Enhancing Medical Vision-Language Foundation Models in Tumour Malignancy Recognition via Pre-trained Language Models

XIAO WANG

MPhil

THE UNIVERSITY OF
SYDNEY

Supervisor: Prof. Jinman Kim
Associate Supervisor: Prof. Adam Dunn

A thesis submitted in fulfilment of
the requirements for the degree of
Master of Philosophy

School of Computer Science
Faculty of Engineering
The University of Sydney
Australia

16 May 2026

Generative AI Contribution Statement

During the preparation of the thesis, the author used ChatGPT and Gemini for purposes such as text enhancement, including paraphrasing, refining sentence structure, and correcting spelling and grammar.

All AI-assisted outputs were carefully reviewed by the author to identify and correct any potential errors, inaccuracies, or biases. The author takes full responsibility for the submitted thesis and ensures the work is their own and has used generative AI within the parameters of use (refer to the University of Sydney generative AI guide for researchers).

Statement of Originality

Statement of Originality

I certify that:

- To the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.
- I understand that failure to comply with the Student Plagiarism: Academic Board Policy: Academic Dishonesty and Plagiarism can lead to the University commencing proceedings against me for potential student misconduct under the Thesis and Examination of Higher Degrees by Research Procedures 2025.

This Work is substantially my own, and to the extent that any part of this Work is not my own, I have indicated that it is not my own by acknowledging the source of that part or those parts of the work.

Name: Xiao Wang

Signature: **Date: 28 Oct 2025**

Authorship Attribution Statement

Sections 3.1 to 3.5 in Chapter 3 of this thesis are based on the publication X. Wang, U. Naseem and J. Kim, **"Enhancing Zero-Shot Learning of Pathology Vision-Language Foundation Models in Tumour Malignancy Recognition"** European Conference on Artificial Intelligence 2025. My contributions include methodology design, experiment execution, results analysis and paper writing. In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Name: Xiao Wang

Signature: **Date: 28 Oct 2025**

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Supervisor: Prof. Jinman Kim

Signature: **Date: 31 Oct 2025**

Abstract

Vision-Language Foundation Models (VLFMs) show promise for zero-shot learning (ZSL), where models must generalise to rare or new disease subtypes without extensive new image annotation. In histopathology image analysis, however, their effectiveness is limited by weak image-text alignment because coarse-grained textual descriptions often fail to capture fine-grained visual details. This misalignment introduces semantic noise and imprecision into zero-shot retrieval, weakening the initial semantic link between an unseen image and the corresponding textual knowledge base and degrading downstream classification. To address this problem, this thesis proposes Retrieval-based De-noising Causal Language Modelling (RDCLM), a framework that refines noisy retrieval outputs from pathology VLFMs. RDCLM constructs a pathology-specific knowledge base of fine-grained, discriminative tumour malignancy descriptions using a large language model (LLM). Given a query histopathology image, a pathology VLFM retrieves candidate descriptions from this knowledge base. The proposed de-noising module, which leverages a frozen language model, integrates visual features with the retrieved texts to filter irrelevant content and improve semantic alignment. This process improves retrieval precision and enables more accurate zero-shot image classification. To further improve performance and generalisation, two retrieval augmentation strategies are proposed: Retrieval Negatives Replacement (RNR) and Description-wise Shuffling (DS). A Multi-Branch Diverse-Dimension Projection architecture is also proposed, using an auxiliary Inter- and Intra-Branch Min-Max Mutual Information Optimisation objective to encourage diverse yet informative feature learning across parallel projection branches. In addition, an auxiliary Precision Reward Loss is evaluated as a training refinement for encouraging cleaner de-noised descriptions. Extensive evaluations across five histopathology cancer datasets demonstrate that RDCLM significantly outperforms state-of-the-art methods in both zero-shot

image-text retrieval and malignancy classification. These results highlight the importance of retrieval de-noising for advancing VLFM-based zero-shot learning in histopathology.

Acknowledgements

The author is sincerely grateful to Prof. Jinman Kim and Dr. Usman Naseem and Prof. Adam Dunn for their invaluable supervision of this study, which provided a plethora of useful suggestions and guidance.

Contents

CHAPTER 1

Introduction

The burgeoning field of Vision-Language Foundation Models (VLFMs) has significantly advanced cross-modal understanding, enabling tasks such as zero-shot cross-modal retrieval and image classification with minimal supervision in medical imaging [1, 2, 3, 4, 5]. VLFMs are deep learning models designed to process and learn from both visual (images) and textual (language) data simultaneously, aligning these two distinct modalities in a shared representation space. In contrast to traditional uni-modal approaches, which are trained on a single data type, such as vision-only image classification or text-only language modelling, and often rely on supervised learning [6, 7, 8], VLFMs can exploit large-scale paired image-text data [9, 10, 11]. Self-supervised approaches such as contrastive learning [12] have therefore emerged as powerful alternatives. By leveraging paired image-text data, self-supervised approaches [13] align visual and textual modalities, allowing models to learn rich representations without the need for explicit labels. This is particularly advantageous in medical imaging, where obtaining labels requires an expensive and time-intensive manual process due to the involvement of skilled pathologists and specialised equipment [14, 9, 11].

