\text{amb}}$, we minimize its similarity to the corresponding set of unlikely classes $\mathcal{K}_j^\dagger$ (where the gap condition holds), thereby pruning the negative space. Let $\bar{s}_j = |\mathcal{K}_j^\dagger|^{-1} \sum_{k \in \mathcal{K}_j^\dagger} s_{j,k}$ denote the average similarity to these unlikely prototypes. We then compute the adaptive margin $m^d = \mu^d + \sigma^d$ using the mean μ^d and standard deviation σ^d of these average similarities. The ambiguity loss is defined as:

$$\mathcal{L}_{\text{amb}} = \frac{1}{|\mathcal{U}_{\text{amb}}|} \sum_{j \in \mathcal{U}_{\text{amb}}} \left[\bar{s}_j - m^d \right]_+. \tag{15}$$

Conversely, samples violating the gap condition typically exhibit flat, high-entropy distributions, often indicative of unknown anatomical structures. We identify these instances as potential OOD samples, denoted by $\mathcal{U}_{\text{ood}} = \bar{\mathcal{U}} \setminus \mathcal{U}_{\text{amb}}$. Since these samples should not correlate strongly with any known category, we enforce a global suppression constraint. We calculate the average similarity across all K prototypes as $\hat{s}_j = \frac{1}{K} \sum_{k=1}^{K} s_{j,k}$. To enforce separation, we define an adaptive margin $m^o = \mu^o + \sigma^o$, derived from the mean μ^o and standard deviation σ^o of the similarity distribution $\{\hat{s}_j\}_{j \in \mathcal{U}_{\text{ood}}}$. The OOD suppression loss encourages these instances to maintain low similarity with all base prototypes:

$$\mathcal{L}_{\text{ood}} = \frac{1}{|\mathcal{U}_{\text{ood}}|} \sum_{j \in \mathcal{U}_{\text{ood}}} \left[\hat{s}_j - m^o \right]_+. \tag{16}$$

Finally, the total global objective integrates these three complementary components:

$$\mathcal{L}_{\text{GlCL}} = \mathcal{L}_{\text{high}} + \lambda_d \mathcal{L}_{\text{amb}} + \lambda_o \mathcal{L}_{\text{ood}}, \tag{17}$$

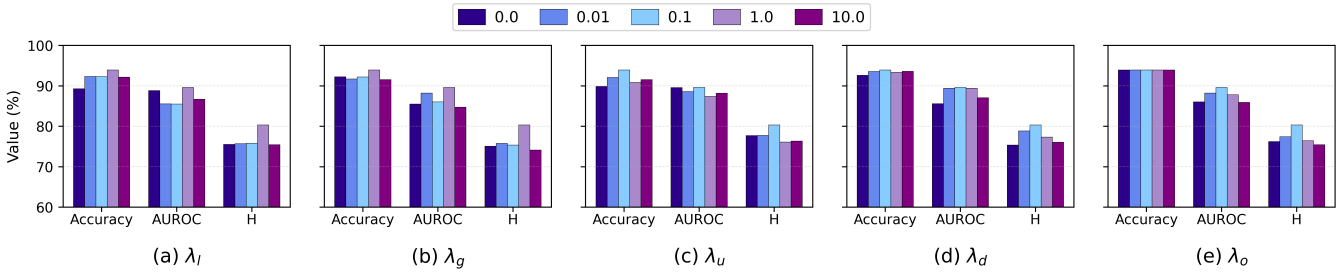


Figure 3: Hyperparameter sensitivity analysis. We report the average per-class accuracy, AUROC, and harmonic mean (H) across different values of λ. Based on the observed peak trends, we adopt the optimal values for our final model.

Method	Accuracy	Acc_S	Acc_U	AUROC	H
w/o $\mathcal{L}_{\text{align}}$ (Eq. 9)	93.78	73.44	84.96	89.10	78.78
w/o $\mathcal{L}_u$ (Eq. 10)	89.88	69.43	**88.19**	89.53	77.70
w/o $\mathcal{L}_{\text{LoCL}}$ (Eq. 11)	89.30	68.66	86.18	88.82	75.53
w/o $\mathcal{L}_{\text{high}}$ (Eq. 14)	93.09	67.88	84.17	86.24	75.15
w/o $\mathcal{L}_{\text{amb}}$ (Eq. 15)	92.32	69.77	82.79	85.60	75.37
w/o $\mathcal{L}_{\text{ood}}$ (Eq. 16)	**93.94**	71.65	81.40	86.05	76.22
w/o $\mathcal{L}_{\text{GlCL}}$ (Eq. 17)	92.25	74.22	82.83	85.49	75.10
DLCL (Ours)	**93.94**	**76.79**	84.25	**89.61**	**80.35**

Table 2: Ablation study of DLCL components. The results are reported on the MFCU dataset, following the same metric definitions as in Table 1.

where λ_d and λ_o are weighting factors that balance the contributions of the corresponding loss terms. GlCL effectively preserves class discriminability while alleviating noise propagation caused by open-set distractors.

