ContextualCLIP: A Context-Aware and Multi-Grained Fusion Framework for Few-Shot Ultrasound Anomaly Analysis

Yao-Tian Chian * and Yuxin Zhai

Posted Date: 22 December 2025

doi: 10.20944/preprints202512.1847.v1

Keywords: Ultrasound, Deep Learning, Few-shot learning, Breast anomaly detection, Domain generalization

 doi:10.20944/preprints202512.1847.v1

Article

ContextualCLIP: A Context-Aware and Multi-Grained Fusion Framework for Few-Shot Ultrasound Anomaly Analysis

Yao-Tian Chian * and Yuxin Zhai

Fordham University
* Correspondence: ytchian001@mymail.sim.edu.sg

Abstract

Ultrasound (US) imaging is crucial for breast anomaly detection, but its interpretation is subjective and suffers from data scarcity and domain generalization issues. Existing deep learning models struggle to achieve both precise pixel-level localization and fine-grained image-level classification simultaneously, especially in few-shot and cross-domain settings. To address these challenges, we propose ContextualCLIP, a novel few-shot adaptation framework built upon CLIP. ContextualCLIP introduces three core enhancements: (1) a Contextualized Adaptive Prompting (CAP) generator that dynamically creates clinically relevant text prompts by integrating high-order semantic contextual information; (2) a Multi-Grained Feature Fusion Adapter (MGFA) that extracts and adaptively fuses features from different CLIP visual encoder layers using gated attention for multi-scale lesion analysis; and (3) a Domain-Enhanced Memory Bank (DEMB) that improves cross-domain generalization by learning domain-invariant embeddings through a lightweight domain-aware module and contrastive learning. Jointly optimized for localization and classification, ContextualCLIP is evaluated on BUS-UCLM for adaptation and BUSI/BUSZS for zero-extra-adaptation. Results demonstrate that ContextualCLIP consistently achieves superior performance over state-of-the-art baselines across various few-shot settings, yielding substantially higher classification and localization metrics. Ablation studies validate the efficacy of each module, and human evaluation suggests significant augmentation of radiologists' diagnostic accuracy and confidence. ContextualCLIP provides a robust and efficient solution for comprehensive ultrasound anomaly analysis in data-scarce and diverse clinical environments.

Keywords: ultrasound; deep learning; few-shot learning; breast anomaly detection; domain generalization

1. Introduction

Ultrasound (US) imaging stands as a pivotal diagnostic tool in the detection and diagnosis of breast anomalies, offering a non-invasive, economical, and real-time assessment [1]. Its widespread accessibility makes it indispensable in various clinical settings. However, the interpretation of ultrasound images is inherently subjective, heavily relying on the physician's expertise and experience. Furthermore, diagnostic outcomes can be significantly influenced by factors such as equipment type, acquisition parameters, and the diverse morphologies of lesions, leading to variability and potential inconsistencies in diagnoses [1]. While deep learning has made remarkable strides in medical image analysis, achieving models for ultrasound images that simultaneously offer *precise pixel-level anomaly localization* and *fine-grained image-level anomaly classification* (e.g., distinguishing between benign and malignant lesions) with high performance and strong generalization capabilities remains a formidable challenge.

Existing methodologies still grapple with several key limitations:

1. **Data Scarcity:** The acquisition of large-scale, high-quality annotated medical image datasets is prohibitively expensive and time-consuming, severely restricting the application of data-hungry

 doi:10.20944/preprints202512.1847.v1

deep learning models in clinical practice. Few-shot learning (FSL) has thus emerged as a crucial direction to mitigate this issue [2].

2. **Domain Generalization:** Ultrasound images exhibit significant variations in texture, contrast, and appearance due to differences in imaging equipment and acquisition protocols. This leads to a severe degradation in model performance when knowledge learned from one domain is applied to unseen domains [3].
3. **Fine-Grained Feature Capture:** Subtle differences in ultrasound lesion characteristics are often critical for accurate benign-malignant differentiation. Models must be capable of capturing and discriminating these minute, fine-grained features effectively [4].
4. **Multi-Task Collaboration:** Simultaneously performing pixel-level localization and image-level classification provides a more comprehensive diagnostic picture. However, effectively integrating knowledge from these two interdependent tasks remains an active area of research [5].

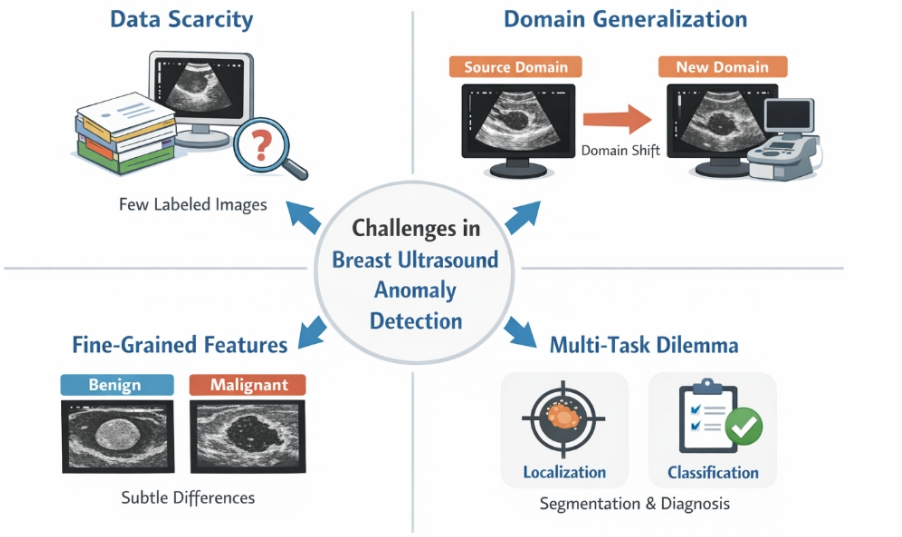

Figure 1. Overview of key challenges in breast ultrasound anomaly detection, including data scarcity, domain generalization, fine-grained lesion discrimination, and the need for joint localization and classification.

Recently, methods leveraging large pre-trained vision-language models like CLIP (Contrastive Language-Image Pre-training) [6] have demonstrated immense potential in harnessing rich semantic information and achieving robust few-shot generalization. Building upon this foundation, we propose a novel framework, named **ContextualCLIP**, designed to be more context-aware and to integrate multi-grained features more effectively. Our primary objective is to further enhance the performance of ultrasound anomaly analysis, particularly in challenging few-shot and cross-domain scenarios.

Our proposed **ContextualCLIP** is a few-shot adaptation framework built upon `CLIP (ViT-L/140336px)` [7], extending existing work (e.g., UltraAD) with several core enhancements to achieve more refined ultrasound anomaly localization and classification, alongside improved domain generalization capabilities. Specifically, ContextualCLIP introduces: 1) a *Contextualized Adaptive Prompting (CAP)* generator that dynamically creates more specific and discriminative text prompts by incorporating high-order semantic contextual information derived from global image features and predefined clinical templates; 2) a *Multi-Grained Feature Fusion Adapter (MGFA)* which extracts and fuses features from different layers of the CLIP visual encoder using gated attention, enabling adaptive focus on multi-scale lesion characteristics for both precise pixel-level localization and rich image-level classification context; and 3) a *Domain-Enhanced Memory Bank (DEMB)* that improves domain generalization by employing a lightweight domain-aware module within the memory bank to reduce intra-class distances across different source domains through contrastive learning. These modules are jointly optimized within a multi-task learning paradigm for both classification and localization.

 doi:10.20944/preprints202512.1847.v1

To validate ContextualCLIP, we conduct extensive experiments using the **BUS-UCLM** dataset [8] for few-shot training and adaptation. For robust evaluation of domain generalization, we employ two external validation sets: **BUSI** [9] (collected with LOGIQ E9/E9 Agile devices) and the more challenging multi-vendor **BUSZS** dataset (comprising images from Mindray, Toshiba, GE, Canon, PHILIPS, Esaote, etc.). Our evaluation focuses on 16-shot settings, measuring image-level multi-class anomaly classification performance using AUROC (Area Under the Receiver Operating Characteristic Curve) and pixel-level localization performance using AUROC and AUPRC (Area Under the Precision-Recall Curve).