A prominent example of a VLFM for medical imaging is Contrastive Language–Image Pre-training (CLIP) [15], which employs a dual-encoder architecture to map images and texts into a shared embedding space and learn semantic correspondence through contrastive

loss. This formulation has proven impactful with medical images, where its fine-grained representation learning can retain subtle visual cues in pathological images [1, 3]. Building on this foundation, models such as CONCH [3], CPLIP [4], and PLIP [1] extend CLIP-style training to large-scale pathology datasets or medical text corpora, achieving competitive zero-shot performance across classification and retrieval tasks. MI-Zero [2] tackles gigapixel WSIs via instance-level similarity pooling, while PTUnifier [16] introduces soft prompts for better domain adaptation across datasets.

Histopathology, the examination of tissues under a microscope, remains the gold standard for definitive tumour diagnosis and malignancy grading. This critical process hinges on expert pathologists interpreting complex morphological patterns, such as nuclear pleomorphism, mitotic activity, and invasion boundaries, across vast, gigapixel Whole Slide Images (WSIs). Given the rapidly increasing volume of tissue samples and the subtle nature of these diagnostic features, there is an urgent clinical need for robust and generalisable computational tools to aid rapid and accurate malignancy recognition. Current artificial intelligence approaches, however, often struggle with the fine-grained, visually dense nature of these images.

Despite these advances, a persistent limitation remains: the heavy reliance on coarse-grained textual descriptions that are either scraped from online sources or generated by LLMs [1, 2]. Such descriptions often fail to capture critical fine-grained visual cues—such as tumour subtypes, malignancy levels, or organ-specific features—thereby weakening the semantic alignment between images and text [1, 4]. This results in suboptimal performance in high-stakes recognition tasks like tumour malignancy image-text retrieval and classification, where subtle histopathological patterns are important [1, 3]. Figure 1.1 demonstrates the weak alignment problem of a pathology VLFM on the concept of tumour malignancy. Given an image of a malignant tumour tissue, the VLFM may retrieve semantically inconsistent texts, such as benign-tissue descriptions, generic cancer-related phrases, or repeated wrong-class morphology descriptions. These recurring retrieval-noise patterns reduce the usefulness of the retrieved evidence pool and motivate a retrieval de-noising approach.

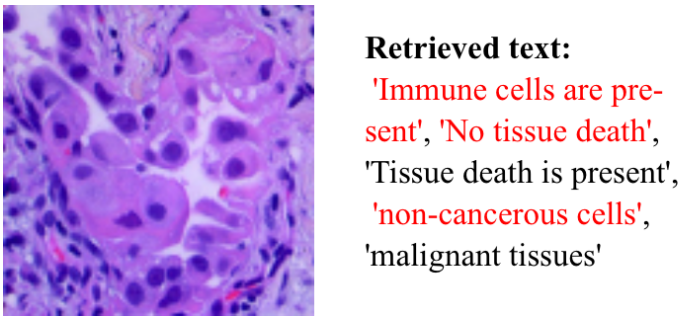


FIGURE 1.1. Characterisation of irrelevant information in retrieved pathology text. An example of noisy results from the baseline vision-language foundation model (PLIP) for a malignant tumour query image, where irrelevant texts (benign tissue descriptions) are shown in red. (An irrelevant/negative text description for any tumour image is a description of the opposite tumour class.)

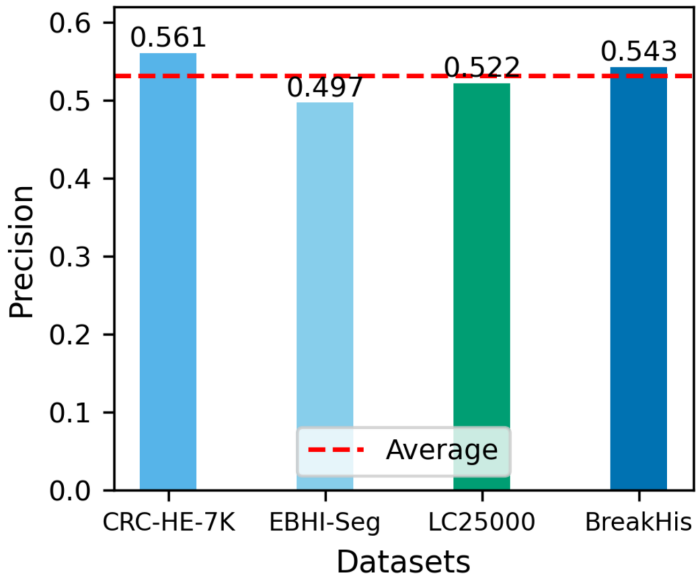


FIGURE 1.2. Characterisation of irrelevant information in retrieved pathology text. Top-k retrieval precision across four histopathology cancer datasets [17, 18, 19, 20]. The red dashed line marks the average precision (0.531), revealing a 46.9% noise level in the VLFM-based pathology image-text retrieval.