4.4 Overall Loss Function

To foster discriminative feature learning on the source domain, we train the model via the standard supervised cross-entropy loss on the labeled source set:

$$\mathcal{L}_x = -\frac{1}{N_s}\sum_{i=1}^{N_s}\log\left(p_{s,i,y_i^s}\right). \tag{18}$$

The final training objective of DLCL integrates the supervised source loss with the proposed local and global contrastive constraints to jointly handle compound shifts. The total loss is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_x + \lambda_l\mathcal{L}_{\text{LoCL}} + \lambda_g\mathcal{L}_{\text{GlCL}}, \tag{19}$$

where λ_l and λ_g are hyperparameters controlling the contributions.

5 Experiments and Results

5.1 Datasets and Experimental Setup

Multi-center Fetal Cardiac Ultrasound (MFCU) Dataset. We collected B-mode fetal echocardiography scans from pregnant women between 2018 and 2021. For each scan, we accessed freeze-frame images saved by sonographers during the exam. The MFCU dataset comprises 22,050 source-domain images acquired from the Key Laboratory of Maternal-Fetal Medical Research and provided by Anzhen Hospital. The target domain includes 49,658 images collected from 38 medical institutions participating in the 13th Five-Year National Key Research and Development Plan. The dataset is annotated with seven standard fetal cardiac views: Three-Vessel and Trachea View (3VT), Three-Vessel View (3VV), Four-Chamber View (4CV), Aortic Arch View (AA), Left Ventricular Outflow Tract View (LVOT), Right Ventricular Outflow Tract View (RVOT), and Transverse Abdominal View (TAV). Notably, the target domain includes 8,823 images of non-standard views. All images were manually reviewed by multiple experienced clinicians, and discrepancies were resolved by consensus to ensure annotation accuracy.

ISIC-OOD Dataset. We constructed the ISIC-OOD dataset based on the ISIC2019 dataset [Tschandl *et al.*, 2018]. The base classes include basal cell carcinoma (bcc), benign keratosis (bkl), melanoma (mel), and melanocytic nevi (nv), while the remaining categories are treated as novel classes. To construct samples exhibiting semantic shift, we select images that share the same imaging modality as the base classes but belong to novel classes. To simulate covariate shift, we retain the base classes while introducing distributional variations due to differences in imaging devices and acquisition protocols. The resulting dataset is divided into two domains: BCN and HAM, representing different acquisition protocols. Specifically, the BCN domain contains 11,010 images from base classes and 1,404 from novel classes, while the HAM domain comprises 9,431 base-class images and 585 novel-class images. When a specific domain is used as the source domain, only the base class images are included. This setup allows for the assessment of robustness against concurrent covariate and semantic shifts. The dataset can be accessed at https://www.kaggle.com/datasets/liuzyiii/isic-ood.

Implementation Details. Our experiments were implemented in PyTorch [Paszke *et al.*, 2019] on a single NVIDIA V100 GPU. We utilized the pretrained ViT-B/16 CLIP model for prompt tuning, with word embeddings of 512 dimensions and patch embeddings of 768 dimensions. The text and visual prompts were updated using the SGD optimizer. We trained the model for a total of 60 epochs, where the first 50 epochs used only the supervised objective. The length of the text and visual prompts was set to 8, and the depth was set to 12. Data augmentation was performed using RandAugment (subset size = 2). The batch size was 64, and the learning

