Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Training-Free Retrieval-Augmented Medical Vision-Language Inference Across Multiple Imaging Domains

Shahd M. Noman

sh.fekry@nu.edu.eg

Nile University

Mohammad Yaqub

Mohamed bin Zayed University of Artificial Intelligence

Mustafa Elattar

Nile University

Research Article

Keywords: Medical vision--language models, Zero-shot learning, Retrieval-augmented inference, Medical image classification, Foundation models, Test-time augmentation

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11118061/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Training-Free Retrieval-Augmented Medical Vision-Language Inference Across Multiple Imaging Domains

Shahd M. Noman[1,2*], Mohammad Yaqub[3] and Mustafa Elattar[1,2]

[1*]Medical Imaging and Image Processing Research Group, Center for Informatics Science, Nile University, Giza, Egypt.
[2]School of Information Technology and Computer Science, Nile University, Giza, Egypt.
[3]Department of Computer Vision, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates.

*Corresponding author(s). E-mail(s): sh.fekry@nu.edu.eg;

Abstract

Background: Medical vision–language models enable zero-shot image classification without task-specific parameter optimization, but their predictions can be sensitive to visual variation and rely primarily on image–text similarity without using available domain-specific reference cases. We investigated whether inference-time visual stabilization and retrieval from labeled support examples could improve the performance of a frozen medical vision–language model across heterogeneous medical imaging tasks.

Methods: We developed a training-free inference framework combining multi-scale test-time augmentation with retrieval-augmented prediction. Multiple centrally cropped views of each image were encoded using a frozen MedSigLIP model and averaged to obtain a robust visual representation. The same representation was used to retrieve visually similar examples from a labeled training-set gallery. A task-adaptive retrieval prior was constructed using temperature-scaled class voting for multi-class tasks and independent per-label evidence aggregation for multi-label tasks. Retrieval evidence was subsequently combined with prompt-based zero-shot logits using calibrated log-probability fusion. Fusion hyperparameters were selected using masked self-retrieval on the support set without access to the held-out test data. The framework was evaluated on NCT-CRC-HE-100K/CRC-VAL-HE-7K for histopathology, ISIC 2018 for dermoscopy, and NIH ChestX-ray14 for chest radiography.

Results: Retrieval augmentation consistently improved performance across all three imaging domains. On NCT-CRC, accuracy increased from 50.46% with standard zero-shot inference to 94.57%. On ISIC 2018, accuracy increased from 15.25% to 54.24%. On NIH ChestX-ray14, macro-AUC increased from 68.97% to 77.55%. Improvements were statistically significant according to McNemar's test for the multi-class tasks and DeLong's test for the multi-label task. Ablation experiments showed that retrieval provided the dominant improvement, whereas the contribution of test-time augmentation varied across domains. Support-gallery experiments further demonstrated substantial differences in the amount of labeled reference data required across imaging modalities.
Conclusion: Retrieval from domain-matched reference examples can substantially improve frozen medical vision–language inference without task-specific parameter updates. The results indicate that non-parametric support memories can provide an effective inference-time adaptation mechanism across both single-label and multi-label medical imaging tasks, although their effectiveness depends on the coverage and representativeness of the available support gallery.

Keywords: Medical vision–language models, Zero-shot learning, Retrieval-augmented inference, Medical image classification, Foundation models, Test-time augmentation

1 Introduction

Artificial Intelligence (AI) has become an increasingly important tool in medical image analysis, where deep learning systems have demonstrated strong performance across a wide range of diagnostic tasks, including diabetic retinopathy, classifying skin lesions, and identifying metastases in histopathology slides [1–3]. These systems have shown the potential to support clinical decision-making, improve diagnostic consistency, and reduce the workload of specialists in data-intensive settings. However, most of these advances have been achieved in fully supervised settings, where success depends heavily on the availability of large, carefully annotated datasets. In medicine, constructing such datasets is costly, time-consuming, and often impractical, particularly for rare diseases, newly emerging diagnostic categories, or institutions with limited annotation resources [4]. To reduce reliance on task-specific labels, recent research has increasingly shifted toward foundation models and vision-language models (VLMs), which offer a more flexible alternative to conventional supervised pipelines. Through contrastive pre-training on large collections of image-text pairs, models such as CLIP [5] and medical adaptations PLIP [6], and MedSigLIP [7] learn a shared embedding space in which visual content can be directly aligned with natural language descriptions. This makes zero-shot classification possible, allowing a model to infer target labels from text prompts without requiring explicit training on the downstream dataset itself. This paradigm is especially well suited to medicine, where expert-labeled data are often scarce, but clinically meaningful descriptions of diseases, tissues, and imaging findings are typically available [8]. However, despite the capabilities of these foundation models, their direct application in clinical workflows is hindered by two fundamental challenges: instability under test-time shifts and lack of contextual grounding. First,

standard zero-shot models are surprisingly sensitive to small changes in the image. Unlike supervised models that are trained to learn invariance through extensive augmentation during training, a frozen zero-shot model might change its diagnosis if the image undergoes minor shifts in scale, cropping, or orientation [9]. Second, and perhaps more critically, the standard paradigm relies exclusively on abstract textual descriptions (e.g., "a photo of melanoma") to classify complex pathologies. This approach effectively ignores the wealth of information available in labeled support sets or historical hospital databases, forcing the model to depend exclusively on its pre-trained memory rather than grounding its decision in concrete, domain-specific examples [10]. Domain-specific vision-language models (VLMs) have enabled zero-shot medical inference by learning a shared embedding space for images and clinical text. In histopathology, models such as CPLIP [11] and CONCH [12] have demonstrated strong zero-shot performance across classification and retrieval tasks. More recent approaches, including PathologyVLM [13], further extended these capabilities to applications such as visual question answering and tissue-level recognition. In parallel, post-hoc strategies such as multi-resolution prompting and cosine calibration have shown that zero-shot performance can be substantially improved even without model fine-tuning [14]. In radiology, models such as CXR-CLIP [15], and LLaVA-RadZ [16] have advanced zero-shot abnormality detection in chest X-rays, while in dermatology, PanDerm [17] demonstrated the effectiveness of large-scale multimodal pre-training across a broad range of skin imaging tasks. To address the domain shift often encountered in zero-shot inference, adaptation methods have evolved from optimization-heavy procedures toward increasingly lightweight, and in some cases training-free, interventions. Prompt learning remains a prominent strategy, with methods such as CoOp and CoCoOp [18] learning context vectors, while later variants including MetaPrompt [19] aim to improve domain robustness. Another line of work introduces lightweight adapters to refine pre-trained features such as CLIP-Adapter [10]. More recently, training-free and post-hoc approaches such as LFA [20] have shown that zero-shot performance can be improved without updating model weights. The idea of using a retrieval prior derived from visually similar examples has become an important inference-time strategy for improving classification. In generic visual recognition, methods such as Retrieval-Augmented Classification (RAC) [21] use a non-parametric memory bank to estimate a neighbor-based label distribution, which is then fused with the base classifier's logits. In medicine, the same principle underlies Content-Based Image Retrieval (CBIR) systems [22], where similar historical cases are used to support diagnosis. However, most existing formulations treat retrieval as a single-label prior, which is not well suited to multi-label settings where multiple findings may co-occur, as in chest radiography. Test-Time Augmentation (TTA) is a widely used strategy for improving robustness by aggregating predictions across multiple transformed views of the same input [23]. In medical imaging, it has been applied in supervised tasks such as segmentation and lesion classification to reduce prediction noise and improve stability under acquisition-related variability [24]. More recently, TTA has also been explored in zero-shot Vision-Language Models (VLMs), where methods such as MTA [25] use augmented test views to better align frozen models with the target distribution. These works suggest that multi-view inference can serve not only as a robustness mechanism,

but also as a lightweight form of test-time adaptation under domain shift. To address these challenges without the computational burden of fine-tuning, we propose a unified, training-free framework that acts upstream during the inference process. This work represents a strategic evolution of our previous research [14], shifting the focus from simply calibrating the model's final outputs (post-hoc) to enriching the visual representation itself. Our approach combines two powerful mechanisms: Test-Time Augmentation (TTA) and Retrieval-Augmented Inference. First, we employ multi-scale TTA to generate multiple views of the input image at varying scales, allowing the model to form a "visual consensus" that smooths out noise and reduces sensitivity to size or crop artifacts. Second, we bridge the context gap by employing a k-Nearest Neighbor (KNN) retrieval mechanism. This mechanism uses the available training data as a dynamic "memory bank," allowing the model to identify and reference similar historical cases before making a decision. This process mimics clinical practice, analogous to a pathologist comparing a difficult sample against a reference atlas, thereby injecting concrete, domain-specific evidence directly into the zero-shot pipeline.
Throughout this work, *zero-shot* refers to the frozen prompt-based VLM prediction branch, which performs no task-specific parameter optimization. The complete retrieval-augmented framework additionally uses labeled reference examples from the target-domain support gallery at inference time. We therefore describe the proposed system as *training-free*, rather than label-free.
The main contributions of this work are

- We propose a unified, training-free framework that improves zero-shot medical vision-language inference through multi-scale test-time augmentation and retrieval-augmented fusion.
- We introduce a task-adaptive retrieval prior that supports both multi-class and multi-label prediction.
- We demonstrate consistent and statistically significant improvements across histopathology, dermoscopy, and radiology benchmarks.