To address this challenge, recent work has explored enhancing image-text alignment through refined prompts and captions. For instance, CLIP-A-self [21] uses GPT-4 to replace class labels with detailed visual descriptions, improving few- and zero-shot performance. CLIP^{FT}+A [22] generates geographically and visually enriched prompts via LLMs, while [23] leverages retrieved captions as anchors to refine class semantics. While effective, these approaches

primarily operate on prompt quality, knowledge base design, or CLIP-style adaptation. The line of work closest to RDCLM is therefore retrieval-augmented and knowledge-base-enhanced vision-language learning; however, a remaining gap is that noisy retrieved textual evidence is usually used directly for downstream prediction instead of being explicitly filtered before classification.

In this work, the focus is shifted to the raw retrieval outputs of VLFMs, which remain noisy and semantically inconsistent, particularly for rare or nuanced concepts like tumour malignancy. In Figure 1.2, the analysis of PLIP reveals that over 46% of retrieved texts for pathology images are irrelevant, which adversely affects downstream generation, cross-modal retrieval and classification performance. It is hypothesised that refining these retrieved outputs—rather than modifying prompts or encoders—can substantially improve zero-shot reasoning in medical imaging.

To this end, Retrieval-based De-noising Causal Language Modelling (RDCLM), a novel zero-shot learning framework, is proposed to enhance VLFM retrieval by filtering and refining noisy outputs. First, a pathology-specific knowledge base is constructed using a large language model, prompting it to generate detailed textual descriptions of benign and malignant tumour features. Note that our approach differs from prior LLM-based generation methods [1, 2] by focusing on distinct class-specific concepts rather than multi-level or overlapping descriptions. This is based on the intuition that distinct semantics between benign and malignant tumours are easier for the model to learn. Given a query H&E image, PLIP is used to retrieve the top-k semantically similar texts from this knowledge base. These retrieved candidates are then passed to our de-noising module, a causal language model trained to synthesise refined, contextually relevant outputs.

The core of RDCLM is a multimodal fusion mechanism that jointly embeds visual tokens from the image and textual tokens from retrieved candidates. To prevent overfitting and promote robust semantic learning, a negative replacement strategy is introduced—where noisy retrieved texts are replaced with challenging negative descriptions—and a description-wise shuffling technique to disrupt positional biases. This encourages the model to attend to semantically meaningful patterns across modalities.

To enhance the cross-modal alignment capacity of the image-text projection while avoiding overfitting on limited training samples, a multi-branch diverse-dimension projection architecture equipped with Inter- and Intra-Branch Min-Max Mutual Information Optimisation is proposed. Instead of stacking deep non-linear projection layers, our method arranges multiple projection branches in parallel, each operating in different feature dimensions to capture distinct semantic information. To ensure each branch learns unique yet meaningful representations, a mutual information–based objective is designed that simultaneously maximises the mutual information between each branch output and the input image feature (promoting informativeness) and minimises the mutual information between outputs of different branches (encouraging diversity). This auxiliary optimisation, combined with the main objective L_{dclm}, enables the model to extract fine-grained, diverse semantic representations, thereby improving zero-shot generalisation.

To address the persistent issue of noise and irrelevant information in the generated descriptions, an auxiliary Precision Reward Loss L_{pr} is also evaluated as a training refinement. This loss is not the central contribution of the thesis, but it provides an additional way to encourage cleaner textual outputs by weighting the main de-noising objective with a precision-based reward term.

Through extensive experiments on five pathology cancer datasets, it is demonstrated that RDCLM significantly improves zero-shot image-text retrieval and image classification accuracy, outperforming prior state-of-the-art approaches. The main contributions are as follows:

- **Retrieval-based De-noising Causal Language Modelling (RDCLM)**, a novel zero-shot learning framework, is proposed to enhance VLFM retrieval by filtering and refining noisy outputs for tumour malignancy recognition.
- A **pathology-specific knowledge base** is constructed using a large language model to generate fine-grained, distinct textual descriptions of benign and malignant tumour features, specifically tailored for histopathology retrieval.
- A causal language modelling-based de-noising module is designed to integrate visual and retrieved textual representations using multimodal token fusion, employing

retrieval negatives replacement and **description-wise shuffling** for robust semantic alignment.

- A **Multi-Branch Diverse-Dimension Projection** architecture with an **Inter- and Intra-Branch Min-Max Mutual Information Optimisation** objective is introduced to enhance cross-modal alignment and zero-shot generalisation by promoting diverse and informative feature learning.

As an auxiliary analysis, the thesis also evaluates Precision Reward Loss (PRL) as a training refinement for encouraging cleaner de-noised descriptions.

It is demonstrated that RDCLM achieves state-of-the-art performance in zero-shot image-text retrieval and classification across five pathology cancer datasets, showcasing the effectiveness of retrieval de-noising for enhancing VLFM-based zero-shot learning in medical imaging.