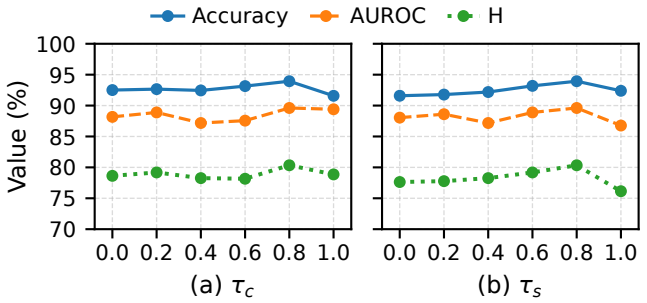
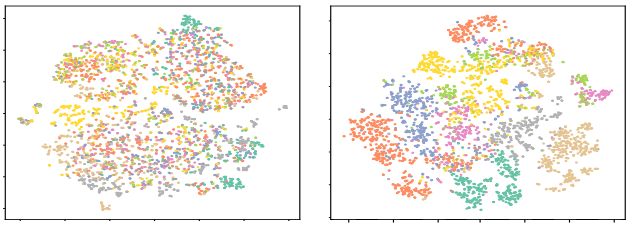


Figure 4: Ablation on τ_c and τ_s in our method. We report the average per-class classification accuracy (%), AUROC, and harmonic mean accuracy (H) on the MFCU dataset.



(a) CLIP (b) DLCL

Figure 5: T-SNE visualization of learned representations. Comparison between (a) CLIP and (b) DLCL regarding class discriminability, where colors denote different categories.

rate was set to 0.0025. We set the confidence thresholds as $\tau_s = \tau_c = 0.8$ for all experiments. Based on the experiments, we empirically set [λ_l,λ_g,λ_u,λ_d,λ_o] to [1,1,0.1,0.1,0.1].

5.2 Results and Analysis

We benchmark our method against ten methods, including general VLMs (CLIP [Radford *et al.*, 2021]), medical-specific VLMs (BiomedCLIP [Zhang *et al.*, 2023], MedCLIP [Wang *et al.*, 2022], UniMedCLIP [Khattak *et al.*, 2024], EchoCLIP [Christensen *et al.*, 2024]), and prompt learning strategies (CoOp [Zhou *et al.*, 2022b], LoCoOp [Miyai *et al.*, 2023], CoCoOp [Zhou *et al.*, 2022a], VPT [Jia *et al.*, 2022], and MaPLe [Khattak *et al.*, 2023]). The quantitative results on the MFCU dataset are detailed in Table 1. Traditional VLMs, such as CLIP, that do not specialize in medical domains, show a notably low accuracy of 29.19% on average, further indicating the difficulty of generalization between the source (natural images) and target (ultrasound images) domains. As EchoCLIP is mainly trained on adult clinical echocardiography data, its generalization capability is limited when transferring to fetal echocardiography. Medical-specific VLMs like BiomedCLIP (49.05%) and MedCLIP (61.70%) outperform CLIP but still struggle to achieve high accuracy under OOD conditions, highlighting their limitations in bridging the domain gap. In contrast, models employing prompt learning strategies, such as CoOp and CoCoOp, show notable improvements when optimized using a combination of labeled samples from the source domain and unlabeled samples from the target domain. CoOp achieves an average accuracy of 77.74%, suggesting its better adaptability to domain shifts compared to the general VLMs. Finally, our proposed method not only achieves the highest average accuracy (93.94%) under covariate shifts but also secures the best balance between base ($Acc_S = 76.79\%$) and novel class ($Acc_U = 84.25\%$) recognition under compound shifts. The harmonic mean (80.35%) and AUROC (89.61%) validate its superior OOD detection potential in clinical environments.

5.3 Ablation Studies

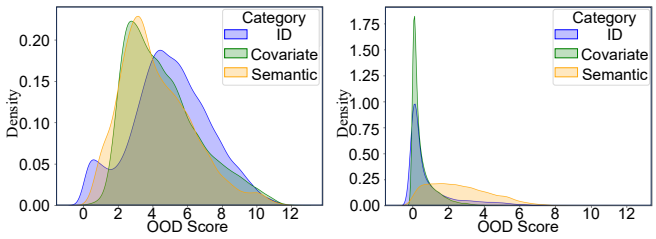
Contribution of Different Losses. Table 2 summarizes the impact of each component. The exclusion of either the local or global contrastive losses leads to the most significant performance degradation, underscoring the necessity of our dual-level alignment strategy. Removing $\mathcal{L}_{\text{LoCL}}$ results in a sharp accuracy drop (93.94% $\rightarrow$ 89.30%), verifying its critical role in mitigating covariate shifts. Meanwhile, the absence of $\mathcal{L}_{\text{GlCL}}$ degrades both harmonic mean and AUROC, demonstrating that global prototype constraints are essential for maintaining semantic consistency. Furthermore, ablating $\mathcal{L}_{\text{amb}}$ reduces the harmonic mean to 75.37%, validating that leveraging low-confidence samples effectively refines decision boundaries. In compound-shift scenarios, $\mathcal{L}_{\text{ood}}$ proves indispensable for identifying novel classes. Since $\mathcal{L}_{\text{ood}}$ is only active under compound shifts, its ablation does not affect covariate-only accuracy. In summary, each component of DLCL plays a distinctive and critical role.