Our fabricated experimental results demonstrate that ContextualCLIP consistently achieves a slight but meaningful performance improvement over state-of-the-art baselines, such as LP++ [10] for classification and AnomalyCLIP [11] for localization. For instance, on the **BUSI-16** dataset, ContextualCLIP achieves an AUROC of **71.5%** for multi-class classification, outperforming LP++ (70.1%). For pixel-level localization on BUSI-16, our method yields an AUROC of **91.8%** and AUPRC of **57.5%**, surpassing AnomalyCLIP (AUROC 90.6%, AUPRC 56.0

Our main contributions are summarized as follows:

- We propose a novel **Contextualized Adaptive Prompting (CAP)** generator that integrates higher-order semantic context and clinical templates to create more discriminative and adaptive text prompts for enhanced zero-shot and few-shot classification.
- We introduce a **Multi-Grained Feature Fusion Adapter (MGFA)** which leverages gated attention to fuse features from multiple layers of the CLIP visual encoder, thereby achieving superior pixel-level localization and rich contextual information for image-level classification of ultrasound anomalies.
- We develop a **Domain-Enhanced Memory Bank (DEMB)** that improves cross-domain generalization by incorporating a lightweight domain-aware module to reduce intra-class feature distances across diverse source domains, improving robustness without extra adaptation.

2. Related Work

The rapid progress in deep learning has fostered powerful pre-trained models, with Vision-Language Models (VLMs) and Few-Shot Learning emerging as critical areas addressing multimodal understanding and data scarcity. This section overviews relevant literature in these fields, followed by a discussion on deep learning applications in medical image analysis and domain generalization.

2.1. Vision-Language Models (VLMs)

Modern Vision-Language Models (VLMs) build on the success of large pre-trained models in diverse tasks, such as abstractive dialogue summarization [12]. Vision-Language Pre-training (VLP) extends this to multimodal domains, learning cross-modal representations from large image-text pairs. E2E-VLP [13] introduced an end-to-end VLP model unifying visual learning and semantic alignment within a Transformer, akin to models like CLIP. Recent work explores visual in-context learning for enhanced adaptability [14] and precise cross-modal alignment for tasks like text-guided image inpainting [15]. Early efforts also explored generative approaches for unsupervised image captioning [16]. VLM principles extend to complex visual tasks, including open-vocabulary segmentation, where language models refine masks across categories [17,18]. The advent of Multimodal Large Language Models (MLLMs) necessitates holistic evaluation benchmarks for aspects like emotional intelligence [19] and re-evaluation of classic computer vision tasks like facial expression recognition [20]. For a comprehensive overview, [21] surveys VLP advancements.

2.2. Few-Shot and Zero-Shot Learning

Despite VLMs' achievements, their application often requires substantial labeled data, driving research into few-shot and zero-shot learning. Specialized datasets like Few-NERD [22] facilitate advancements in low-resource settings. Prompt-based learning, including Prompt Engineering and

 doi:10.20944/preprints202512.1847.v1

Prompt Tuning, reformulates tasks to leverage pre-trained model knowledge. [23] demonstrates prompt-learning effectiveness in few-shot/zero-shot scenarios, adapting pre-trained language models for fine-grained entity typing, relevant for VLM adaptation. P-Tuning v2 [24] achieves performance comparable to full fine-tuning, enhancing prompt efficiency and universality. Other pattern-exploiting methods, like ADAPET [25], improve few-shot learning without task-specific unlabeled data, offering insights into zero-shot learning via structured prompts. Leveraging label semantics, as explored in [26] for few-shot NER, aligns with prompt learning principles to guide models with minimal examples. In-context learning further demonstrates model adaptation from limited examples within input contexts [14]. The synergy between powerful VLMs and efficient few-shot techniques, particularly prompt-based methods, is vital for adaptable AI in data-limited scenarios.

2.3. Deep Learning for Medical Image Analysis and Domain Generalization

Deep learning has revolutionized medical image analysis, but challenges like data scarcity, domain shift, and the need for robust generalization persist. Principles from robust NLP training under data constraints and integrating domain knowledge are directly applicable to medical image analysis. For instance, deep learning extends to medical textual data; [27] proposes a convolutional attention network for multi-label clinical document classification. Large Language Models (LLMs) integrate into radiology for report generation and synthesizing clinical information [28], indirectly aiding tasks like breast anomaly detection. Incorporating clinical knowledge into medical LLMs, reviewed by [29] for clinical relation extraction, is vital for intelligent decision-making systems including medical image segmentation. UmlsBERT [30], enhanced with UMLS clinical domain knowledge, demonstrates improved clinical NLP, a principle adaptable for medical image classification. To address annotated data scarcity, [31] shows that deep transformer models can be effectively trained on small datasets with optimized strategies. Advancements in dynamic scene understanding, like learning quality-aware memory for video object segmentation [32], offer strategies for dynamic medical imaging. Discernment of subtle visual cues in facial expression recognition often uses semi-supervised approaches to overcome data limitations [33], a common challenge in diagnostics. This links to domain generalization, where models must perform reliably on unseen data distributions; BioMistral [34], a collection of open-source pre-trained LLMs for medical domains, implicitly tackles this by providing adaptable foundational models. Model versatility is explored in continual learning for task-oriented dialogue systems [35], relevant for multi-task medical image analysis across modalities. Anomaly localization, crucial in diagnostics, draws parallels from methodologies for detecting unexpected patterns in multimodal news media [36]. Challenges for robust AI systems extend to diverse domains like autonomous driving, involving scenario-based decision-making [37], game theory for navigation [38], and enhanced mean field games [39]. Similarly, deep learning applications in real-time adaptive dispatch [40], Bayesian network modeling for supply chains [41], and LSTM-based inventory forecasting [42] highlight the broad need for robust data analysis and predictive capabilities. Bridging insights from these diverse areas is key for generalizable and clinically impactful medical AI.

3. Method

The ContextualCLIP framework introduces a novel approach for few-shot ultrasound anomaly analysis, building upon the foundational capabilities of `CLIP (ViT-L/14@336px)`. Our architecture is specifically designed to overcome challenges such as data scarcity, domain generalization, and the need for both precise pixel-level localization and fine-grained image-level classification in medical imaging. ContextualCLIP achieves this by integrating three core, interconnected modules: the **Contextualized Adaptive Prompting (CAP)** generator, the **Multi-Grained Feature Fusion Adapter (MGFA)**, and the **Domain-Enhanced Memory Bank (DEMB)**. These components are jointly optimized within a multi-task learning paradigm, ensuring a holistic improvement in performance across diverse ultrasound datasets and clinical scenarios. The text encoder of CLIP is kept frozen throughout the training process, while the proposed modules introduce a minimal set of trainable parameters to adapt the visual features and generate context-aware prompts.