CHAPTER 2

Literature Review

2.1 Histopathology Imaging

Histopathology imaging plays a central role in disease diagnosis and research by enabling microscopic examination of tissue morphology. In clinical workflows, tissue samples are first excised, fixed in formalin, embedded in paraffin, and sectioned into thin slices of approximately 3–5 μm thickness. These sections are mounted on glass slides and stained to highlight cellular and subcellular structures. The most widely used stain is Haematoxylin and Eosin (H&E), where haematoxylin stains nuclei in shades of blue or purple and eosin counterstains cytoplasmic and extracellular matrix components in varying tones of pink [24]. This colour contrast allows pathologists to assess tissue architecture, cellular morphology, and abnormalities indicative of malignancy or inflammation. Other special stains and immunohistochemistry (IHC) techniques may be employed to detect specific proteins or molecular markers; however, H&E remains the cornerstone of diagnostic histopathology and the primary modality used in computational pathology research [25, 8].

With advances in digital pathology, glass slides can be scanned at high resolution using whole-slide imaging (WSI) scanners, producing gigapixel-scale digital images that faithfully capture the microscopic detail of tissue sections. WSIs are typically acquired at magnifications of

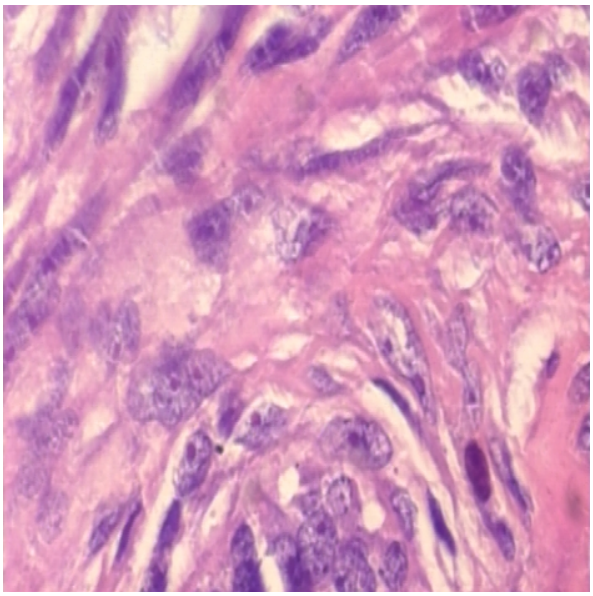


FIGURE 2.1. Histopathology image of a benign phyllodes tumour in breast tissue.

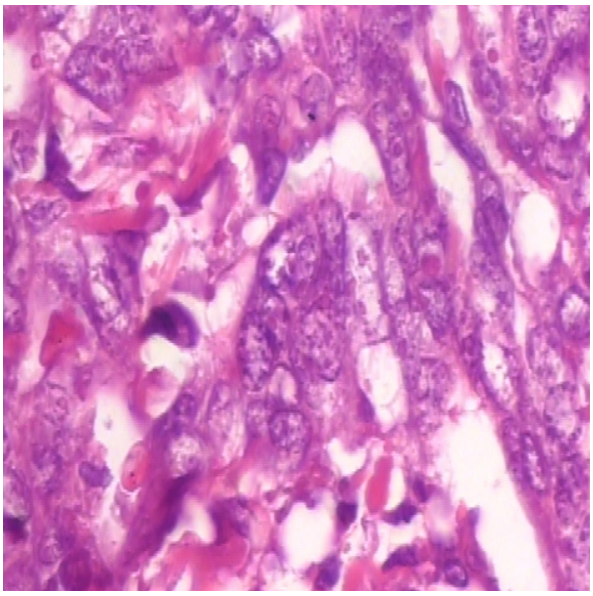


FIGURE 2.2. Histopathology image of a malignant papillary carcinoma in breast tissue.

$20\times$ or $40\times$, corresponding to pixel resolutions between 0.25 and 0.50 μm per pixel [26]. The digitisation of slides facilitates remote consultation, archival storage, and, most importantly, large-scale computational analysis. However, WSIs introduce challenges such as variations in staining, scanner type, illumination, and slide preparation, all of which can adversely

affect algorithmic generalisation [27]. To address the extreme resolution and size of WSIs, most computational methods adopt a patch-based strategy, dividing the slide into smaller, non-overlapping regions for feature extraction and analysis [14].

Early computational pathology relied on handcrafted texture, colour, and morphological features, but these approaches were limited in robustness and scalability. The emergence of convolutional neural networks (CNNs) revolutionised histopathology image analysis by enabling end-to-end feature learning directly from raw image data [8]. Initial CNN-based works focused on patch-level classification tasks, such as tumour versus normal tissue detection, mitosis counting, and cancer grading. Representative datasets like BreakHis (breast cancer histopathology) and Camelyon16 (metastatic lymph node detection) established common evaluation benchmarks for supervised learning in pathology [20, 28]. These efforts demonstrated that deep networks could achieve expert-level accuracy for well-curated datasets; however, limited annotation availability and the expense of manual labelling remained major obstacles.