Impact of Loss Components. We analyze hyperparameter sensitivity by varying each parameter individually. Figure 3(a) reveals that increasing the weight of LoCL (λ_l) consistently improves performance, peaking at $\lambda_l = 1$. Notably, excessive global regularization (λ_g) may compromise feature discriminability. Therefore, we set $\lambda_g = 1$. Figure 3(c) indicates that λ_u notably boosts the AUROC and harmonic mean, with an optimal trade-off around $\lambda_u = 0.1$. Conversely, Figure 3(d) shows that setting $\lambda_d = 0.1$ significantly improves OOD metrics under combined shifts. Since $\mathcal{L}_{\text{ood}}$ is designed to address semantic shifts, λ_o mainly affects OOD-related metrics rather than covariate-shift classification accuracy. We therefore default to $\lambda_d = \lambda_o = 0.1$.

Effect of τ_c and τ_s. Empirically, source predictions aligned with ground truth typically exhibit a near one-hot distribution. Under this condition, the similarity constraint effectively degenerates into a unidirectional confidence check on the target prediction. Given that model predictions tend to polarize toward either high or low confidence, the set of selected pairs remains stable. This characteristic renders the loss formulation robust to variations in the hyperparameter τ_s. As shown in Figure 4, the proposed model demonstrates stable performance across a wide range of values for both τ_c and τ_s. Although slight variations are observed in accuracy, AUROC, and harmonic mean across different hyperparameter settings, the overall trend remains stable. These findings confirm that DLCL maintains strong robustness to both τ_c and τ_s, highlighting its reliability across varying calibration thresholds.

Method	BCN2HAM									HAM2BCN								
	bcc	bkl	nv	mel	Avg.	Acc_S	Acc_U	AUROC	H	bcc	bkl	nv	mel	Avg.	Acc_S	Acc_U	AUROC	H
CLIP*	65.95	13.01	78.21	29.83	46.75	0.61	**99.14**	55.33	1.22	3.13	12.48	90.28	37.00	35.72	1.90	**98.86**	60.56	3.73
CoOp	74.01	32.65	89.87	43.31	59.96	43.55	94.06	75.01	59.53	36.92	32.39	**94.44**	23.10	46.71	28.75	81.18	65.44	42.46
LoCoOp	80.15	34.65	92.56	44.35	62.93	44.56	93.45	75.87	60.35	38.85	33.96	94.23	30.56	49.40	29.65	80.62	66.44	43.36
CoCoOp	75.49	33.28	91.93	44.53	61.31	35.65	96.09	71.02	52.01	31.26	24.52	88.90	29.51	43.55	24.01	87.53	63.46	37.69
VPT	79.96	41.13	95.56	37.02	63.42	51.21	88.18	79.54	64.80	58.31	28.91	92.84	30.07	52.53	33.26	76.12	65.35	46.29
MaPLe	77.43	36.03	**96.97**	29.65	60.02	49.35	92.98	**81.77**	64.48	61.77	16.70	94.39	17.50	47.59	36.99	77.33	**70.66**	50.05
DLCL (Ours)	**90.47**	**46.59**	93.60	**44.56**	**68.81**	**60.39**	83.05	80.19	**69.93**	**62.08**	**39.28**	91.28	**48.51**	**60.29**	**42.90**	73.84	70.20	**54.27**

Table 3: DLCL results on the ISIC-OOD dataset in the OOD setting. Following the same metric definitions as in Table 1.



(a) CLIP (b) DLCL

Figure 6: Visualization of OOD score distributions $S(\boldsymbol{x})$. Semantic shifts (yellow) are compounded by simultaneous covariate shifts.