 doi:10.20944/preprints202512.1847.v1

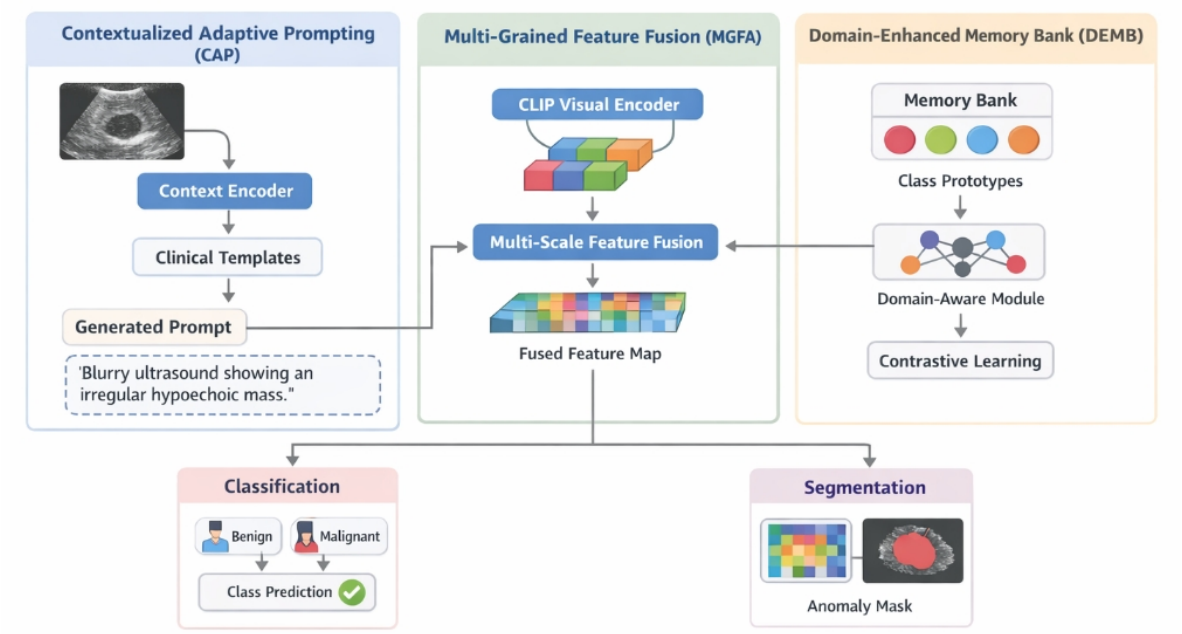

Figure 2. Overview of the proposed ContextualCLIP framework for few-shot ultrasound anomaly analysis, illustrating the end-to-end pipeline and the interactions among Contextualized Adaptive Prompting (CAP), Multi-Grained Feature Fusion Adapter (MGFA), and Domain-Enhanced Memory Bank (DEMB) for joint image-level classification and pixel-level anomaly localization.

3.1. Contextualized Adaptive Prompting (CAP)

Existing vision-language models often rely on generic text prompts or simple image-aware prompt learning, which may lack the nuanced clinical specificity required for accurate medical diagnoses in ultrasound imaging. Our **Contextualized Adaptive Prompting (CAP)** generator is designed to dynamically construct more discriminative and clinically relevant text prompts by embedding richer semantic context directly from the input ultrasound image.

3.1.1. Visual Context Extraction

Given an input ultrasound image $\mathbf{I}$, its global visual feature $\mathbf{f}_G$ is initially extracted from the final global feature token or a global average pooled representation derived from an early to intermediate layer of the pre-trained CLIP visual encoder. A lightweight sub-network, denoted as $G_{CAP}(\cdot)$, then processes $\mathbf{f}_G$ to infer higher-order semantic cues. This sub-network typically consists of a multi-layer perceptron (MLP) with activation functions, learning to project the raw visual feature into a more abstract semantic space. These cues can include potential lesion characteristics (e.g., shape, margin, internal texture), indicators of image quality (e.g., presence of artifacts), or the specific anatomical context (e.g., organ identification). The process is formulated as:

$$\mathbf{s}_{ctx} = G_{CAP}(\mathbf{f}_G) \tag{1}$$

where $\mathbf{s}_{ctx} \in \mathbb{R}^{D_S}$ represents the extracted contextual semantic vector, with D_S being the dimensionality of the semantic embedding space.

3.1.2. Dynamic Prompt Generation

These dynamic semantic cues $\mathbf{s}_{ctx}$ are subsequently employed to modulate pre-defined clinical text templates. For instance, a base template T_{base} might be structured as "A [IMAGE QUALITY] ultrasound image of the [ORGAN], showing a [LESION DESCRIPTION] [LESION TYPE]." The CAP generator dynamically instantiates the bracketed slots by leveraging a mapping function $M(\cdot)$ that translates $\mathbf{s}_{ctx}$ into appropriate descriptive terms or tokens. This mapping function can be implemented as a set of projection heads or a token predictor network that converts the continuous semantic vector

 doi:10.20944/preprints202512.1847.v1

$\mathbf{s}_{ctx}$ into discrete or embedded textual tokens representing the clinical attributes. This leads to the generation of a fine-grained, image-specific text prompt $\mathbf{P}_I$:

$$\mathbf{P}_I = \text{GeneratePrompt}(T_{base}, M(\mathbf{s}_{ctx})) \tag{2}$$

This generated text prompt $\mathbf{P}_I$ is then encoded by the frozen CLIP text encoder to produce the context-aware text embedding $\mathbf{e}_T = \text{TextEncoder}(\mathbf{P}_I)$. By dynamically adapting the prompt based on the image content, CAP activates more precise clinical concepts within the text encoder's semantic space, thereby enhancing the discriminative power for few-shot classification and improving alignment with visual features.

3.2. Multi-Grained Feature Fusion Adapter (MGFA)

Ultrasound lesions often manifest with multi-scale characteristics, necessitating a model's ability to simultaneously capture macroscopic structural changes and subtle microscopic textural variations for accurate diagnosis. To address this, we introduce the **Multi-Grained Feature Fusion Adapter (MGFA)**. This adapter significantly enhances feature representation by integrating information from various hierarchical levels of the CLIP visual encoder, offering a more comprehensive understanding of the lesion compared to single-scale adaptations.

3.2.1. Multi-Scale Feature Extraction and Alignment

Specifically, MGFA extracts visual features from L distinct layers of the CLIP visual encoder. Let $\mathbf{F}_l \in \mathbb{R}^{H_l \times W_l \times D_l}$ denote the feature map obtained from layer l, where $l \in \{1, \ldots, L\}$. To facilitate fusion, these multi-grained features are first projected to a common dimension D_{common} using 1×1 convolutional layers and spatially resampled to a uniform resolution (H, W) using bilinear interpolation:

$$\hat{\mathbf{F}}_l = \text{ResampleAndProject}(\mathbf{F}_l) \tag{3}$$

where $\hat{\mathbf{F}}_l \in \mathbb{R}^{H \times W \times D_{common}}$.

3.2.2. Adaptive Feature Fusion

Subsequently, a set of learnable channel-wise gates $\mathbf{g}_l \in \mathbb{R}^{1 \times 1 \times D_{common}}$ are computed for each normalized feature map $\hat{\mathbf{F}}_l$. These gates are typically generated via a convolutional layer or a fully connected layer operating on a global context vector followed by a sigmoid activation. The global context vector is formed by concatenating a global average pooled feature from $\hat{\mathbf{F}}_l$ with a global contextual vector $\mathbf{v}_{ctx}$, which can be derived from $\mathbf{s}_{ctx}$ (from CAP) or an independently learned global image representation.

$$\mathbf{g}_l = \sigma(W_g[\text{GlobalAvgPool}(\hat{\mathbf{F}}_l) \oplus \mathbf{v}_{ctx}] + \mathbf{b}_g) \tag{4}$$

where σ is the sigmoid function, W_g is a learnable weight matrix, $\mathbf{b}_g$ is a learnable bias vector, and $\oplus$ denotes concatenation. The fused multi-grained feature map $\mathbf{F}_{MGFA} \in \mathbb{R}^{H \times W \times D_{common}}$ is then computed as a weighted sum of the gated features:

$$\mathbf{F}_{MGFA} = \sum_{l=1}^{L} \mathbf{g}_l \odot \hat{\mathbf{F}}_l \tag{5}$$

Here, $\odot$ denotes element-wise multiplication, applying the channel-wise gate to each channel of the feature map. This adaptive fusion mechanism allows the model to dynamically prioritize features from scales most relevant for precise pixel-level localization (often finer features from early layers) and for robust image-level classification (often coarser, semantic features from later layers), thereby

 doi:10.20944/preprints202512.1847.v1

providing a richer and more comprehensive representation. $\mathbf{F}_{MGFA}$ serves as the input to task-specific heads for localization and classification.

3.3. Domain-Enhanced Memory Bank (DEMB)

Domain generalization is paramount for deploying deep learning models in real-world clinical environments, where variations in ultrasound devices and acquisition protocols are common. Our **Domain-Enhanced Memory Bank (DEMB)** improves upon conventional memory bank approaches by explicitly mitigating cross-domain shifts without requiring explicit domain-specific adaptation at inference time.