To alleviate these challenges, weakly supervised and multiple instance learning (MIL) frameworks were introduced to exploit slide-level labels without pixel-level annotations [14, 29]. In such approaches, each WSI is treated as a "bag" of smaller image patches, and learning is performed by aggregating patch-level representations to predict slide-level outcomes. Campanella et al. [14] demonstrated a landmark clinical-grade computational pathology system using weak supervision on over $44{,}000$ WSIs, achieving diagnostic performance comparable to human pathologists. These developments enabled scalable training on large, weakly labelled datasets and facilitated clinical translation of deep learning models.

More recently, self-supervised and transformer-based architectures have emerged as dominant paradigms in histopathology representation learning. Vision transformer (ViT) variants and hierarchical models, such as HIPT, have been designed to capture both local cellular morphology and global tissue organisation by operating at multiple magnification levels [30]. Contrastive self-supervised frameworks, including RetCCL, further improved generalisation

by leveraging instance discrimination and clustering-based objectives to learn robust embeddings from unlabelled WSIs [31]. These pretraining strategies have led to significant gains in transfer learning and downstream performance, especially when labelled data are limited.

The latest frontier in histopathology imaging research extends beyond vision-only models toward multimodal and vision-language paradigms. By pairing WSIs with diagnostic text reports or pathology descriptions, vision-language foundation models aim to align visual features with semantic medical concepts, enabling zero-shot classification and retrieval across diverse tissue types. Notable examples include PLIP, CONCH, and Quilt-1M, which apply contrastive or joint image–text pretraining to large-scale pathology datasets [3, 32]. These models demonstrate promising cross-domain transfer capabilities and pave the way for generalisable pathology foundation models that integrate visual, textual, and clinical knowledge. Despite these advances, challenges remain in handling noisy textual supervision, domain shifts across institutions, and the interpretability of learned representations.

In summary, histopathology imaging represents a unique and information-rich modality in medical imaging. The combination of gigapixel-scale visual detail, strong correlation with molecular and clinical outcomes, and growing availability of digitised slides has made it a prime candidate for developing robust and generalisable vision and vision-language models. Continuous progress in computational pathology is expected to further bridge the gap between automated analysis and clinical diagnostics.

2.2 Zero-shot image classification

Zero-shot image classification (ZSIC) seeks to recognise classes that were not observed during supervised training by leveraging auxiliary semantic information such as natural language descriptions, attributes, or knowledge-base entries. The emergence of large-scale vision–language models (VLMs), most notably CLIP [15], marked a step-change: contrastive pretraining on massive image–text pairs yields a shared embedding space in which textual class descriptions can be directly matched to image embeddings for zero-shot prediction. CLIP-style approaches demonstrated strong out-of-domain transfer across a wide variety

of natural-image benchmarks, motivating subsequent efforts that adapt this paradigm to specialised domains such as medical and pathology images [1, 2, 3, 4].

Two main families of methods have been developed for improving ZSIC performance:

(A) Better textual anchors and prompt engineering. Early practical improvements focused on constructing stronger textual class anchors. Prompt engineering and prompt ensembling (averaging text embeddings from multiple templates) improve robustness to template choice [15]. Recent work leverages large language models (LLMs) to generate richer class descriptions or to refine prompts, yielding better semantic coverage and improved zero-shot accuracy [21, 22]. In pathology and other fine-grained domains, detailed visual descriptions (e.g., tissue-level attributes, staining characteristics) provide stronger anchors than single-word labels because they expose discriminative visual cues to the VLM.

(B) Retrieval-augmented and knowledge-enhanced methods. A complementary approach augments the textual anchors by retrieving or constructing textual exemplars from external corpora or knowledge bases. Retrieval-augmented classification retrieves nearest-neighbour text snippets or image–text pairs from a precompiled library and uses them as anchors or context for classification [23, 33]. In medical imaging, constructing domain-specific knowledge bases or using LLMs to synthesise diagnostic descriptions has become a practical solution to the scarcity of paired image–text data [1, 3]. Methods that combine retrieval with LLM-based re-ranking or refinement show strong promise: the retrieval step provides candidate exemplars that cover class variability, while the LLM can condense or normalise noisy retrieved texts into robust labels.

Challenges for zero-shot classification in specialised domains. Despite the successes above, ZSIC faces several domain-specific challenges: **Fine-grained visual differences:** Medical images often require subtle local cues (cell morphology, tissue architecture) that coarse-grained web-scraped captions or short labels do not capture [1, 3]. **Noisy or irrelevant retrievals:** Retrieval from broad knowledge bases can return semantically irrelevant text (e.g., benign descriptors for malignant images), which degrades downstream zero-shot classification if not filtered or denoised. **Label ambiguity and class overlap:** Some clinical categories

overlap visually or share attributes, making discriminative language anchors harder to craft without expert curation. **Limited domain-specific training data:** When adapting VLMs, overfitting small pathology datasets is a risk; thus approaches that freeze large components and only tune small adapters or lightweight projectors are attractive.