2 Methods

We propose a unified, training-free inference framework that improves zero-shot medical vision-language prediction by stabilizing the visual representation and incorporating retrieval-grounded evidence. As shown in Fig. 1, the framework operates entirely at inference time and requires no updates to model weights.

The core of our framework is the **MedSigLIP** foundation model [7], denoted as $\mathcal{M}$. MedSigLIP consists of two parallel encoders: an image encoder E_I that maps visual inputs to a d-dimensional embedding space and a text encoder E_T that maps clinical text prompts to the same shared space. In a standard zero-shot setting, given a test image x and a set of candidate class descriptions $\mathcal{C} = \{t_1, \ldots, t_M\}$, the prediction is based solely on the cosine similarity between the image embedding and the text embeddings. However, this standard inference path suffers from visual instability and a lack of access to domain-specific historical data. We extend this process with two components: (1) multi-scale visual ensembling, which produces a more stable query embedding using test-time augmentation, and (2) retrieval-aware logit fusion, which

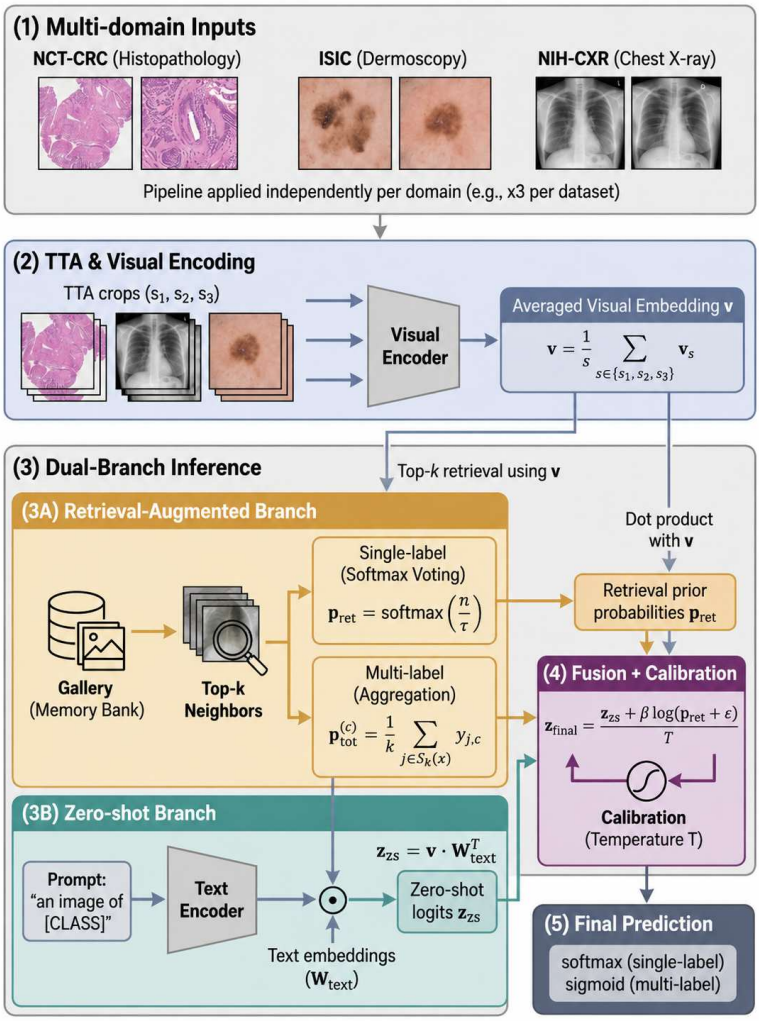


Fig. 1 Overview of the proposed training-free retrieval-augmented inference pipeline. A query image is processed using multi-scale test-time augmentation (TTA) and encoded to obtain an averaged visual embedding $\mathbf{v}$. This same embedding $\mathbf{v}$ is used to query the training gallery, retrieving the Top-k nearest neighbors and inducing a task-adaptive retrieval prior (softmax voting for single-label tasks; independent aggregation for multi-label tasks). In parallel, a zero-shot branch computes logits via image-text similarity using prompt embeddings. Finally, the prior retrieval and zero-shot logits are fused at the logit level to produce the final prediction.

incorporates evidence from similar labeled training examples before making the final prediction.

2.1 Multi-Scale Visual Ensembling

Standard zero-shot inference typically relies on a single forward pass of the input image. In medical imaging, this approach can be fragile: small changes in tissue scale,

scanner resolution, or acquisition artifacts can lead to inconsistent feature representation. To reduce this sensitivity and obtain a more stable visual representation, we use test-time augmentation (TTA) with multi-scale ensembling, combining embeddings from several resized views of the same image into a single, robust embedding. For a given test image x, we generate a set of views $\mathcal{V} = \{v_1, \ldots, v_N\}$ by applying central cropping at varying scales. Specifically, we define a set of scale factors $\mathcal{S} = \{1.0, 0.9, 0.8\}$, where a scale of $s = 1.0$ corresponds to the original image (no crop), while $s = 0.9$ and $s = 0.8$ correspond to crops that retain the central 90% and 80% of the image *width and height*, respectively (i.e., approximately $0.9^2 = 81\%$ and $0.8^2 = 64\%$ of the image area). All views are resized to the native input resolution of MedSigLIP (448×448 pixels). Each view v_s (corresponding to scale $s \in \mathcal{S}$) is independently processed by the image encoder E_I to produce a normalized feature vector:

$$\mathbf{f}_s = \frac{E_I(v_s)}{\|E_I(v_s)\|_2}. \tag{1}$$

To form the final robust representation, we compute the centroid of these multi-scale features. This aggregation creates a "visual consensus," reducing the impact of outliers or scale-specific noise. The final robust embedding $\mathbf{q}_{img}$ is obtained by averaging the view embeddings and re-normalizing to the unit hypersphere:

$$\mathbf{q}_{img} = \text{norm}\left(\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} \mathbf{f}_s\right), \tag{2}$$

where $\text{norm}(\mathbf{z}) = \mathbf{z}/\|\mathbf{z}\|_2$. This resulting vector $\mathbf{q}_{img}$ serves as the query for both the zero-shot classifier and the retrieval module.

2.2 Retrieval-Augmented Inference

To overcome the "contextual blindness" of zero-shot models, where predictions are made in isolation from historical data, we leverage the available training set $\mathcal{D}_{train}$ to construct a non-parametric memory bank (or gallery). This allows the model to ground its predictions in concrete, domain-specific examples.

2.2.1 Gallery construction and retrieval

We use the labeled training split as a lightweight, non-parametric memory that provides domain-matched reference evidence at inference time. Formally, let $\mathcal{D}_{\text{train}} = \{(x_i, \mathbf{y}_i)\}_{i=1}^{L}$, where x_i is a training image and $\mathbf{y}_i$ is its label (a single class index for multi-class tasks, or a binary vector for multi-label tasks). During an offline indexing phase, we embed each training image with the *same* multi-scale TTA embedding function used at test time (Section 2.1). This ensures that both queries and references live in the same feature space and benefit from the same robustness gains of augmentation-based ensembling.

We store each embedded training instance as a key-value pair, where the key is the TTA embedding and the value is the label:

$$\mathbf{k}_i = f_{\text{TTA}}(x_i), \qquad \mathcal{G} = \{(\mathbf{k}_i, \mathbf{y}_i)\}_{i=1}^{L}. \tag{3}$$

Here, $\mathcal{G}$ denotes the *gallery* of L labeled reference embeddings. At inference, a test image x is embedded identically to produce a query vector $\mathbf{v} = f_{\mathrm{TTA}}(x)$.
We then retrieve the k most similar gallery entries (where k is the number of nearest neighbors) using cosine similarity:

$$\begin{aligned} \mathcal{I}_k(x) &= \mathrm{TopK}_i(\cos(\mathbf{v}, \mathbf{k}_i)) \,, \\ \mathcal{N}_k(x) &= \{(\mathbf{k}_i, \mathbf{y}_i) \ : \ i \in \mathcal{I}_k(x)\} \,. \end{aligned} \tag{4}$$

Intuitively, $\mathcal{I}_k(x)$ is the index set of the Top-k nearest neighbors of x in the gallery, and $\mathcal{N}_k(x)$ is the corresponding retrieved neighbor set (their embeddings and labels). In the next step, we convert this neighborhood into a task-specific *retrieval prior* over labels, which acts as an additional evidence source that complements the prompt-based zero-shot prediction. All gallery embeddings are computed once offline; test-time computation consists of a single query embedding followed by a Top-k search.