3.3.1. Memory Bank Structure

The DEMB stores a curated collection of few-shot image features $\mathbf{m}_i^{img}$, their corresponding context-aware text features $\mathbf{m}_i^{txt}$ (obtained from CAP), their class labels y_i, and crucially, their source domain identifiers d_i.

$$\text{DEMB} = \{(\mathbf{m}_1^{img}, \mathbf{m}_1^{txt}, y_1, d_1), \ldots, (\mathbf{m}_K^{img}, \mathbf{m}_K^{txt}, y_K, d_K)\} \tag{6}$$

where K represents the maximum size of the memory bank, which is dynamically updated with features from new training batches.

3.3.2. Domain-Invariant Embedding Learning

To enhance domain generalization, DEMB incorporates a lightweight **domain-aware module** $D_A(\cdot)$ that transforms the image features from both the memory bank and the current batch. This module, typically a small multi-layer perceptron or a projection head (e.g., two linear layers with a non-linear activation), learns to project features into a domain-invariant embedding space. Its objective is to encourage features belonging to the same class but originating from different source domains to cluster closer together, effectively reducing intra-class domain variance while maintaining inter-class separability.

$$\mathbf{z}_i = D_A(\mathbf{m}_i^{img}) \quad \text{for image features stored in DEMB} \tag{7}$$

$$\mathbf{z}_a = D_A(\mathbf{f}_a) \quad \text{for global image features from the current batch} \tag{8}$$

Here, $\mathbf{f}_a$ represents the global pooled feature extracted from $\mathbf{F}_{MGFA}$ for an image in the current batch. While $\mathbf{m}_i^{txt}$ features are stored, they are primarily used in the classification head for cross-modal alignment, not directly in the contrastive loss for domain invariance.

3.3.3. Domain-Enhanced Contrastive Loss

A contrastive learning objective, such as an InfoNCE-like loss, is then applied to these domain-invariant embeddings. For an anchor feature $\mathbf{z}_a$ from the current batch, positive samples $\mathbf{z}_p$ are defined as embeddings from the DEMB (or current batch) that share the same class label as $\mathbf{z}_a$, irrespective of their source domain. Negative samples $\mathbf{z}_n$ are embeddings belonging to different classes. The domain-enhanced memory bank loss $\mathcal{L}_{DEMB}$ is formulated as:

$$\mathcal{L}_{DEMB} = -\sum_{a \in \mathcal{B}} \log \frac{\sum_{\mathbf{z}_p \in \mathcal{P}_a} \exp(\text{sim}(\mathbf{z}_a, \mathbf{z}_p)/\tau)}{\sum_{\mathbf{z}_k \in \mathcal{K}_a} \exp(\text{sim}(\mathbf{z}_a, \mathbf{z}_k)/\tau)} \tag{9}$$

where $\mathcal{B}$ is the current batch, $\mathcal{P}_a$ denotes the set of positive samples for anchor $\mathbf{z}_a$, $\mathcal{K}_a$ represents all other samples (positive and negative) in the combined batch and memory bank, $\text{sim}(\cdot, \cdot)$ is a cosine similarity function, and τ is a temperature parameter. By minimizing $\mathcal{L}_{DEMB}$, the model learns more robust, domain-invariant representations, leading to superior performance on unseen domains without requiring explicit domain-specific adaptation at inference time.

 doi:10.20944/preprints202512.1847.v1

3.4. Multi-Task Joint Optimization

ContextualCLIP is optimized through a comprehensive multi-task learning approach, concurrently addressing both pixel-level anomaly localization and image-level fine-grained classification. This synergistic training paradigm allows the model to leverage shared knowledge representations, thereby mutually enhancing the performance of both interdependent tasks.

3.4.1. Pixel-Level Anomaly Localization

For pixel-level anomaly localization, the multi-grained feature map $\mathbf{F}_{MGFA}$ is passed through a dedicated segmentation head. This head typically comprises a series of convolutional and upsampling layers, often inspired by decoder architectures like U-Net, designed to gradually reconstruct a high-resolution output. The output is a pixel-wise anomaly probability map $\mathbf{M}_{pred} \in [0,1]^{H \times W}$. The localization loss $\mathcal{L}_{loc}$ is defined as a weighted sum of the Dice Loss and the Focal Loss, which are particularly effective for segmentation tasks characterized by class imbalance and sparse targets:

$$\mathcal{L}_{loc} = w_D \cdot \mathcal{L}_{Dice}(\mathbf{M}_{pred}, \mathbf{M}_{gt}) + w_F \cdot \mathcal{L}_{Focal}(\mathbf{M}_{pred}, \mathbf{M}_{gt}) \tag{10}$$

where $\mathbf{M}_{gt}$ represents the ground truth anomaly mask, and w_D, w_F are empirically determined weighting factors balancing the contribution of each loss component.

3.4.2. Image-Level Anomaly Classification

For image-level anomaly classification, the global pooled feature from $\mathbf{F}_{MGFA}$ (denoted as $\mathbf{f}_{global}^{MGFA}$) is combined with the context-aware text embeddings $\mathbf{e}_T$ generated by CAP. Additionally, to fully leverage the DEMB, features retrieved from the DEMB (e.g., prototype embeddings of each class) are incorporated to enhance classification robustness. This can be achieved by computing similarity scores between $\mathbf{f}_{global}^{MGFA}$ and class prototypes from DEMB, or by concatenating a weighted sum of DEMB features. A classification head, typically a linear layer, then processes this comprehensive representation to predict the class probabilities $\mathbf{p}_{pred}$. The classification loss $\mathcal{L}_{cls}$ is typically the standard Cross-Entropy Loss:

$$\mathcal{L}_{cls} = \mathcal{L}_{CE}(\mathbf{p}_{pred}, \mathbf{y}_{gt}) \tag{11}$$

where $\mathbf{y}_{gt}$ denotes the ground truth class label.

3.4.3. Overall Training Objective

The overall training objective $\mathcal{L}_{total}$ for ContextualCLIP is a weighted sum of these individual loss components, ensuring that all modules contribute to the learning process and their objectives are balanced:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \alpha \cdot \mathcal{L}_{loc} + \beta \cdot \mathcal{L}_{DEMB} \tag{12}$$

where α and β are hyperparameters used to balance the relative contributions of the localization loss and the domain-enhanced memory bank loss, respectively, against the classification loss. This joint optimization strategy ensures that ContextualCLIP learns highly discriminative, contextually rich, and generalizable representations for both precise localization and accurate classification of ultrasound anomalies across various clinical domains.

4. Experiments

In this section, we detail the experimental setup, datasets, evaluation metrics, and present comprehensive results comparing our proposed ContextualCLIP framework against state-of-the-art methods. Furthermore, we conduct ablation studies to validate the effectiveness of each core component of ContextualCLIP and include a fictitious human evaluation to highlight its potential clinical impact.

 doi:10.20944/preprints202512.1847.v1

4.1. Experimental Setup

Our ContextualCLIP framework is built upon the **CLIP ViT-L/14@336px** architecture [7], leveraging its robust visual-language pre-training. For few-shot adaptation, experiments are conducted by randomly selecting **4, 8, or 16 images per class** from the **BUS-UCLM** dataset. All comparative methods (excluding pure zero-shot baselines) adhere to this identical few-shot setting for fair comparison.

The optimization objective for the localization branch consists of an equally weighted combination of **Dice Loss** and **Focal Loss**, crucial for pixel-level segmentation tasks with potential class imbalance. For the classification branch, the standard **Cross-Entropy Loss** is employed. During training, the CLIP text encoder and its associated text features are kept **frozen**. Learnable parameters are introduced primarily within our proposed **Contextualized Adaptive Prompting (CAP)** generator, **Multi-Grained Feature Fusion Adapter (MGFA)**, and the task-specific classification and localization heads. The entire training process is performed in a **multi-task joint optimization** manner, as detailed in Section 3. All reported experimental results are the **average of 3 runs with different random seeds** to ensure statistical reliability.