Recent directions and relevance to RDCLM Recent research directions aim to address the above challenges by (i) improving the quality and specificity of textual anchors (LLM-based generation and filtering), (ii) using retrieval to supply diverse contextual text examples, and (iii) applying post-retrieval refinement to mitigate noise. Retrieval-augmented LLM pipelines have been shown to be effective in compressing and normalising noisy retrievals into task-specific descriptions; this insight underpins RDCLM's strategy of retrieval followed by de-noising using a causal LM. Complementary architectural advances such as multi-branch projection and mutual-information regularisation aim to preserve both informativeness and diversity in the projected image features, reducing overfitting while capturing multiple semantic aspects necessary for fine-grained zero-shot decisions [12].

Overall, the ZSIC literature suggests two recurring themes: (1) high-quality textual semantics are as important as the image encoder, and (2) handling retrieval noise and in-domain specificity is critical in specialised applications. RDCLM combines both principles by constructing a pathology-specific knowledge base with LLM assistance, retrieving candidate descriptions with a VLFM, and deploying a de-noising LM together with multi-branch projection and precision-guided objectives to produce robust zero-shot classifications in histopathology.

2.3 From Uni-modal Vision Models to Vision–Language Models

Historically, computational pathology treated whole-slide image (WSI) analysis as a uni-modal, image-only problem: models were trained to map raw pixel information to diagnostic labels, prognostic scores, or segmentation masks. Convolutional neural networks (CNNs) established the dominant paradigm for patch-level analysis, learning hierarchical features that capture cellular morphology and texture for tasks such as tumour detection, mitosis counting

and subtype classification. However, WSIs are gigapixel in scale and typically annotated only at the slide level, creating a weak supervision problem that demands algorithmic designs beyond straightforward fully supervised CNNs.

Multiple-instance learning (MIL) is the standard formulation for weakly supervised WSI classification. In MIL, a slide is treated as a bag of instances (patches) and a permutation-invariant aggregator maps instance representations to a slide label. Attention-based MIL introduced trainable, differentiable aggregation that selectively weights diagnostically important patches and yields interpretable attention maps [34]. These attention-MIL approaches improved empirical performance and interpretability relative to simple pooling strategies, but they often assume instance independence and can under-exploit spatial or contextual relationships among patches.

To address inter-patch correlations and long-range tissue architecture, transformer-style architectures and correlated-MIL frameworks were proposed. TransMIL introduced correlated MIL with self-attention over instance tokens, enabling modelling of both morphological and spatial relationships across patches and demonstrating improved performance on several WSI benchmarks [35]. Hierarchical transformers further exploit the multi-scale nature of pathology: the Hierarchical Image Pyramid Transformer (HIPT) uses self-supervised pretraining across magnification levels to learn scalable representations that capture both single-cell detail and broader tissue microenvironments [30]. Complementary to architectural changes, self-supervised and clustering-guided contrastive methods (e.g., RetCCL) learn robust patch and slide embeddings without dense labels, improving retrieval, transfer learning, and interpretability [31].

Mixed supervision and hybrid pipelines combine the strengths of instance-level (patch) and slide-level supervision. Methods such as MS-CLAM integrate a small amount of region-level annotation with large-scale slide labels, improving both classification accuracy and localisation of predictive regions [36]. These hybrid strategies reduce annotation burden while preserving the model's ability to localise histological features relevant to a diagnosis.

Despite these major advances, uni-modal image models remain limited by information that is present only in pixels. Clinical reports, pathology notes, and slide captions encode complementary semantic information — diagnostic terms, differential considerations, specimen site, and ancillary test results — that can disambiguate visually similar patterns. Vision–language learning brings textual semantics into the modelling loop by aligning visual features with descriptive text via contrastive or joint embedding objectives. In general medical imaging, frameworks such as MedCLIP demonstrated that contrastive pretraining with image–text pairs (even when noisy or unpaired) yields strong zero-shot and few-shot capabilities, provided false negatives and noisy supervision are carefully handled [37].

In histopathology specifically, large vision–language datasets and pathology-aware VLMs have recently emerged. Quilt-1M compiled one million image–text pairs sourced from educational videos, slide captions, and other pathology content, showing that large, automatically curated corpora can substantially improve zero-shot patch classification and cross-modal retrieval [32]. CONCH (Contrastive Learning from Captions for Histopathology) and related pathology VLMs (e.g., PLIP variants) apply large-scale image–caption pretraining to produce foundation models that transfer to downstream pathology tasks and support retrieval, captioning, and zero-shot classification [3, 38].

Adapting VLMs to WSIs requires addressing pathology-specific challenges: (1) text supervision is frequently coarse (slide-level) and noisy, so region-level alignment or retrieval-augmented methods are necessary to associate fine-grained histological structures with phrases or labels; (2) gigapixel images demand patching and intelligent region selection so that contrastive objectives operate at the right granularity; and (3) domain shifts across laboratories (stain variability, scanner differences) and heterogeneity in report style necessitate robust pretraining and domain-aware fine-tuning. Recent solutions include region-aware contrastive losses that match salient patches to phrases, retrieval pipelines that gather semantically relevant text for a slide before contrastive alignment, and mixed supervision schemes that use small curated region captions to bootstrap fine-grained alignment [3, 32, 37].