2.2.2 Task-adaptive prior formulation

The retrieved neighborhood $\mathcal{N}_k(x)$ provides local, example-based evidence about the likely label(s) of the query. We summarize this evidence as a *retrieval prior* $\mathbf{p}_{\mathrm{ret}}(x) \in \mathbb{R}^M$, where M is the number of labels. A key consideration is that medical benchmarks may follow either a *single-label* (mutually exclusive classes) or a *multi-label* (co-occurring findings) formulation. Since these settings require different probability structures, we define $\mathbf{p}_{\mathrm{ret}}$ using two aggregation rules.

Case A: Single-label (multi-class) prior.

For mutually exclusive classes $c \in \{1, \ldots, M\}$, we first count the number of retrieved neighbors assigned to each class:

$$n_c(x) = \sum_{(\mathbf{k}_j, y_j) \in \mathcal{N}_k(x)} \mathbb{1}(y_j = c). \tag{5}$$

We then convert these counts into a categorical distribution using a temperature-scaled softmax:

$$p_{\mathrm{ret}}^{(c)}(x) = \frac{\exp\big(n_c(x)/\tau\big)}{\sum_{m=1}^{M} \exp\big(n_m(x)/\tau\big)}. \tag{6}$$

Here, $\tau > 0$ controls the sharpness of the prior: smaller values yield a more peaked distribution when neighbors agree, while larger values preserve uncertainty when the neighborhood contains mixed labels.

Case B: Multi-label (independent) prior.

For multi-label problems, labels are not mutually exclusive and can co-occur. A softmax normalization would incorrectly enforce competition between labels. Instead, we estimate each label probability independently by averaging binary indicators across the retrieved neighborhood. Each neighbor has a binary label vector $\mathbf{y}_j \in \{0, 1\}^M$,

where $\mathbf{y}_{j,c} = 1$ indicates that label c is present in neighbor j (and $\mathbf{y}_{j,c} = 0$ otherwise). The retrieval prior for label c is then the fraction of the k neighbors that contain it:

$$p_{\text{ret}}^{(c)}(x) = \frac{1}{k} \sum_{j \in \mathcal{I}_k(x)} \mathbf{y}_{j,c}. \tag{7}$$

This formulation produces an empirical neighborhood prevalence for each label and naturally preserves co-occurrence patterns present in the reference cohort.

2.3 Logit Fusion Strategy

The final stage of our inference pipeline synergizes the general medical knowledge encapsulated in the pre-trained VLM with the specific domain evidence derived from the retrieval prior.
First, we compute the standard Zero-Shot Logits $\mathbf{z}_{zs} \in \mathbb{R}^M$. Let $\mathbf{W}_{text}$ denote the normalized text embeddings derived from class-specific prompt templates (e.g., "*a histopathological image of* $\{class\}$"). The zero-shot score is calculated as the cosine similarity between the robust visual embedding $\mathbf{v}$ and the text prototypes:

$$\mathbf{z}_{zs} = \mathbf{v} \cdot \mathbf{W}_{text}^{\top}. \tag{8}$$

To integrate the retrieval evidence, we employ a weighted log-probability injection. Let $\mathbf{p}_{ret}$ be the task-adaptive prior distribution calculated in Section 2.2. The final enhanced logits $\mathbf{z}_{final}$ are computed as:

$$\mathbf{z}_{final} = \frac{\mathbf{z}_{zs} + \beta \cdot \log(\mathbf{p}_{ret} + \epsilon)}{T}, \tag{9}$$

where ϵ is a small constant (10^{-6}) ensuring numerical stability. This formulation introduces two critical hyperparameters governing the fusion:

- β: Controls the strength of the retrieval signal. A higher β indicates greater reliance on the support set neighbors compared to the abstract text description.
- T: A temperature scaling parameter that calibrates the sharpness of the final probability distribution.

The final logits are converted into task-specific prediction probabilities according to the label structure of each dataset. For the single-label multi-class tasks, the final class probabilities are obtained using the softmax function:

$$p_{\text{final}}^{(c)}(x) = \frac{\exp\left(z_{\text{final}}^{(c)}\right)}{\sum_{m=1}^{M} \exp\left(z_{\text{final}}^{(m)}\right)}, \tag{10}$$

and the predicted class is

$$\hat{y} = \arg \max_{c \in \{1,\ldots,M\}} p_{\text{final}}^{(c)}(x). \tag{11}$$

For the multi-label task, each label is treated independently. The final logits are converted into per-label probabilities using the sigmoid function:

$$p_{\text{final}}^{(c)}(x) = \sigma\left(z_{\text{final}}^{(c)}\right) = \frac{1}{1 + \exp\left(-z_{\text{final}}^{(c)}\right)}. \tag{12}$$

For operating-point metrics, label c is predicted as positive when $p_{\text{final}}^{(c)}(x) \geq 0.5$. Continuous probabilities from Eq. 12 are used for ROC-AUC computation.

2.4 Hyperparameter Calibration via Masked Self-Retrieval

A distinct advantage of our framework is that it operates in a training-free manner, requiring no gradient updates to the model backbone. However, selecting optimal values for β and T is crucial for performance. To determine these parameters without peeking at the test set, we propose a Masked Self-Retrieval strategy that utilizes only the training support set $\mathcal{D}_{train}$.

We treat the support set as a validation query set against itself. For each sample $(\mathbf{k}_i, \mathbf{y}_i) \in \mathcal{G}$, we execute the retrieval process against the gallery. Crucially, to prevent data leakage (where a sample trivially retrieves itself as a perfect match), we enforce a masking constraint on the similarity matrix:

$$\text{Sim}(\mathbf{k}_i, \mathbf{k}_j) = \begin{cases} -\infty & \text{if } i = j \\ \mathbf{k}_i \cdot \mathbf{k}_j & \text{otherwise} \end{cases} \tag{13}$$

Using this masked retrieval, we perform a grid search over $\beta \in [0, 1]$ and $T \in [0.5, 2.0]$. We select the pair (β^*, T^*) that maximizes the domain-specific evaluation metric on the support set. These optimal parameters are then frozen and applied during the final inference on the held-out test set.

2.5 Datasets and study partitions

We evaluate the framework on three medical imaging datasets spanning histopathology, dermatology, and radiography. Table 1 summarizes the statistics and data splits for each dataset. We also report normalized Shannon entropy (H_{norm}) computed from the training label distribution, where values closer to 1 indicate more balanced class distributions.

2.5.1 NCT-CRC (Pathology)

We use NCT-CRC-HE-100K [26] as the source dataset for the support gallery and evaluate on the independent CRC-VAL-HE-7K cohort. From the 100,000 available H&E-stained patches in NCT-CRC-HE-100K, 80,002 images were randomly selected to construct the labeled support gallery used for retrieval. The gallery contains examples from all nine colorectal tissue classes. CRC-VAL-HE-7K includes 7,180 patches from unseen patients and is used exclusively for final evaluation. No test images are included in the retrieval gallery or used during hyperparameter calibration, ensuring a strict separation between support and evaluation data.

2.5.2 ISIC (Dermoscopy)

For skin lesion analysis, we utilize a subset of the ISIC archive hosted on kaggle [27] containing 2,357 dermoscopic images across nine diagnostic categories, including Melanoma, Basal Cell Carcinoma, and Vascular Lesions. The dataset is characterized by significant class imbalance. We follow a patient-wise split with 2,239 training images and 118 testing images.

2.5.3 NIH Chest X-ray (Radiology)

We employ the NIH ChestX-ray14 dataset [28], which contains 112,120 frontal chest X-ray images annotated with 14 thoracic pathology labels (e.g., Pneumonia, Effusion, and Cardiomegaly). This is a multi-label classification task. We follow the official patient-wise split based on `train_val_list.txt` and `test_list.txt`, yielding 86,524 training images and 25,596 testing images, which avoids patient-level leakage between training and testing.

Table 1 Summary of datasets and experimental partitions used in the study.