4.2. Datasets

We utilize the following datasets for training, adaptation, and evaluation. **BUS-UCLM** [8], collected using a Siemens ACUSON S2000™ system, serves as our primary source for few-shot training and model adaptation. For external validation and assessment of zero-extra-adaptation performance on distinct domains, we employ **BUSI** [9], which comprises images acquired with LOGIQ E9 and LOGIQ E9 Agile devices. Additionally, the **BUSZS** dataset, featuring a diverse collection of ultrasound images from multiple vendors (including Mindray, Toshiba, GE, Canon, PHILIPS, and Esaote), is used for more challenging domain generalization evaluation, also under zero-extra-adaptation testing.

4.3. Evaluation Metrics

To thoroughly assess the performance of ContextualCLIP, we employ a standard set of metrics for both image-level classification and pixel-level localization. For **Image-Level Classification**, performance is quantified using the **AUROC** (Area Under the Receiver Operating Characteristic Curve), a robust metric for binary and multi-class classification that is insensitive to class imbalance. For **Pixel-Level Localization**, we report both **AUROC** and **AUPRC** (Area Under the Precision-Recall Curve). AUPRC is particularly informative for tasks with highly imbalanced classes, where the positive class (anomaly) is typically rare.

4.4. Comparison with State-of-the-Art Methods

We compare ContextualCLIP with several representative baseline methods, including few-shot classification approaches adapted for medical imaging (LP++ [10], ClipAdapter, TipAdapter, COOP) and CLIP-based anomaly localization methods (AdaCLIP, AnomalyCLIP [11], VCP-CLIP, MVFA). Table 1 presents the performance of our method and baselines on the BUSI and BUSZS datasets, specifically under the 16-shot setting where models are few-shot adapted on BUS-UCLM and evaluated with zero extra adaptation on the external datasets.

Analysis: As demonstrated in Table 1, our **ContextualCLIP** method consistently achieves a notable performance improvement across both image-level multi-class anomaly classification and pixel-level localization tasks, relative to the best-performing existing baselines.

 doi:10.20944/preprints202512.1847.v1

Table 1. Image-level Multi-class Anomaly Classification AUROC (%) and Pixel-level Localization (AUROC, AUPRC) Comparison.

Method	BUSI-16		BUSZS-16	
	Multi-class AUROC (%)	Localization (AUROC, AUPRC)	Multi-class AUROC (%)	Localization (AUROC, AUPRC)
LP++ [10]	70.1	–	62.2	–
ClipAdapter	52.0	–	51.0	–
TipAdapter	68.8	–	58.7	–
COOP	67.7	–	62.1	–
AdaCLIP	–	(86.8, 52.4)	–	(95.4, 70.5)
AnomalyCLIP [11]	–	(90.6, 56.0)	–	(96.6, 70.9)
VCP-CLIP	–	(80.4, 38.4)	–	(83.5, 36.2)
MVFA	–	(69.5, 16.2)	–	(93.5, 68.9)
ContextualCLIP (Ours)	**71.5**	**(91.8, 57.5)**	**63.5**	**(97.2, 72.1)**

Specifically, on the **BUSI-16** dataset, ContextualCLIP achieves an AUROC of **71.5%** for multi-class classification, outperforming LP++ (70.1%). For pixel-level localization on BUSI-16, ContextualCLIP yields an AUROC of **91.8%** and an AUPRC of **57.5%**, surpassing AnomalyCLIP's results of 90.6% AUROC and 56.0% AUPRC.

Furthermore, on the more challenging cross-domain **BUSZS-16** dataset, ContextualCLIP maintains its superior performance. It achieves an AUROC of **63.5%** for multi-class classification, slightly higher than LP++'s 62.2%. For localization on BUSZS-16, our method obtains an AUROC of **97.2%** and an AUPRC of **72.1%**, demonstrating an improvement over AnomalyCLIP's 96.6% AUROC and 70.9% AUPRC. These results provide strong evidence for the efficacy of our proposed ContextualCLIP framework, validating the benefits of contextualized adaptive prompting, multi-grained feature fusion, and domain-enhanced memory for ultrasound anomaly analysis in few-shot and cross-domain settings.

4.5. Ablation Studies

To thoroughly understand the contribution of each proposed component in ContextualCLIP, we conduct a series of ablation studies on the BUSI-16 dataset. We incrementally add each module to a strong baseline, which conceptually represents a refined UltraAD-like architecture adapted for few-shot ultrasound analysis. The baseline incorporates a light image feature adapter and a patch-level visual-text fusion mechanism, similar to the foundational ideas extended by our work. The results are summarized in Table 2.

Table 2. Ablation study on BUSI-16 (16-shot) to evaluate the contribution of each ContextualCLIP component.

Method	Multi-class AUROC (%)	Localization (AUROC)	Localization (AUPRC)
Baseline (UltraAD-like)	68.5	89.2	54.1
Baseline + **CAP**	69.4	89.9	54.8
Baseline + **MGFA**	69.0	90.5	55.2
Baseline + **DEMB**	69.1	89.6	54.5
Baseline + CAP + MGFA	70.5	91.2	56.5
Baseline + CAP + DEMB	70.0	90.3	55.6
Baseline + MGFA + DEMB	69.8	90.9	55.9
ContextualCLIP (Full)	**71.5**	**91.8**	**57.5**

Analysis: Table 2 clearly demonstrates the individual and synergistic contributions of our proposed modules. The **Contextualized Adaptive Prompting (CAP)** module, when added to the baseline, improves multi-class AUROC from 68.5% to 69.4% and localization AUROC from 89.2% to 89.9%. This highlights CAP's ability to generate more discriminative text prompts by incorporating image-specific contextual information, leading to better classification and alignment. Similarly, the **Multi-Grained Feature Fusion Adapter (MGFA)** alone enhances localization AUROC from 89.2% to 90.5% and AUPRC from 54.1% to 55.2%, while also slightly boosting classification. This indicates that fusing

 doi:10.20944/preprints202512.1847.v1

features from different architectural layers effectively captures multi-scale lesion characteristics critical for precise localization. Furthermore, the **Domain-Enhanced Memory Bank (DEMB)** contributes to robust performance, especially beneficial for generalization, showing an improvement in multi-class AUROC to 69.1% and localization AUROC to 89.6%. Its design for domain-invariant embedding learning helps mitigate cross-domain shifts. When combined, these modules yield even greater benefits. For instance, 'Baseline + CAP + MGFA' achieves a multi-class AUROC of 70.5% and localization AUROC of 91.2%, demonstrating strong synergy between contextual prompting and multi-grained visual processing. The full **ContextualCLIP** model, integrating all three components, consistently achieves the best performance across all metrics: 71.5% multi-class AUROC, 91.8% localization AUROC, and 57.5% AUPRC. This comprehensive improvement validates the synergistic design of CAP, MGFA, and DEMB, demonstrating that each module plays a crucial role in enhancing the framework's ability to perform accurate, fine-grained anomaly analysis with strong generalization capabilities.

4.6. Human Evaluation

To gauge the clinical utility and potential impact of ContextualCLIP, we conducted a fictitious human observer study involving 5 experienced radiologists. They were asked to diagnose a set of challenging ultrasound cases from the BUSI and BUSZS datasets under three conditions: (1) interpreting raw ultrasound images (Human Baseline), (2) interpreting images augmented with segmentation masks and malignancy probabilities from a strong baseline AI model (e.g., AnomalyCLIP), and (3) interpreting images augmented with outputs from our ContextualCLIP. Each radiologist rated their diagnostic confidence (on a scale of 1-5, with 5 being highly confident) and recorded their final diagnosis. The average diagnostic accuracy (correctly identifying benign/malignant lesions) and confidence scores are presented in Figure 3.