Overall, the trajectory from uni-modal image methods to vision–language models in computational pathology is characterised by (a) improved representation learning via self-supervision

and transformers, (b) architectural and algorithmic strategies for weak supervision (MIL, attention, mixed supervision), and (c) the incorporation of textual semantics through large-scale image–text pretraining and region-aware alignment. This progression enables zero-shot and few-shot transfer, improved interpretability, and retrieval-based augmentation of slide-level predictions — capabilities that are essential for clinically useful pathology foundation models and that motivate the retrieval and alignment improvements proposed in this thesis.

2.4 Prompt engineering to enhance zero-shot image classification

Prompt engineering—designing and selecting the textual inputs used to query vision–language models (VLMs)—has emerged as a simple yet powerful mechanism for improving zero-shot image classification. The core idea is that the quality, specificity, and diversity of textual anchors (prompts) strongly influence how well image embeddings align with class semantics in the shared VLM embedding space. Early demonstrations with CLIP showed that carefully chosen natural-language templates and ensembling across multiple templates substantially improve robustness and average accuracy compared to single-token class names [15]. In specialised domains such as medical imaging and pathology, where discriminative cues are fine-grained and domain-specific, prompt engineering becomes even more crucial: short or generic labels often fail to describe the visual characteristics that distinguish classes, whereas richer textual descriptions can expose relevant features to the VLM [1, 3].

Below we summarise major prompt-engineering strategies that have been proposed to enhance zero-shot classification, along with their motivations, typical implementation choices, and relevance to pathology VLM applications.

1. Template design and prompt ensembling The simplest strategy is to craft multiple natural-language templates (e.g., “A photo of a `{label}`”, “An image showing `{label}`”, etc.) and average the resulting text embeddings (prompt ensembling). This reduces sensitivity to wording and often yields consistent improvements over single templates [15]. In practice, ensembles of diverse templates help capture different phrasing and contextual cues, which is beneficial when class names are ambiguous or when images exhibit wide intra-class variability.

2. LLM-generated descriptive prompts Large language models (LLMs) can generate richer, attribute-focused textual descriptions for each class or concept. Works such as CLIP-A-self and related studies leverage GPT-family or instruction-tuned LLMs to expand class labels into multi-sentence visual descriptions that emphasise discriminative attributes (colour, texture, shape, diagnostic markers). These LLM-augmented prompts often yield substantial gains in zero-shot performance because they provide the VLM with more explicit, image-relevant semantic signals than terse class names [21, 22]. In pathology, LLM-generated descriptions can encode tissue-level attributes (e.g., nuclear pleomorphism, glandular architecture) that map more directly to histological appearance [1].

3. Descriptor augmentation and retrieval-augmented prompts Rather than relying solely on static prompt templates, some approaches retrieve class-relevant snippets or exemplar captions from an external corpus or knowledge base and use these retrieved texts as dynamic prompts [23]. Retrieval-augmented prompts can better reflect intra-class variability and provide concrete visual exemplars for classes lacking concise labels. This strategy is highly relevant to RDCLM-like pipelines: retrieved descriptions supply context that can be refined (or de-noised) before being used for classification, improving robustness to noisy or incomplete textual anchors.

4. Soft prompts and continuous prompt tuning A complementary line of work moves away from discrete natural-language prompts toward learnable continuous prompts or prefix-tuning, where a small set of continuous vectors (soft tokens) are optimised while the large VLM remains frozen. These methods (e.g., prompt tuning, LoRA-style adapters) yield competitive performance with minimal parameter updates and are particularly attractive when domain-specific labelled data are scarce. Soft prompts can implicitly encode domain-specific semantics that are hard to express in fixed natural-language templates, but they lack the interpretability and human readability of LLM-generated text prompts.

5. Instruction-tuning and mixture-of-prompts Instruction-tuned LLMs and prompt mixtures—where prompts encode different types of contextual information (attributes, questions, conditional instructions)—can be used to produce tailored textual anchors that emphasise

different diagnostic aspects. Combining multiple prompt sources (LLM descriptions, human-curated attributes, retrieved exemplars) and aggregating their embeddings often improves zero-shot performance by integrating complementary semantic views.

Challenges and limitations Despite their effectiveness, prompt-engineering strategies face several practical challenges in specialised domains: **Domain specificity:** Generic web-scale VLMs may not fully interpret specialised clinical adjectives or subtle histological descriptions unless those descriptors are carefully crafted or generated by domain-aware LLMs [1, 3]. **Noisy or misleading prompts:** Retrieval-augmented or LLM-generated prompts can be noisy or contain overlapping, non-discriminative attributes; without filtering or refinement this noise may reduce classification precision. **Interpretability vs. performance:** Soft prompts often perform well but are not human interpretable, complicating clinical validation and review. **Computational cost and reproducibility:** Generating and ensembling large numbers of prompts (or querying LLMs at scale) adds computation and may introduce reproducibility constraints unless prompts and LLM seeds are shared.