Dataset	Task	Classes/labels	Gallery	Test	H_{norm}	Split
NCT-CRC-HE	Multi-class	9	80,002	7,180	0.994	Patient-wise
NIH ChestX-ray14	Multi-label	14	86,524	25,596	0.849	Patient-wise
ISIC 2018	Multi-class	9	2,239	118	0.917	Patient-wise

2.6 Implementation Details

All experiments were conducted on a single NVIDIA GPU with 24GB memory, using MedSigLIP-448 as the backbone vision–language model. The framework is strictly training-free: model weights remain frozen, no adapters are fine-tuned, and no gradients are computed. Accordingly, all gains arise solely from inference-time procedures, namely multi-scale test-time augmentation (TTA), retrieval-based prior estimation, and logit-level fusion. For fair comparison, all baselines and proposed variants use the same backbone, class-name prompts, and preprocessing pipeline, differing only in whether TTA and retrieval are enabled. All images are resized to 448×448. The zero-shot baseline computes image–text cosine similarity logits directly from the frozen embeddings, without TTA or retrieval. TTA uses deterministic center crops at three scales (100%, 90%, and 80%), followed by embedding extraction and arithmetic averaging as described in Section 2.1. Image and text embeddings are ℓ_2-normalized before similarity computation. For retrieval-augmented inference, the support gallery is constructed exclusively from the training split, with each training image embedded using the same TTA procedure to ensure consistency between query and gallery representations. Nearest neighbors are retrieved using cosine similarity with brute-force Top-k search, and the resulting retrieval prior is fused with zero-shot logits at the log-probability level. Hyperparameters (β, T) are selected using masked self-retrieval on

the training split (Section 2.4), with no test-set access.
To ensure strict separation between training and evaluation data, the retrieval gallery is constructed exclusively from the training split, and test images are never indexed or used during neighbor search. Hyperparameter selection is performed solely via masked self-retrieval within the training data. For the datasets, splits are patient-wise to prevent cross-patient leakage. No duplicate or overlapping samples are shared between training and test splits. This guarantees that all reported test results reflect true out-of-sample evaluation. Since the framework is training-free and uses frozen embeddings, repeated runs under identical settings produced negligible variation.

2.7 Evaluation Metrics

Performance is evaluated separately for the multi-class NCT-CRC and ISIC datasets and the multi-label NIH ChestX-ray14 dataset. Macro-averaging denotes computing each metric independently per class or label and then averaging across classes, thereby reducing the influence of class-frequency imbalance.
For the multi-class tasks, we report accuracy, macro recall (sensitivity), macro F1-score, and Cohen's kappa. Accuracy measures overall classification correctness, while macro recall and macro F1 provide class-balanced measures of predictive performance. Cohen's kappa quantifies agreement beyond chance and is particularly informative for imbalanced class distributions.
For NIH ChestX-ray14, all metrics are calculated independently for each of the 14 findings and then macro-averaged. Macro-AUC is used as the primary metric to quantify ranking performance from continuous prediction scores. For operating-point evaluation, we additionally report macro F1-score, macro label accuracy, and macro Matthews correlation coefficient (MCC) using a fixed threshold of 0.5. MCC is included because it remains informative under class imbalance and asymmetric error profiles [29].

2.8 Statistical analysis

To determine whether improvements over the standard zero-shot baseline are statistically significant, hypothesis tests appropriate to each task formulation are applied. For the single-label NCT-CRC and ISIC datasets, McNemar's test is used to compare paired predictions from the baseline and proposed framework. For NIH ChestX-ray14, where macro-AUC is the primary evaluation metric, DeLong's test is used to compare correlated ROC curves. Statistical significance is determined at $p < 0.05$.

2.9 Neighborhood-size sensitivity analysis

The retrieval neighborhood size k is treated as a dataset-dependent inference hyperparameter. To evaluate sensitivity to this choice, we sweep

$$k \in \{1, 3, 5, 10, 20, 50, 64, 100, 200\}. \tag{14}$$

During this analysis, β, T, and τ remain fixed. The experiment is designed to quantify how retrieval performance varies as the local neighborhood expands and to

assess whether the optimal neighborhood size differs according to gallery size, intra-class variability, and embedding structure. The operating points used in the main experiments are $k = 5$ for NCT-CRC, $k = 10$ for ISIC 2018, and $k = 64$ for NIH ChestX-ray14.

2.10 Support-gallery-size analysis

To quantify the dependence of retrieval-augmented inference on the amount of labeled reference data, we progressively reduce the support gallery to

$$\{1, 5, 10, 25, 50, 100\}\%$$

of its original size. This corresponds to galleries ranging from approximately 800 to 80,002 images for NCT-CRC, 23 to 2,239 images for ISIC 2018, and 865 to 86,524 images for NIH ChestX-ray14. For the single-label NCT-CRC and ISIC datasets, reduced galleries are constructed using class-stratified sampling so that all classes remain represented. For NIH ChestX-ray14, repeated subsets are sampled from the multi-label support pool, with all 14 findings represented at every evaluated fraction. For each reduced-gallery configuration from 1% to 50%, the experiment is repeated using five random seeds (42, 123, 456, 789, and 2026), and the mean and standard deviation are reported. The complete 100% gallery is deterministic and is evaluated once. Importantly, the evaluation cohort, frozen model, prompt configuration, image preprocessing, and retrieval-fusion hyperparameters remain unchanged throughout this analysis. The neighborhood sizes used in the main experiments are retained: $k = 5$ for NCT-CRC, $k = 10$ for ISIC 2018, and $k = 64$ for NIH ChestX-ray14. The fusion parameters β, T, and τ, where applicable, are likewise fixed to their main-experiment values and are not re-optimized for any gallery fraction. Therefore, this experiment isolates the effect of support-gallery size without introducing additional test-set tuning.

3 Results

We evaluate the proposed framework across three medical imaging domains and compare it against the standard zero-shot MedSigLIP baseline. Results are reported for both multi-class and multi-label settings, with statistical significance assessed as described in Section 2.8.

3.1 Performance on Multi-Class Tasks

Table 2 summarizes the classification performance on the NCT-CRC and ISIC 2018 datasets. **Histopathology (NCT-CRC):** On the independent CRC-VAL-HE-7K evaluation set, the pure zero-shot MedSigLIP baseline achieves 50.46% accuracy, 47.33% macro-F1, 53.77% macro recall, and a Cohen's Kappa of 0.44. In comparison, our framework substantially improves performance, reaching 94.57% accuracy, 93.01% macro-F1, 93.25% macro recall, and a Cohen's Kappa of 0.94, corresponding to an absolute accuracy gain of 44.11 percentage points. McNemar's test confirms that the improvement is highly significant ($\chi^2 = 2903.35$, $p < 0.001$). At the individual-sample level, the proposed framework corrects 3,332 cases that are misclassified by

the pure zero-shot baseline, while introducing only 151 new errors among samples that were previously classified correctly. The strong improvement therefore indicates that retrieval-based grounding remains highly effective for structured histopathological morphology even under this more challenging evaluation protocol. The behavior of the framework on NCT-CRC is examined further in Section 4, with particular emphasis on morphological consistency and the effectiveness of local neighborhood evidence.
Dermoscopy (ISIC 2018): On ISIC 2018, the baseline achieves 15.25% accuracy and 11.40% macro-F1. Our framework improves accuracy to 54.24% and macro-F1 to 51.71%, with Cohen's Kappa increasing from 0.05 to 0.47. McNemar's test confirms that the improvement is statistically significant ($\chi^2 = 40.5, p < 0.001$). While the magnitude of improvement is lower than in histopathology, the consistent gains indicate that retrieval provides beneficial contextual grounding even in visually heterogeneous dermatological images.

Table 2 Performance of the standard zero-shot baseline and the proposed framework on the multi-class datasets. All metrics are reported as percentages, with Cohen's kappa scaled by 100 for presentation consistency.

Dataset	Method	Accuracy	Macro-F1	Kappa	Macro recall
NCT-CRC	Zero-shot	50.46	47.33	44.42	53.77
	Proposed framework	**94.57***	**93.01**	**93.77**	**93.25**
ISIC 2018	Zero-shot	15.25	11.40	4.80	12.50
	Proposed framework	**54.24***	**51.71**	**47.20**	**53.47**

*Statistically significant improvement over the zero-shot baseline in accuracy according to McNemar's test ($p < 0.001$).

3.2 Performance on the multi-label task

For NIH ChestX-ray14, macro-AUC was used as the primary evaluation metric. The proposed framework improved macro-AUC from 68.97% with standard zero-shot inference to 77.55%, corresponding to an absolute gain of 8.58 percentage points. Macro-MCC increased from 0.13 to 0.25, while macro-F1 increased from 0.18 to 0.30. DeLong's test indicated statistically significant AUC improvements across all 14 findings ($p < 0.001$). The largest gains were observed for Mass, with an increase of 22.1 AUC points, and Emphysema, with an increase of 20.6 AUC points. Table 3 reports the per-label results for all 14 thoracic findings.

3.3 Comparison with representative state-of-the-art methods

Table 4 compares the proposed framework with representative supervised and zero-shot results reported in the literature for NCT-CRC and ISIC 2018. Because the cited studies may differ in dataset composition, class definitions, preprocessing, data splits, and training protocols, these results should be interpreted as contextual reference points rather than direct head-to-head comparisons. On NCT-CRC, supervised

Table 3 Performance of the standard zero-shot baseline and the proposed framework on the NIH ChestX-ray14 test set. Higher values indicate better performance for all metrics.