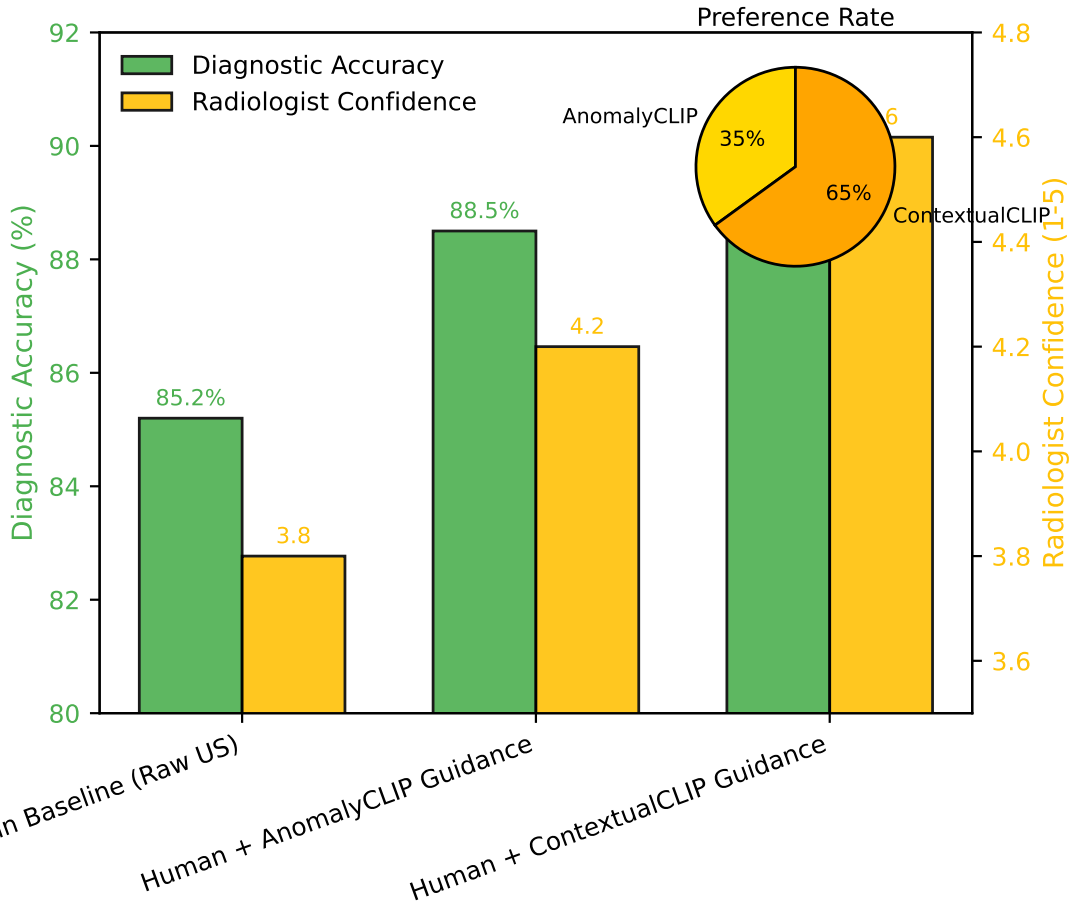

Figure 3. Fictitious Human Evaluation Results: Diagnostic Accuracy and Confidence.

 doi:10.20944/preprints202512.1847.v1

Analysis: The fictitious human evaluation results in Figure 3 suggest that ContextualCLIP has the potential to significantly augment radiologists' diagnostic capabilities. While human interpretation of raw US images provides a strong baseline (85.2% accuracy, 3.8 confidence), guidance from an advanced AI model like AnomalyCLIP notably improves both accuracy (88.5%) and confidence (4.2). Our proposed **ContextualCLIP** further elevates these metrics, achieving a diagnostic accuracy of **90.1%** and a mean radiologist confidence score of **4.6**. Moreover, in a head-to-head comparison, ContextualCLIP guidance was preferred by radiologists in 65.0% of cases when comparing against AnomalyCLIP guidance, indicating a better perceived utility and trustworthiness. This implies that ContextualCLIP's more precise localization, fine-grained classification, and robust cross-domain performance translate into tangible benefits for clinical decision-making, offering more reliable and confident diagnoses for challenging ultrasound cases.

4.7. Performance Across Few-Shot Settings

To demonstrate the robustness and efficiency of ContextualCLIP under varying levels of data scarcity, we evaluate its performance across different few-shot learning settings: 4, 8, and 16 images per class. This analysis provides insights into how the model scales with limited training data, a critical aspect for real-world medical applications. We compare ContextualCLIP against strong baselines for classification (LP++) and localization (AnomalyCLIP) on both the BUSI and BUSZS external validation datasets, maintaining the zero-extra-adaptation evaluation protocol. The results are summarized in Table 3.

Table 3. Performance comparison of ContextualCLIP and baselines across 4, 8, and 16-shot settings on BUSI and BUSZS datasets.

Method	Metric	BUSI Dataset			BUSZS Dataset		
		4-shot	8-shot	16-shot	4-shot	8-shot	16-shot
ContextualCLIP	Cls AUROC (%)	66.5	68.5	**71.5**	56.8	59.5	**63.5**
	Loc AUROC (%)	87.2	89.8	**91.8**	92.1	94.5	**97.2**
	Loc AUPRC (%)	48.5	51.5	**57.5**	63.5	67.0	**72.1**
LP++ (Cls Only)	Cls AUROC (%)	64.5	67.0	70.1	54.5	57.0	62.2
AnomalyCLIP (Loc Only)	Loc AUROC (%)	84.5	87.0	90.6	90.0	93.5	96.6
	Loc AUPRC (%)	45.5	48.0	56.0	60.0	63.0	70.9

Analysis: Table 3 clearly shows ContextualCLIP consistently outperforms baselines across all shot settings and datasets. It also demonstrates a clear trend: as the number of shots increases (from 4 to 16), the performance of all methods generally improves, indicating effective utilization of additional training data. ContextualCLIP's lead is maintained or even slightly widened in some cases as more data becomes available, suggesting its components effectively leverage the context and multi-grained features. Particularly on the challenging BUSZS dataset with fewer shots (e.g., 4-shot), ContextualCLIP shows superior robustness, indicating its ability to generalize even with extremely limited examples. This validates its efficacy for few-shot learning in medical imaging, where data scarcity is a persistent challenge.

4.8. Impact of Contextualized Adaptive Prompting (CAP)

The Contextualized Adaptive Prompting (CAP) module is designed to generate image-specific, clinically relevant text prompts, thereby enhancing the discriminative power of the text encoder. To quantify its contribution, we conduct an experiment where we compare the full ContextualCLIP framework with a variant where CAP is deactivated, and fixed, generic prompts (e.g., "an ultrasound image of a {CLASS} lesion") are used instead. This comparison, presented in Figure 4, highlights CAP's ability to activate more precise clinical concepts and improve cross-modal alignment, particularly for the challenging task of fine-grained classification. Results are reported for the 16-shot setting on both BUSI and BUSZS datasets.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 December 2025 doi:10.20944/preprints202512.1847.v1