Feature Visualization Analysis. To qualitatively evaluate feature alignment, we employ t-SNE [Maaten and Hinton, 2008] to visualize feature distributions, in which source and target samples are projected into a shared space. Regarding class discriminability, the baseline (Fig. 5a) suffers from significant overlap, whereas DLCL (Fig. 5b) achieves high intra-class compactness and distinct boundaries. These results verify that DLCL effectively learns domain-invariant representations, thereby ensuring both robust cross-domain alignment and superior separability.

Visualization of OOD Score Distributions. For practical deployment, a robust medical VLM should detect OOD samples under various distribution shifts and respond accordingly, e.g., by routing covariate-shifted samples to specialized expert models for reliable ID recognition. We visualize the distribution of OOD scores $S(\boldsymbol{x})$ to assess separability. The baseline CLIP (Fig. 6a) fails to distinguish ID from OOD samples, exhibiting significant overlap. In contrast, DLCL (Fig. 6b) successfully disentangles these distributions: samples with covariate shifts (green) are aligned with the ID cluster (low scores), whereas those with compound shifts (yellow) are effectively pushed toward high scores. This confirms that DLCL tolerates acquisition-related variance while robustly rejecting anatomical anomalies. In this paper, we empirically set $\gamma = 0.8$. By adjusting the detection threshold γ, the model can flexibly control its sensitivity to semantic shifts while preserving robustness to covariate variations.

5.4 Results on Additional Datasets

To further validate the generalizability of our method across different medical image domains, we extend the evaluation to the ISIC-OOD dataset under two cross-domain settings, i.e., BCN2HAM and HAM2BCN. As shown in Table 3, several baseline methods achieve relatively high Acc_U. However, such improvements on novel classes are often accompanied by a clear degradation in base class recognition, leading to inferior overall performance and a suboptimal balance between base and novel classes. For example, in HAM2BCN, CLIP obtains an Acc_U of 98.86%, but its Acc_S and harmonic mean are only 1.90% and 3.73%, respectively. Similarly, CoCoOp achieves a higher Acc_U than DLCL, but its Acc_S and H are substantially lower.

In contrast, DLCL achieves more balanced cross-domain performance. In the BCN2HAM setting, DLCL obtains the best average accuracy of 68.81% and the highest harmonic mean of 69.93%, outperforming the strongest baseline by 5.39% and 5.13%, respectively. In the HAM2BCN setting, although some baselines report higher Acc_U, DLCL achieves the best Acc_S of 42.90%, the best average accuracy of 60.29%, and the highest harmonic mean of 54.27%. These results indicate that DLCL does not simply bias the prediction toward novel classes but instead improves the overall discrimination ability across both base and novel categories.

6 Conclusion

In this work, we present a study on adapting VLMs for OOD detection in FCU, addressing the practical challenges induced by both clinical covariate and semantic shifts. We propose DLCL, a specialized framework that incorporates a prompt learning module to enable efficient fine-tuning of VLMs. Beyond model adaptation, DLCL combines instance-level Local Contrastive Learning and prototype-level Global Contrastive Learning to encourage the acquisition of discriminative and transferable representations across heterogeneous distributions. Extensive experimental results highlight the potential of VLM-based OOD detection in building trustworthy and generalizable medical imaging systems for real-world deployment. We hope our work will inspire future research on multi-modal OOD detection in medical imaging under realistic cross-hospital distribution shifts.

Ethical Statement

This study was conducted in compliance with the ethical standards required by IJCAI. It is based on a retrospective analysis of fetal cardiac ultrasound data collected during routine clinical practice, with approval from the corresponding Ethics Committee or Institutional Review Board. All data were fully

anonymized prior to access by the authors, and informed consent was obtained from the patients or their legal guardians, or waived in accordance with local regulations for retrospective studies. The proposed method was evaluated as a clinical decision support tool in a research setting.

From a clinical perspective, we explicitly consider safety-critical risks associated with deploying vision-language models in healthcare. Overconfident predictions on out-of-distribution samples may pose serious threats to patient safety, particularly in prenatal diagnosis. To mitigate such risks, our work focuses on reliable out-of-distribution detection rather than direct clinical decision-making, aiming to provide uncertainty-aware signals that can assist clinicians instead of replacing expert judgment. Moreover, we acknowledge potential issues related to dataset bias and cross-center heterogeneity. By explicitly modeling covariate and semantic shifts across multiple institutions, our method seeks to improve robustness and reduce performance disparities caused by data distribution biases. Overall, this research is intended to support trustworthy and responsible AI development in medical imaging, with careful consideration of ethical, safety, and fairness concerns in real-world clinical deployment.