	Zero-shot				Proposed framework			
Label	**AUC**	**MCC**	**F1**	**Accuracy**	**AUC**	**MCC**	**F1**	**Accuracy**
Atelectasis	0.68	0.20	0.31	0.79	**0.76***	0.28	0.39	0.78
Cardiomegaly	0.83	0.35	0.37	0.93	**0.87***	0.36	0.38	0.95
Consolidation	0.56	0.02	0.13	0.63	**0.73***	0.16	0.23	0.82
Edema	0.79	0.13	0.16	0.89	**0.84***	0.22	0.25	0.94
Effusion	0.77	0.33	0.47	0.76	**0.82***	0.41	0.53	0.80
Emphysema	0.68	0.07	0.12	0.82	**0.89***	0.38	0.40	0.93
Fibrosis	0.70	0.11	0.11	0.92	**0.77***	0.12	0.13	0.97
Hernia	0.69	0.00	0.00	1.00	**0.73***	0.24	0.15	1.00
Infiltration	0.66	0.15	0.31	0.72	**0.70***	0.23	0.46	0.56
Mass	0.54	0.02	0.04	0.92	**0.76***	0.26	0.31	0.89
Nodule	0.62	0.05	0.09	0.91	**0.69***	0.15	0.21	0.85
Pleural Thickening	0.66	0.11	0.13	0.64	**0.73***	0.17	0.21	0.89
Pneumonia	0.65	0.06	0.06	0.68	**0.69***	0.06	0.09	0.94
Pneumothorax	0.82	0.21	0.26	0.43	**0.87***	0.45	0.51	0.87
Macro average	**0.69**	**0.13**	**0.18**	**0.79**	**0.78**	**0.25**	**0.30**	**0.87**

*Statistically significant improvement in AUC over the zero-shot baseline according to DeLong's test ($p < 0.001$).

approaches such as ResNet, DenseNet, EfficientNet-B0, and CRCCN-Net report accuracies ranging from approximately 96% to 99%. Prior zero-shot methods including CONCH and PLIP also achieve strong performance. The proposed framework achieved 94.57% accuracy, substantially improving over the standard MedSigLIP zero-shot baseline and narrowing the gap to supervised approaches. On ISIC 2018, supervised CNN-based approaches achieve substantially higher performance than the proposed training-free framework. Nevertheless, the proposed method increased the MedSigLIP zero-shot baseline from 15.25% to 54.24% accuracy.

For NIH ChestX-ray14, the proposed framework achieved a mean AUC of 0.776. The resulting value was numerically higher than the standard MedSigLIP zero-shot baseline and several reported zero-shot or label-efficient results listed in Table 5; however, these studies used different experimental protocols and are included only for contextual comparison.

3.4 Sensitivity to Neighborhood Size

We analyze sensitivity to the neighborhood size k by sweeping $k \in 1, 3, 5, 10, 20, 50, 64, 100, 200$ and reporting performance in Figure 2. The neighborhood size used for the main experiments is fixed prior to evaluation on the final test sets; the sweep is reported only to assess the robustness of the retrieval mechanism to this choice. For NCT-CRC, $k = 5$ was selected using the internal calibration split and subsequently kept fixed for evaluation on the independent CRC-VAL-HE-7K cohort. On CRC-VAL-HE-7K, accuracy increases from 92.32% at $k = 1$ to 94.58% at the fixed

Table 4 Comparison with representative literature results on the multi-class datasets.

Dataset	Method	Supervision	Accuracy (%)
NCT-CRC	ResNet/DenseNet [30]	Supervised	99.76
	EfficientNet-B0 [31]	Supervised	99.87
	CRCCN-Net [32]	Supervised	96.26
	Post-Hoc CONCH [14]	Zero-shot	92.63
	MR-PHE [33]	Zero-shot	80.22
	MedSigLIP baseline	Zero-shot	50.46
	Proposed framework	**Training-free**	**94.57**
ISIC 2018	AlexNet [34]	Supervised	98.70
	MSRNet [35]	Supervised	85.40
	MSLC-14 [36]	Supervised	82.50
	Inception [37]	Supervised	89.90
	MedSigLIP baseline	Zero-shot	15.25
	Proposed framework	**Training-free**	**54.24**

Table 5 Comparison with representative literature results on NIH ChestX-ray14.

Method	Supervision	Backbone	Mean AUC
Fully supervised			
CONVFCMAE [38]	Supervised	ConvNeXt	0.852
CheXNet [39]	Supervised	DenseNet	0.841
ResNet-38-large [40]	Supervised	ResNet-38	0.806
Zero-shot / label-efficient			
Partial CLIP [41]	Few-shot	CLIP	0.754
MoCoCLIP [42]	Zero-shot	ResNet-50	0.742
MedSigLIP baseline	Zero-shot	MedSigLIP	0.690
Proposed framework	**Training-free**	**MedSigLIP**	**0.776**

$k = 5$ configuration (Figure 2A) and remains within a relatively narrow range for moderate neighborhood sizes, reaching 95.48% at $k = 20$ before gradually decreasing for larger values. For ISIC 2018, accuracy improves with small neighborhoods, reaches its maximum at $k = 10$, and decreases beyond $k = 20$(Figure 2B). For NIH-CXR14, macro-AUC increases steadily with larger neighborhoods in the tested range, and we use $k = 64$ as the default setting for the main results (Figure 2C). Unless otherwise stated, all experiments use $k = 5$ for NCT-CRC, $k = 10$ for ISIC 2018, and $k = 64$ for NIH-CXR14.

3.5 Ablation Study: Impact of TTA and Retrieval

To isolate the contributions of our two inference-time components, we evaluate the zero-shot baseline (ZS), ZS with multi-scale test-time augmentation (ZS+TTA), and

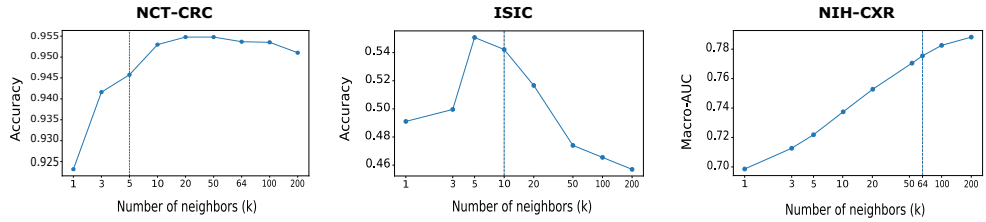


Fig. 2 Sensitivity to neighborhood size k across datasets. (A) NCT-CRC accuracy increases rapidly for small neighborhoods and remains relatively stable across moderate neighborhood sizes; the dashed line indicates the fixed $k = 5$ operating point. (B) ISIC 2018 shows a clear optimum at $k = 10$, with performance decreasing as k increases beyond moderate neighborhood sizes. (C) NIH-CXR14 Macro-AUC improves steadily with larger k, with $k = 64$ used as the default configuration in the main experiments.

the full system with retrieval-aware fusion (ZS+TTA+Ret). Table 6 summarizes results across all domains.

Table 6 Component ablation across the three imaging domains. Accuracy is reported for the multi-class datasets and macro-AUC for NIH ChestX-ray14. Δ denotes the absolute change relative to the standard zero-shot baseline.

Dataset	Metric	ZS	ZS+TTA	Δ	ZS+TTA+Ret	Δ
NCT-CRC	Accuracy	50.46	49.17	−1.29	**94.57**	+45.40
ISIC 2018	Accuracy	15.25	25.42	+10.17	**54.24**	+38.99
NIH-CXR14	Macro-AUC	69.00	69.10	+0.10	**77.60**	+8.60

Effect of TTA. The effect of TTA is domain dependent. On NCT-CRC, adding multi-scale TTA slightly decreases accuracy from 50.46% to 49.17% (−1.29 percentage points) on the independent CRC-VAL-HE-7K evaluation set. This indicates that center-crop feature averaging does not provide an advantage for this histopathology cohort in isolation and may remove useful tissue context in some patches. In contrast, on ISIC 2018, TTA improves accuracy from 15.25% to 25.42% (+10.17), showing clear benefits in visually heterogeneous dermoscopic images where scale and framing vary significantly. On NIH-CXR14, its effect remains marginal (69.00% to 69.10%, +0.10), consistent with the more standardized acquisition geometry of chest radiographs.
Effect of Retrieval. Adding retrieval-aware fusion produces the dominant improvements in all settings. On NCT-CRC, retrieval acts as a strong corrective prior, boosting performance from 49.17% (ZS+TTA) to 94.57%. On ISIC, retrieval further raises accuracy from 25.42% to 54.24%. On NIH-CXR14, retrieval increases macro-AUC from 69.0% to 77.6%, indicating that incorporating in-domain visual evidence is substantially more effective than augmentation alone.

3.6 Effect of Labeled Support Gallery Size

We next evaluate how strongly retrieval-augmented inference depends on the amount of labeled support data. Figure 3 shows markedly different support-size behavior across the three medical imaging domains.