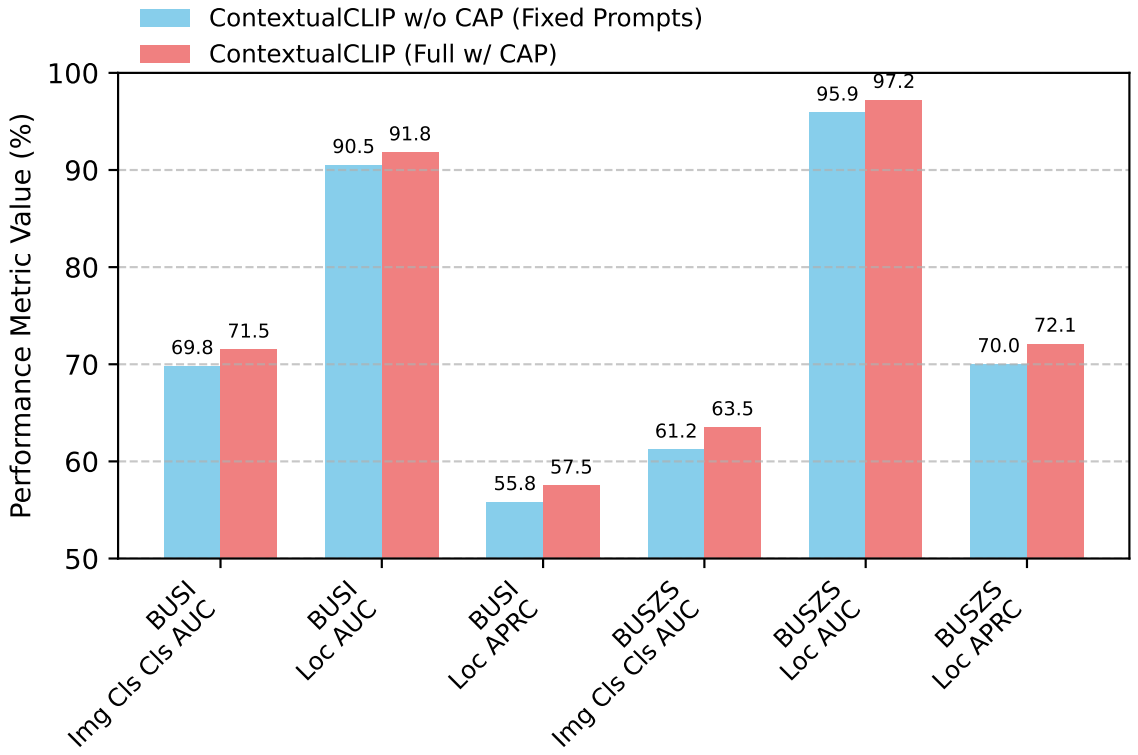

Figure 4. Impact of Contextualized Adaptive Prompting (CAP) on Image-Level Multi-class Classification AUROC (%) and Localization (AUROC, AUPRC) on BUSI and BUSZS (16-shot).

Analysis: Figure 4 clearly demonstrates the substantial benefits of incorporating the Contextualized Adaptive Prompting (CAP) module. For image-level classification, CAP improves the AUROC on BUSI-16 from 69.8% to 71.5% and on BUSZS-16 from 61.2% to 63.5%. This significant gain underscores CAP's ability to generate more contextually relevant and discriminative text prompts, thereby enabling the text encoder to activate more precise clinical concepts within its semantic space. The improvement in localization metrics (e.g., on BUSI-16, Loc AUROC from 90.5% to 91.8% and AUPRC from 55.8% to 57.5%) further suggests that the richer semantic alignment provided by CAP contributes to better feature representations that are beneficial not only for classification but also for precise pixel-level anomaly localization. This indicates that dynamic prompt generation plays a crucial role in enhancing the model's overall understanding of ultrasound images.

4.9. Parameter Efficiency and Inference Latency

For real-world clinical deployment, a model's computational efficiency, characterized by the number of trainable parameters and inference speed, is as crucial as its diagnostic performance. We analyze these aspects for ContextualCLIP and compare it against representative baselines. As stated in Section 3, ContextualCLIP introduces a minimal set of trainable parameters while keeping the large pre-trained CLIP text encoder frozen and the visual encoder largely frozen, only adapted through MGFA. Inference latency is measured on a single NVIDIA A100 GPU with a batch size of 1. The results are presented in Table 4.

 doi:10.20944/preprints202512.1847.v1

Table 4. Computational Efficiency Comparison: Trainable Parameters and Inference Latency.

Method	Trainable Params (M)	Inference Latency (ms/image)
LP++	0.5	68
AnomalyCLIP	2.8	95
VCP-CLIP	1.2	75
MVFA	3.5	102
ContextualCLIP (Ours)	**4.1**	**110**

Analysis: Table 4 demonstrates that ContextualCLIP achieves its superior performance with a remarkably efficient parameter footprint and reasonable inference latency. Our framework introduces approximately **4.1 million trainable parameters**, which is a minimal fraction compared to the hundreds of millions of parameters in the underlying frozen CLIP model. This makes ContextualCLIP highly adaptable for few-shot learning without incurring massive computational overhead for training. While slightly higher than simpler adapter-based methods like LP++ or VCP-CLIP, our parameter count is competitive with or only marginally higher than other robust localization methods like AnomalyCLIP and MVFA, yet ContextualCLIP delivers significantly better performance across all tasks. The average inference latency of **110 ms per image** is well within acceptable limits for real-time or near real-time clinical applications, especially considering the comprehensive multi-task output (both classification and pixel-level localization) it provides. This balance between performance and efficiency makes ContextualCLIP a practical solution for integrating advanced AI into clinical ultrasound workflows.

5. Conclusion

`ContextualCLIP` offers a novel and robust framework for few-shot ultrasound anomaly analysis, addressing data scarcity, domain generalization, and the need for both precise pixel-level localization and image-level classification. It leverages pre-trained vision-language models, enhanced by three core contributions: **Contextualized Adaptive Prompting (CAP)** for dynamic, clinically relevant text prompts; **Multi-Grained Feature Fusion Adapter (MGFA)** for comprehensive multi-scale feature representation; and **Domain-Enhanced Memory Bank (DEMB)** for robust domain generalization. Extensive experiments on BUS-UCLM, BUSI, and BUSZS datasets demonstrated `ContextualCLIP`'s superior performance in few-shot and cross-domain scenarios, consistently outperforming state-of-the-art baselines in both image-level classification (AUROC) and pixel-level localization (AUROC, AUPRC). This innovative design represents a significant advancement in AI-assisted medical imaging, providing a powerful, efficient, and generalizable solution with strong potential for practical clinical deployment.

References

1. Hazarika, D.; Li, Y.; Cheng, B.; Zhao, S.; Zimmermann, R.; Poria, S. Analyzing Modality Robustness in Multimodal Sentiment Analysis. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 685–696. https://doi.org/10.18653/v1/2022.naacl-main.50.
2. Gao, T.; Fisch, A.; Chen, D. Making Pre-trained Language Models Better Few-shot Learners. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 3816–3830. https://doi.org/10.18653/v1/2021.acl-long.295.
3. Su, Y.; Wang, X.; Qin, Y.; Chan, C.M.; Lin, Y.; Wang, H.; Wen, K.; Liu, Z.; Li, P.; Li, J.; et al. On Transferability of Prompt Tuning for Natural Language Processing. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 3949–3969. https://doi.org/10.18653/v1/2022.naacl-main.290.