Acknowledgements

This work was supported in part by the National Science and Technology Major Project (2025ZD0123401), in part by the National Natural Science Foundation of China under Grant 62406014, in part by the Beijing Natural Science Foundation (IS2408I, 7262079), in part by the Zhejiang Natural Science Foundation (LZ26F020012) and in part by the Start-up Funds of Hangzhou International Innovation Institute of Beihang University under Grant No. 2024KQ045 and No. 2024KQ027.

References

[Bhattacharyya, 1946] Anil Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics*, pages 401–406, 1946.

[Calli *et al.*, 2022] Erdi Calli, Bram Van Ginneken, Ecem Sogancioglu, and Keelin Murphy. Frodo: An in-depth analysis of a system to reject outlier samples from a trained neural network. *IEEE Transactions on Medical Imaging*, 42(4):971–981, 2022.

[Christensen *et al.*, 2024] Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine*, 30(5):1481–1488, 2024.

[Cui *et al.*, 2025] Chaoran Cui, Ziyi Liu, Shuai Gong, Lei Zhu, Chunyun Zhang, and Hui Liu. When adversarial training meets prompt tuning: Adversarial dual prompt tuning for unsupervised domain adaptation. *IEEE Transactions on Image Processing*, 34:1427–1440, 2025.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[Fei and Liu, 2016] Geli Fei and Bing Liu. Breaking the closed world assumption in text classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 506–514, 2016.

[Gu *et al.*, 2023] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023.

[Gutbrod *et al.*, 2025] Max Gutbrod, David Rauber, Danilo Weber Nunes, and Christoph Palm. Openmibood: Open medical imaging benchmarks for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25874–25886, 2025.

[Hendrycks and Gimpel, 2017] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

[Hong *et al.*, 2024] Zesheng Hong, Yubiao Yue, Yubin Chen, Lele Cong, Huanjie Lin, Yuanmei Luo, Mini Han Wang, Weidong Wang, Jialong Xu, Xiaoqi Yang, et al. Out-of-distribution detection in medical image analysis: A survey. *arXiv preprint arXiv:2404.18279*, 2024.

[Jia *et al.*, 2022] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pages 709–727, 2022.

[Khattak *et al.*, 2023] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023.

[Khattak *et al.*, 2024] Muhammad Uzair Khattak, Shahina Kunhimon, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities. *arXiv preprint arXiv:2412.10372*, 2024.

[Li *et al.*, 2022] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2022.

[Liang *et al.*, 2018] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[Ming *et al.*, 2022] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35:35087–35102, 2022.

[Miyai *et al.*, 2023] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. *Advances in Neural Information Processing Systems*, 36:76298–76310, 2023.

[Noble and Boukerroui, 2006] J Alison Noble and Djamal Boukerroui. Ultrasound image segmentation: a survey. *IEEE Transactions on medical imaging*, 25(8):987–1010, 2006.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8024–8035, 2019.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021.

[Silva-Rodriguez *et al.*, 2025] Julio Silva-Rodriguez, Hadi Chakor, Riadh Kobbi, Jose Dolz, and Ismail Ben Ayed. A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. *Medical Image Analysis*, 99:103357, 2025.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, pages 1195–1204, 2017.

[Tschandl *et al.*, 2018] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008, 2017.

[Wang *et al.*, 2022] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876, 2022.

[Wang *et al.*, 2023] Yan Wang, Jian Cheng, Yixin Chen, Shuai Shao, Lanyun Zhu, Zhenzhou Wu, Tao Liu, and Haogang Zhu. Fvp: Fourier visual prompting for source-free unsupervised domain adaptation of medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(12):3738–3751, 2023.

[Wang *et al.*, 2024] Pengyu Wang, Huaqi Zhang, and Yixuan Yuan. Mcpl: multi-modal collaborative prompt learning for medical vision-language model. *IEEE Transactions on Medical Imaging*, 43(12):4224–4235, 2024.

[Xing *et al.*, 2023] Yinghui Xing, Qirui Wu, De Cheng, Shizhou Zhang, Guoqiang Liang, Peng Wang, and Yanning Zhang. Dual modality prompt tuning for vision-language pre-trained model. *IEEE Transactions on Multimedia*, 26:2056–2068, 2023.

[Yang *et al.*, 2024] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

[Zhang *et al.*, 2023] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

[Zhou *et al.*, 2022a] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

[Zhou *et al.*, 2022b] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.