On NCT-CRC, the framework is highly support-efficient. Using only 1% of the full gallery, corresponding to approximately 800 labeled histopathology patches, already achieves $93.80 \pm 0.81\%$ accuracy, compared with 94.58% using all 80,002 references. Accuracy reaches $95.08 \pm 0.16\%$ with 10% support and $95.10 \pm 0.31\%$ with 25%, after which performance remains within a narrow range. Macro-F1 follows a similar pattern, increasing from $91.35 \pm 1.07\%$ at 1% support to 93.01% with the complete gallery. These results indicate that the retrieval benefit on NCT-CRC saturates relatively early and does not require a dense reference repository.

ISIC 2018 exhibits substantially stronger dependence on support coverage. Accuracy increases from $20.85 \pm 3.03\%$ with only 1% of the gallery (23 reference images) to $32.03 \pm 4.21\%$ at 5%, $40.51 \pm 2.64\%$ at 10%, and $49.15 \pm 3.23\%$ at 25%. Performance reaches $50.51 \pm 0.97\%$ with half of the gallery and 54.24% using all 2,239 support images. Macro-F1 similarly increases from $10.48 \pm 1.27\%$ at 1% to 51.71% with the complete gallery, demonstrating that broader exemplar coverage is particularly important in this dermoscopic setting.

NIH-CXR14 shows an intermediate and more gradual pattern. Macro-AUROC increases from $74.51 \pm 0.33\%$ at 1% support to $75.96 \pm 0.31\%$ at 5%, $76.29 \pm 0.27\%$ at 10%, and $77.13 \pm 0.13\%$ at 25%. With half of the support gallery, performance reaches $77.45 \pm 0.09\%$, approaching the full-gallery result of 77.55%. The decreasing variance and diminishing performance gains at larger fractions indicate that broader radiographic support improves retrieval coverage, while most of the additional benefit has been obtained by approximately half of the available gallery.

Overall, the analysis demonstrates that retrieval effectiveness is not determined by gallery size alone. NCT-CRC retains strong performance with a compact exemplar memory, ISIC requires substantially broader support coverage, and NIH-CXR14 exhibits gradual gains with diminishing returns. These findings motivate treating support-memory requirements as domain dependent rather than assuming that a uniformly large reference gallery is necessary.

4 Discussion

This study addresses a practical deployment challenge in medical imaging: how to improve zero-shot predictions under domain shift without updating model parameters. Across histopathology, dermoscopy, and chest radiography, the proposed framework consistently improves over standard zero-shot inference, although the magnitude and source of these gains vary substantially across domains. In particular, our ablation results show that retrieval provides the most consistent improvement, whereas the contribution of test-time augmentation is strongly dependent on visual characteristics, dataset structure, and label formulation. These domain-specific differences provide insight into when inference-time visual ensembling and retrieval-based grounding are

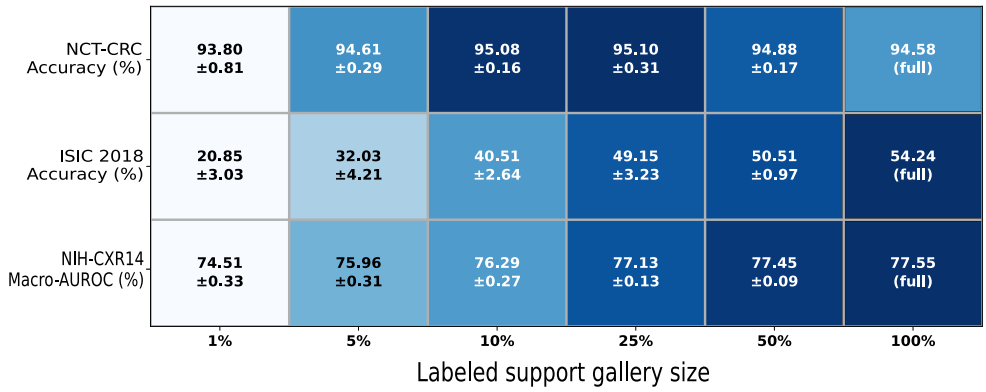


Fig. 3 Performance under reduced labeled support galleries across medical imaging domains. Cell annotations report task-specific performance, while shading is normalized within each dataset row to visualize relative changes in performance as labeled support is reduced. Reduced-gallery settings from 1% to 50% report the mean and standard deviation over five random subsamples, whereas the 100% configuration denotes the deterministic full gallery.

most effective and inform the practical use of training-free adaptation strategies in medical imaging.

The main finding is that the two inference-time components contribute differently across domains. Retrieval-based evidence provides the most consistent corrective signal by grounding predictions in visually similar labeled examples from the target domain, whereas the effect of multi-scale test-time augmentation (TTA) is more domain-dependent. As shown in Table 6, TTA provides a substantial benefit on ISIC 2018, has only a marginal effect on NIH-CXR14, and slightly reduces accuracy on NCT-CRC. In contrast, retrieval produces large improvements across all three datasets. These results suggest that retrieval is the primary source of performance gains when zero-shot semantic predictions are insufficient, while the benefit of TTA depends on whether multi-scale views preserve and enhance the task-relevant visual information.

NCT-CRC exhibits the largest retrieval-driven improvement among the evaluated domains. On the independent CRC-VAL-HE-7K cohort, the proposed framework increases accuracy from 50.46% for zero-shot inference to 94.57%, demonstrating that retrieval remains highly informative even when the evaluation images originate from patients not represented in the support gallery. This behavior is consistent with patch-based histopathology, where tissue categories are characterized by recurring micro-textures and morphological structures that can form coherent neighborhoods in the frozen embedding space [1]. Consequently, visually similar support patches can provide strong class-specific evidence that complements the semantic zero-shot prediction. The neighborhood-size analysis in Fig. 2A further supports this interpretation. Accuracy rises rapidly as k increases from 1 and then remains within a relatively narrow range across moderate neighborhood sizes, indicating that retrieval is not highly sensitive to the exact choice of k. At the fixed operating point of $k = 5$, the framework achieves 94.58% accuracy, while performance remains around 95% across a broader range of neighborhoods. Notably, the strong NCT-CRC performance is primarily

attributable to retrieval rather than TTA: multi-scale TTA alone slightly reduces accuracy from 50.46% to 49.17%, whereas adding retrieval raises performance to 94.57%. This suggests that domain-matched exemplar evidence, rather than visual ensembling, is the dominant source of improvement for this histopathology task. Future work should examine whether the same behavior persists across external institutions, staining protocols, scanners, and whole-slide-level evaluation settings, where morphological and acquisition variability may be substantially greater.

ISIC 2018 shows clear gains from retrieval augmentation and TTA, but it does not approach the near-saturation observed on more uniform datasets such as NCT-CRC. This difference reflects the intrinsic complexity of dermoscopic images. Skin lesions show substantial intra-class variation in morphology, color, and texture [37], along with subtle inter-class boundaries and variable acquisition conditions across sites, including differences in illumination and hardware [2]. In addition, class imbalance can bias the retrieval neighborhood toward common categories. As a result, visually similar images may correspond to different diagnoses, increasing the chance of competing labels within the retrieved set. This is reflected in Fig. 2B, where performance peaks at a moderate k and then declines as additional neighbors add more noise than discriminative signal. In this setting, retrieval is not deterministic evidence but a stabilizing prior that aligns predictions with in-domain patterns. The remaining gap to near-perfect accuracy therefore reflects the ambiguity and high dimensionality of real-world dermoscopic assessment rather than a limitation of the fusion mechanism.

NIH-CXR14 poses a distinct challenge because it is multi-label: clinically related findings can co-occur, and a single image may legitimately contain several labels. In this setting, softmax-style competition across labels is inappropriate because it assumes mutual exclusivity and can suppress true co-occurring findings [15]. The independent retrieval voting scheme is therefore essential, as it estimates per-label support from the neighborhood without forcing labels to compete. Notably, NIH-CXR14 favors a relatively large neighborhood (k=64), consistent with the weaker and more diffuse notion of visual similarity in chest radiographs. A larger k stabilizes per-label evidence by averaging over a broader set of clinically similar cases, reduces sensitivity to noisy or borderline neighbors, and provides a more reliable estimate of label co-occurrence patterns, such as effusion with atelectasis. This helps explain why retrieval-aware fusion improves macro-AUC even when the similarity signal is less discriminative than in the other domains.