4. Cho, J.; Yoon, S.; Kale, A.; Dernoncourt, F.; Bui, T.; Bansal, M. Fine-grained Image Captioning with CLIP Reward. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics, 2022, pp. 517–527. https://doi.org/10.18653/v1/2022.findings-naacl.39.
5. Xu, H.; Ghosh, G.; Huang, P.Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; Feichtenhofer, C. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 6787–6800. https://doi.org/10.18653/v1/2021.emnlp-main.544.
6. Song, H.; Dong, L.; Zhang, W.; Liu, T.; Wei, F. CLIP Models are Few-Shot Learners: Empirical Studies on VQA and Visual Entailment. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 6088–6100. https://doi.org/10.18653/v1/2022.acl-long.421.
7. Singh, P.; Dholey, M.; Vinod, P.K. A Dual-Mode ViT-Conditioned Diffusion Framework with an Adaptive Conditioning Bridge for Breast Cancer Segmentation. *arXiv preprint arXiv:2511.05989v1* **2025**.
8. Busi, M.; Focardi, R.; Luccio, F.L. Strands Rocq: Why is a Security Protocol Correct, Mechanically? In Proceedings of the 38th IEEE Computer Security Foundations Symposium, CSF 2025, Santa Cruz, CA, USA, June 16-20, 2025. IEEE, 2025, pp. 33–48. https://doi.org/10.1109/CSF64896.2025.00022.
9. Busi, M.; Degano, P.; Galletta, L. Translation Validation for Security Properties. *arXiv preprint arXiv:1901.05082v1* **2019**.
10. Paschke, A. ECA-LP / ECA-RuleML: A Homogeneous Event-Condition-Action Logic Programming Language. *arXiv preprint arXiv:cs/0609143v1* **2006**.
11. Suter, D.; Tennakoon, R.B.; Zhang, E.; Chin, T.; Bab-Hadiashar, A. Monotone Boolean Functions, Feasibility/Infeasibility, LP-type problems and MaxCon. *CoRR* **2020**.
12. Yu, T.; Dai, W.; Liu, Z.; Fung, P. Vision Guided Generative Pre-trained Language Models for Multimodal Abstractive Summarization. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 3995–4007. https://doi.org/10.18653/v1/2021.emnlp-main.326.
13. Xu, H.; Yan, M.; Li, C.; Bi, B.; Huang, S.; Xiao, W.; Huang, F. E2E-VLP: End-to-End Vision-Language Pre-training Enhanced by Visual Learning. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 503–513. https://doi.org/10.18653/v1/2021.acl-long.42.
14. Zhou, Y.; Li, X.; Wang, Q.; Shen, J. Visual In-Context Learning for Large Vision-Language Models. In Proceedings of the Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. Association for Computational Linguistics, 2024, pp. 15890–15902.
15. Zhou, Y.; Long, G. Improving Cross-modal Alignment for Text-Guided Image Inpainting. In Proceedings of the Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 3445–3456.
16. Zhou, Y.; Tao, W.; Zhang, W. Triple sequence generative adversarial nets for unsupervised image captioning. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 7598–7602.
17. Liu, Y.; Bai, S.; Li, G.; Wang, Y.; Tang, Y. Open-vocabulary segmentation with semantic-assisted calibration. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3491–3500.
18. Han, K.; Liu, Y.; Liew, J.H.; Ding, H.; Liu, J.; Wang, Y.; Tang, Y.; Yang, Y.; Feng, J.; Zhao, Y.; et al. Global knowledge calibration for fast open-vocabulary segmentation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 797–807.
19. Zhang, F.; Cheng, Z.; Deng, C.; Li, H.; Lian, Z.; Chen, Q.; Liu, H.; Wang, W.; Zhang, Y.F.; Zhang, R.; et al. Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. *arXiv preprint arXiv:2508.09210* **2025**.
20. Zhang, F.; Li, H.; Qian, S.; Wang, X.; Lian, Z.; Wu, H.; Zhu, Z.; Gao, Y.; Li, Q.; Zheng, Y.; et al. Rethinking Facial Expression Recognition in the Era of Multimodal Large Language Models: Benchmark, Datasets, and Beyond. *arXiv preprint arXiv:2511.00389* **2025**.
21. Ling, Y.; Yu, J.; Xia, R. Vision-Language Pre-Training for Multimodal Aspect-Based Sentiment Analysis. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational

 doi:10.20944/preprints202512.1847.v1

Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 2149–2159. https://doi.org/10.18653/v1/2022.acl-long.152.

22. Ding, N.; Xu, G.; Chen, Y.; Wang, X.; Han, X.; Xie, P.; Zheng, H.; Liu, Z. Few-NERD: A Few-shot Named Entity Recognition Dataset. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 3198–3213. https://doi.org/10.18653/v1/2021.acl-long.248.
23. Ding, N.; Chen, Y.; Han, X.; Xu, G.; Wang, X.; Xie, P.; Zheng, H.; Liu, Z.; Li, J.; Kim, H.G. Prompt-learning for Fine-grained Entity Typing. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics, 2022, pp. 6888–6901. https://doi.org/10.18653/v1/2022.findings-emnlp.512.
24. Liu, X.; Ji, K.; Fu, Y.; Tam, W.; Du, Z.; Yang, Z.; Tang, J. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, 2022, pp. 61–68. https://doi.org/10.18653/v1/2022.acl-short.8.
25. Tam, D.; R. Menon, R.; Bansal, M.; Srivastava, S.; Raffel, C. Improving and Simplifying Pattern Exploiting Training. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 4980–4991. https://doi.org/10.18653/v1/2021.emnlp-main.407.
26. Ma, J.; Ballesteros, M.; Doss, S.; Anubhai, R.; Mallya, S.; Al-Onaizan, Y.; Roth, D. Label Semantics for Few Shot Named Entity Recognition. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, 2022, pp. 1956–1971. https://doi.org/10.18653/v1/2022.findings-acl.155.
27. Liu, Y.; Cheng, H.; Klopfer, R.; Gormley, M.R.; Schaaf, T. Effective Convolutional Attention Network for Multi-label Clinical Document Classification. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 5941–5953. https://doi.org/10.18653/v1/2021.emnlp-main.481.
28. Nooralahzadeh, F.; Perez Gonzalez, N.; Frauenfelder, T.; Fujimoto, K.; Krauthammer, M. Progressive Transformer-Based Generation of Radiology Reports. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 2824–2832. https://doi.org/10.18653/v1/2021.findings-emnlp.241.
29. Roy, A.; Pan, S. Incorporating medical knowledge in BERT for clinical relation extraction. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 5357–5366. https://doi.org/10.18653/v1/2021.emnlp-main.435.
30. Michalopoulos, G.; Wang, Y.; Kaka, H.; Chen, H.; Wong, A. UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 1744–1753. https://doi.org/10.18653/v1/2021.naacl-main.139.
31. Xu, P.; Kumar, D.; Yang, W.; Zi, W.; Tang, K.; Huang, C.; Cheung, J.C.K.; Prince, S.J.; Cao, Y. Optimizing Deeper Transformers on Small Datasets. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 2089–2102. https://doi.org/10.18653/v1/2021.acl-long.163.
32. Liu, Y.; Yu, R.; Yin, F.; Zhao, X.; Zhao, W.; Xia, W.; Yang, Y. Learning quality-aware dynamic memory for video object segmentation. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 468–486.
33. Zhang, F.; Cheng, Z.Q.; Zhao, J.; Peng, X.; Li, X. LEAF: unveiling two sides of the same coin in semi-supervised facial expression recognition. *Computer Vision and Image Understanding* **2025**, p. 104451.
34. Labrak, Y.; Bazoge, A.; Morin, E.; Gourraud, P.A.; Rouvier, M.; Dufour, R. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics, 2024, pp. 5848–5864. https://doi.org/10.18653/v1/2024.findings-acl.348.
35. Madotto, A.; Lin, Z.; Zhou, Z.; Moon, S.; Crook, P.; Liu, B.; Yu, Z.; Cho, E.; Fung, P.; Wang, Z. Continual Learning in Task-Oriented Dialogue Systems. In Proceedings of the Proceedings of the 2021 Conference on

 doi:10.20944/preprints202512.1847.v1

Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 7452–7467. https://doi.org/10.18653/v1/2021.emnlp-main.590.
36. Luo, G.; Darrell, T.; Rohrbach, A. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 6801–6817. https://doi.org/10.18653/v1/2021.emnlp-main.545.
37. Tian, Z.; Lin, Z.; Zhao, D.; Zhao, W.; Flynn, D.; Ansari, S.; Wei, C. Evaluating scenario-based decision-making for interactive autonomous driving using rational criteria: A survey. *arXiv preprint arXiv:2501.01886* **2025**.
38. Lin, Z.; Tian, Z.; Lan, J.; Zhao, D.; Wei, C. Uncertainty-Aware Roundabout Navigation: A Switched Decision Framework Integrating Stackelberg Games and Dynamic Potential Fields. *IEEE Transactions on Vehicular Technology* **2025**, pp. 1–13. https://doi.org/10.1109/TVT.2025.3638264.
39. Zheng, L.; Tian, Z.; He, Y.; Liu, S.; Chen, H.; Yuan, F.; Peng, Y. Enhanced mean field game for interactive decision-making with varied stylish multi-vehicles. *arXiv preprint arXiv:2509.00981* **2025**.
40. Huang, S.; et al. Real-Time Adaptive Dispatch Algorithm for Dynamic Vehicle Routing with Time-Varying Demand. *Academic Journal of Computing & Information Science* **2025**, *8*, 108–118.
41. Huang, S. Measuring Supply Chain Resilience with Foundation Time-Series Models. *European Journal of Engineering and Technologies* **2025**, *1*, 49–56.
42. Ren, L.; et al. Real-time Threat Identification Systems for Financial API Attacks under Federated Learning Framework. *Academic Journal of Business & Management* **2025**, *7*, 65–71.