The effect of TTA is strongly domain-dependent rather than uniformly beneficial. TTA is most useful when discriminative cues are sensitive to scale, framing, or localized structure, because multiple views can provide complementary representations of the same image. This is consistent with dermoscopy, where lesion size and framing vary substantially and fine-grained structures may be localized; accordingly, TTA increases accuracy from 15.25% to 25.42% even before retrieval is introduced. By contrast, chest X-rays are acquired under relatively standardized geometry (AP/PA views), and many diagnostic cues are global or spatially diffuse [43]; multi-scale ensembling therefore provides only a marginal improvement in macro-AUC from 69.00% to 69.10%. On NCT-CRC, TTA slightly decreases accuracy from 50.46% to 49.17%. One possible explanation is that center cropping removes peripheral tissue context or that averaging

multiple crop embeddings attenuates localized morphological cues that are already well represented in the original patch. These findings indicate that test-time visual ensembling should not be assumed to benefit every medical domain and motivate domain-aware selection of augmentation strategies, while the consistently larger gains from retrieval suggest that exemplar-based grounding is the more robust inference-time adaptation mechanism across the evaluated tasks.

Figure 4 illustrates how retrieval influences prediction quality across the three domains. For clarity, only the top-3 retrieved neighbors are shown, although the retrieval prior is computed using the domain-specific k from Fig. 2. Successful cases typically occur when the retrieved neighborhood is visually coherent and reinforces the correct label, allowing the retrieval signal to correct baseline zero-shot errors. Failure cases arise when the query lies near a class boundary or when visually similar neighbors carry competing labels. In dermoscopy, this reflects high morphological variability and subtle benign–malignant transitions; in histopathology, errors often occur in patches containing mixed tissue patterns; and in chest X-ray, failures are more likely when findings are subtle or when common co-occurring pathologies dominate the retrieved neighborhood. Overall, the qualitative examples support the quantitative results by showing that retrieval is most reliable when the local neighborhood is label-consistent and less reliable under ambiguity or overlap.

An important distinction is that training-free inference is not equivalent to label-free inference. Although the proposed framework does not update model parameters, retrieval requires a labeled support memory. The gallery-size analysis provides a more precise view of this requirement and shows that support needs differ substantially across domains. NCT-CRC retains most of its full-gallery performance using only 1% of the available references, suggesting that a compact set of representative exemplars can provide sufficient local evidence in this relatively structured patch-level histopathology setting. In contrast, ISIC performance improves substantially as additional reference images are introduced, indicating a stronger dependence on support coverage. NIH-CXR14 lies between these regimes, with performance improving steadily as the gallery grows but exhibiting clear diminishing returns at larger fractions. These observations suggest that the usefulness of additional support examples depends not only on their number, but also on the diversity, redundancy, and label structure of the target domain. Consequently, future work should consider support-memory curation and exemplar selection in addition to simply increasing gallery size.

The benchmarking results (Tables 4 and 5) show that the effectiveness of training-free adaptation depends strongly on domain structure, visual homogeneity, class separability, and label formulation. On NCT-CRC, our framework substantially narrows the gap between standard zero-shot inference and fully supervised classification, reaching 94.57% accuracy while remaining below the strongest supervised models. This strong performance is consistent with patch-level histology, where recurring micro-textures and morphological structures can form coherent neighborhoods in the frozen embedding space, enabling retrieval to provide highly informative exemplar-based evidence. In contrast, ISIC remains well below supervised performance because dermoscopic categories exhibit high intra-class variability, subtle inter-class boundaries, and class imbalance. Even when retrieval recovers useful in-domain signal and corrects

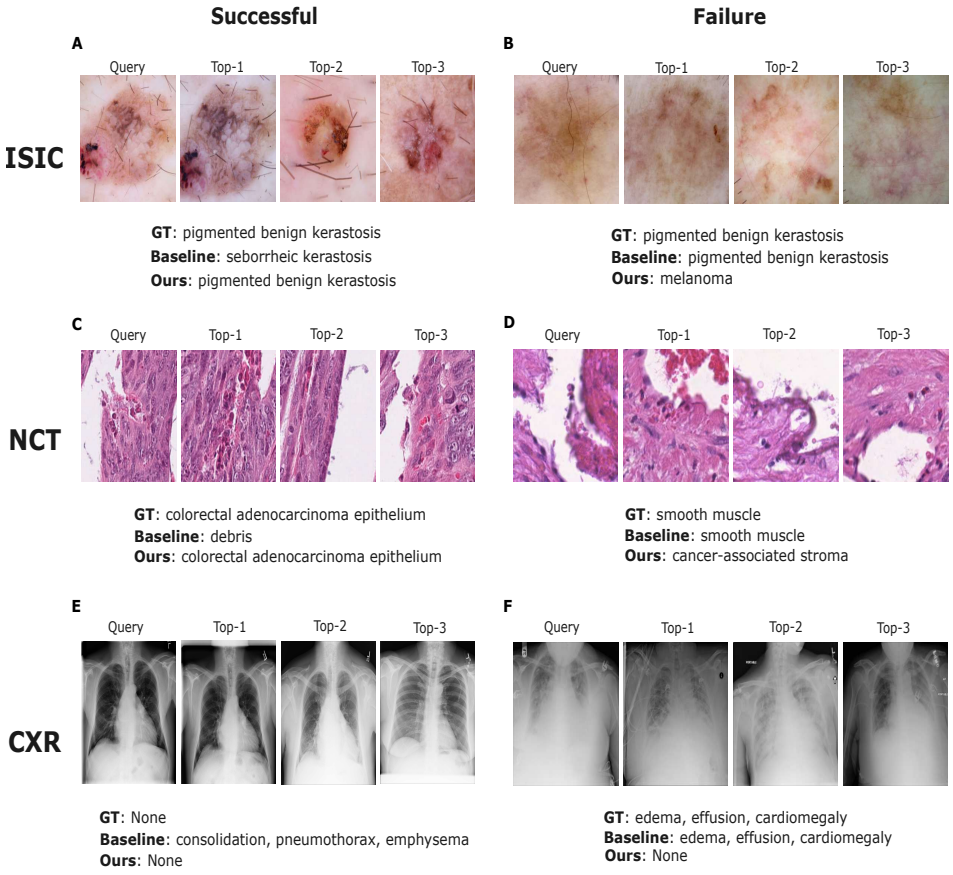


Fig. 4 For each dataset (ISIC, NCT-CRC, NIH-CXR14), we show representative successful (left column) and failure (right column) cases. Each row displays the query image followed by its Top-3 nearest neighbors retrieved from the support memory bank. We report the ground-truth label (GT), the baseline zero-shot prediction, and the prediction of our retrieval-augmented framework.

many prompt-only errors, visually similar lesions may still correspond to competing diagnoses, limiting how closely a non-parametric prior can approach fully trained discriminative models. For NIH-CXR14, the remaining gap to supervised AUC reflects a different challenge: radiographic findings are often subtle, globally distributed, and clinically correlated under noisy multi-label annotations, making visual similarity less precise than in histology. In this setting, retrieval is most useful for aggregating weak per-label evidence and co-occurrence information over broader neighborhoods, which helps explain the consistent improvement over the zero-shot baseline and the competitive performance relative to prior zero-shot approaches.

Despite avoiding gradient-based fine-tuning, the framework introduces practical deployment constraints. Retrieval-augmented inference is more computationally demanding than standard zero-shot prediction because it requires multi-view encoding

for TTA and a k-NN search over a support gallery. Although gallery embeddings can be precomputed and retrieval latency reduced using approximate nearest-neighbor indexing, scaling to very large repositories or real-time applications still requires careful engineering. In addition, label noise, severe class imbalance, or distribution mismatch between the support gallery and the target population can distort neighborhood composition and induce misleading retrieval priors, particularly in heterogeneous domains such as dermoscopy. Retrieval may also provide limited evidence for conditions that are absent or poorly represented in the gallery, making rare or out-of-distribution findings particularly challenging.

The present study also has two important limitations. First, the framework was evaluated using a single medical vision-language backbone, MedSigLIP. Although the proposed inference strategy was assessed across three distinct medical imaging domains, further evaluation using additional medical VLMs is needed to determine how consistently the observed benefits generalize across different pretrained representation spaces. Second, the framework is training-free but not label-free, because retrieval depends on a labeled target-domain support gallery. Its effectiveness therefore remains dependent on the availability and representativeness of suitable reference examples. Future work should evaluate the framework across additional medical VLM backbones and investigate support-memory curation, low-support retrieval, efficient large-scale indexing, and mechanisms for identifying unreliable or out-of-distribution neighborhoods.

5 Conclusion

This paper introduced a unified, training-free framework for augmenting frozen medical vision–language inference without updating model parameters. The approach combines multi-scale test-time augmentation with retrieval-based evidence from a labeled support memory and uses task-adaptive retrieval priors for both single-label and multi-label prediction. Retrieval evidence is integrated with the prompt-based zero-shot logits through calibrated log-probability fusion, with inference hyperparameters determined without gradient-based model training. Across histopathology, dermoscopy, and chest radiography, retrieval-based grounding consistently improves over the frozen zero-shot baseline, whereas the contribution of TTA is strongly domain dependent. The support-gallery-size analysis further shows that the amount of labeled reference data required for effective retrieval varies substantially across domains: NCT-CRC retains strong performance with a highly reduced gallery, NIH-CXR14 exhibits gradual gains with diminishing returns, and ISIC benefits more strongly from broader exemplar coverage. These findings highlight the potential of non-parametric support memories as an inference-time alternative to task-specific fine-tuning while also emphasizing that training-free inference remains dependent on the quality and coverage of labeled reference data. Future work will investigate efficient large-scale retrieval, support-memory curation under distribution shift, mechanisms for detecting unreliable or out-of-distribution neighborhoods, and domain-adaptive test-time augmentation strategies.

Declarations

Funding

No specific funding was received for this study.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable. This study used publicly available, de-identified datasets and did not involve the collection of new human participant data.

Consent for publication

Not applicable.

Data availability

The datasets used in this study are publicly available.

Materials availability

Not applicable.

Code availability

The code used in this study will be made available upon reasonable request.

Author contributions

S.M.N. conceived the study, developed the methodology, implemented the experiments, performed the analysis, and drafted the manuscript. M.E. supervised the study, contributed to the methodological design and interpretation of the results, and critically revised the manuscript. M.Y. contributed to the methodological design. All authors reviewed and approved the final manuscript.

References

[1] Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine **25**(8), 1301–1309 (2019)

[2] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017)

[3] McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., *et al.*: International evaluation of an ai system for breast cancer screening. Nature **577**(7788), 89–94 (2020)

[4] Willemink, M.J., Koszek, W.A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L.R., Summers, R.M., Rubin, D.L., Lungren, M.P.: Preparing medical imaging data for machine learning. Radiology **295**(1), 4–15 (2020)

[5] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PmLR

[6] Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)

[7] Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)

[8] Wang, P., Lu, W., Lu, C., Zhou, R., Li, M., Qin, L.: Large language model for medical images: A survey of taxonomy, systematic review, and future trends. Big Data Mining and Analytics **8**(2), 496–517 (2025)

[9] Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.-M.M., Lin, M.: On evaluating adversarial robustness of large vision-language models. Advances in Neural Information Processing Systems **36**, 54111–54138 (2023)

[10] Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930 (2021)

[11] Javed, S., Mahmood, A., Ganapathi, I.I., Dharejo, F.A., Werghi, N., Bennamoun, M.: Cplip: Zero-shot learning for histopathology with comprehensive vision-language alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11450–11459 (2024)

[12] Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., *et al.*: A visual-language foundation model for computational pathology. Nature medicine **30**(3), 863–874 (2024)

[13] Dai, D., Zhang, Y., Yang, Q., Xu, L., Shen, X., Xia, S., Wang, G.: Pathologyvlm: a large vision-language model for pathology image understanding. Artificial Intelligence Review **58**(6), 1–19 (2025)

[14] Noman, S.M., Henedak, M.T., Elattar, M.: Modular post-hoc enhancements for

zero-shot histopathology classification using vision-language models. Machine Learning: Science and Technology **6**(3), 035056 (2025)

[15] You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 101–111 (2023). Springer

[16] Li, B., Huang, W., Gao, Z., Wang, Y., Shen, Y., Lin, J., You, L., Shen, Y., Lin, S., Ouyang, W., et al.: Llava-radz: Can multimodal large language models effectively tackle zero-shot radiology recognition? arXiv preprint arXiv:2503.07487 (2025)

[17] Yan, S., Yu, Z., Primiero, C., Vico-Alonso, C., Wang, Z., Yang, L., Tschandl, P., Hu, M., Ju, L., Tan, G., Tang, V., Ng, A.B., Powell, D., Bonnington, P., See, S., Magnaterra, E., Ferguson, P., Nguyen, J., Guitera, P., Banuls, J., Janda, M., Mar, V., Kittler, H., Soyer, H.P., Ge, Z.: A multimodal vision foundation model for clinical dermatology. Nature Medicine **31**(8), 2691–2702 (2025) https://doi.org/10.1038/s41591-025-03747-y

[18] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022) https://doi.org/10.1007/s11263-022-01653-1

[19] Zhao, C., Wang, Y., Jiang, X., Shen, Y., Song, K., Li, D., Miao, D.: Learning domain invariant prompt for vision-language models. IEEE Transactions on Image Processing **33**, 1348–1360 (2024) https://doi.org/10.1109/TIP.2024.3362062 . Epub 2024-02-14

[20] Ouali, Y., Bulat, A., Matinez, B., Tzimiropoulos, G.: Black box few-shot adaptation for vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15534–15546 (2023)

[21] Long, A., Yin, W., Ajanthan, T., Nguyen, V., Purkait, P., Garg, R., Blair, A., Shen, C., Hengel, A.: Retrieval augmented classification for long-tail visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6959–6969 (2022)

[22] Saranya, E., Chinnadurai, M.: Medical application driven content based medical image retrieval system for enhanced analysis of x-ray images. Scientific Reports **15**(1), 29115 (2025)

[23] Shanmugam, D., Blalock, D., Balakrishnan, G., Guttag, J.: Better aggregation in test-time augmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1214–1223 (2021). https://doi.org/10.1109/ICCV48922.2021.00125

[24] Sherkatghanad, Z., Abdar, M., Bakhtyari, M., Plawiak, P., Makarenkov, V.:

Baytta: Uncertainty-aware medical image classification with optimized test-time augmentation using bayesian model averaging. arXiv preprint (2024) https://doi.org/10.48550/arXiv.2406.17640 arXiv:2406.17640 [cs.CV]

[25] Zanella, M., Ben Ayed, I.: On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 23783–23793 (2024)

[26] Kather, J.N., Halama, N., Marx, A.: 100,000 histological images of human colorectal cancer and healthy tissue, april 2018. URL https://doi. org/10.5281/zenodo **1214456** (2024)

[27] Gutman, D., Codella, N.C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1605.01397 (2016)

[28] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2097–2106 (2017)

[29] Chicco, D., Jurman, G.: The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC genomics **21**(1), 6 (2020)

[30] Khazaee Fadafen, M., Rezaee, K.: Ensemble-based multi-tissue classification approach of colorectal cancer histology images using a novel hybrid deep learning framework. Scientific Reports **13**(1), 8823 (2023)

[31] Abd El-Ghany, S., Mahmood, M.A., Abd El-Aziz, A.: Adaptive dynamic learning rate optimization technique for colorectal cancer diagnosis based on histopathological image using efficientnet-b0 deep learning model. Electronics **13**(16), 3126 (2024)

[32] Kumar, A., Vishwakarma, A., Bajaj, V.: Crccn-net: Automated framework for classification of colorectal tissue using histopathological images. Biomedical Signal Processing and Control **79**, 104172 (2023) https://doi.org/10.1016/j.bspc.2022.104172

[33] Rahaman, M.M., Millar, E.K., Meijering, E.: Leveraging vision-language embeddings for zero-shot learning in histopathology images. IEEE Journal of Biomedical and Health Informatics (2025)

[34] Hosny, K.M., Kassem, M.A., Fouad, M.M.: Classification of skin lesions into seven

classes using transfer learning with alexnet. Journal of digital imaging **33**(5), 1325–1334 (2020)

[35] Bibi, S., Khan, M.A., Shah, J.H., Damaševičius, R., Alasiry, A., Marzougui, M., Alhaisoni, M., Masood, A.: Msrnet: Multiclass skin lesion recognition using additional residual block based fine-tuned deep models information fusion and best feature selection. Diagnostics **13**(19), 3063 (2023)

[36] Arora, G., Dubey, A.K., Jaffery, Z.A., Rocha, A.: A comparative study of fourteen deep learning networks for multi skin lesion classification (mslc) on unbalanced data. Neural Computing and Applications **35**(11), 7989–8015 (2023)

[37] Shahin, A.H., Kamal, A., Elattar, M.A.: Deep ensemble learning for skin lesion classification from dermoscopic images. In: 2018 9th Cairo International Biomedical Engineering Conference (CIBEC), pp. 150–153 (2018). https://doi.org/10.1109/CIBEC.2018.8641815

[38] Xiong, K., Tu, Y., Rao, X., Zou, X., Du, Y.: Multi-label disease detection in chest x-ray imaging using a fine-tuned convnextv2 with a customized classifier. MDPI **12**(3), 80 (2025)

[39] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)

[40] Baltruschat, I.M., Nickisch, H., Grass, M., Knopp, T., Saalbach, A.: Comparison of deep learning approaches for multi-label chest x-ray classification. Scientific reports **9**(1), 6381 (2019)

[41] Nitinkumar Dalsania, H.: Label-efficient chest x-ray diagnosis via partial clip adaptation. arXiv e-prints, 2507 (2025)

[42] Bhardwaj, P., Bhat, S., Maier, A.: Enhancing zero-shot learning in medical imaging: integrating clip with advanced techniques for improved chest x-ray analysis. arXiv preprint arXiv:2503.13134 (2025)

[43] Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature biomedical engineering **6**(12), 1399–1406 (2022)