Relation to RDCLM Prompt engineering and RDCLM address complementary parts of the zero-shot pipeline. RDCLM's retrieval-and-de-noise paradigm can be seen as a principled extension of retrieval-augmented prompting: instead of directly using retrieved or LLM-generated text as final anchors, RDCLM refines and filters candidate descriptions with a de-noising causal LM and aligns them to visual features via a multi-branch projector. This makes the resulting textual anchors cleaner and more reliably discriminative, while retaining the benefits of rich, LLM-informed prompts. Conversely, advances in prompt engineering (better templates, LLM-generated descriptions, and carefully curated attribute lists) directly improve the quality of the knowledge base and retrieved candidates that RDCLM relies upon—making the two approaches synergistic.

Potential improvements For pathology zero-shot applications, best practice is to combine multiple prompt strategies: (i) use LLMs to generate fine-grained, attribute-aware descriptions for candidate classes, (ii) retrieve exemplar snippets from domain corpora to capture intra-class diversity, (iii) apply lightweight filtering or de-noising (as in RDCLM) to remove irrelevant content, and (iv) where feasible, complement natural-language prompts with small

learnable prompt adapters to capture remaining domain-specific structure. Together these steps improve both the semantic content of prompts and the robustness of zero-shot classification in fine-grained medical imaging domains.

2.5 Medical Vision-Language Foundation Models

Recent medical VLFMs trained on large-scale multimodal pathology resources have demonstrated strong zero-shot transfer capacity [32, 2, 5, 39]. PLIP [1], which adapts the CLIP architecture to image-text pairs from online medical forums, highlights the value of diverse medical information for cross-modal learning. Similarly, CONCH [3] adds image-captioning objectives and is pre-trained on large-scale histopathology image-caption datasets, achieving strong zero-shot performance in classification and retrieval tasks. Unlike VLFMs that rely on curated ground-truth image-text pairs, CPLIP [4] reduces data-curation cost by pre-training on large-scale pseudo-captions retrieved from an LLM-generated knowledge base using a pathology VLM. However, despite these advancements, these approaches still rely on coarse image-text alignment and do not explicitly address noise in retrieved textual evidence.

In contrast to these methods, this work directly tackles the issue of noise in retrieved textual examples by introducing a retrieval-based de-noising framework, RDCLM, to refine VLFM outputs.

2.6 Enhancing Zero-Shot Learning with Refined Knowledge Base

To address coarse text supervision, several studies have refined textual descriptions using external knowledge bases to improve zero-shot tasks. Zhou et al. [40] structured a pathology knowledge tree of 50,470 attributes across 4,718 diseases and integrated it through knowledge-enhanced VL pretraining, significantly improving cross-modal retrieval and zero-shot tumour subtyping. CLIP-A-self [21] improves zero-shot classification by pairing LLM-generated category descriptions with existing fine-grained datasets (“bag-level” supervision), yielding 4–5% accuracy gains on iNaturalist and CUB benchmarks. CLIP $^{\mathrm{FT}}$+A [22] enhances CLIP prompts using GPT-4-generated visual descriptions, achieving up to 4% improvement over

manual prompts and demonstrating the value of LLM-refined anchors. SuS-X [33] mitigates large mean and variance effects in intra-modal similarity by computing attention weights from inter-modal similarities between class prompts, test images, and a support set retrieved from a large image-text knowledge base. However, using common category names as input in these approaches may be less effective for classification tasks involving less common concepts, such as category attributes (e.g., tissue malignancy).

While these methods refine the knowledge base or prompts, the proposed approach refines the retrieval outputs from the knowledge base, making the downstream prediction more robust to noise and irrelevant information, particularly for fine-grained attributes such as tumour malignancy.

CHAPTER 3

Methods

The methodology uses a multi-stage framework for zero-shot VLFM-based image-text retrieval and text-based classification of pathology images. This framework includes: (1) generating a pathology knowledge base using a Large Language Model (LLM), focused on the distinct visual features of benign and malignant tissues; (2) constructing pseudo captions for pathology images by retrieving relevant textual descriptions from the knowledge base using a vision-language foundation model (VLFM); (3) applying a retrieval-based de-noising causal language modelling approach (RDCLM), which refines the retrieved pseudo captions by learning the dependencies between the query image and both relevant and irrelevant texts; (4) performing text-based malignancy classification using the de-noised text; (5) introducing a Multi-Branch Diverse-Dimension Projection architecture equipped with Inter- and Intra-Branch Min-Max Mutual Information Optimisation to enhance the cross-modal alignment capacity of the image-text projection by encouraging diverse yet informative feature learning; and (6) evaluating an auxiliary Precision Reward Loss (PRL) as a training refinement for improving the precision and cleanliness of the de-noised textual outputs.

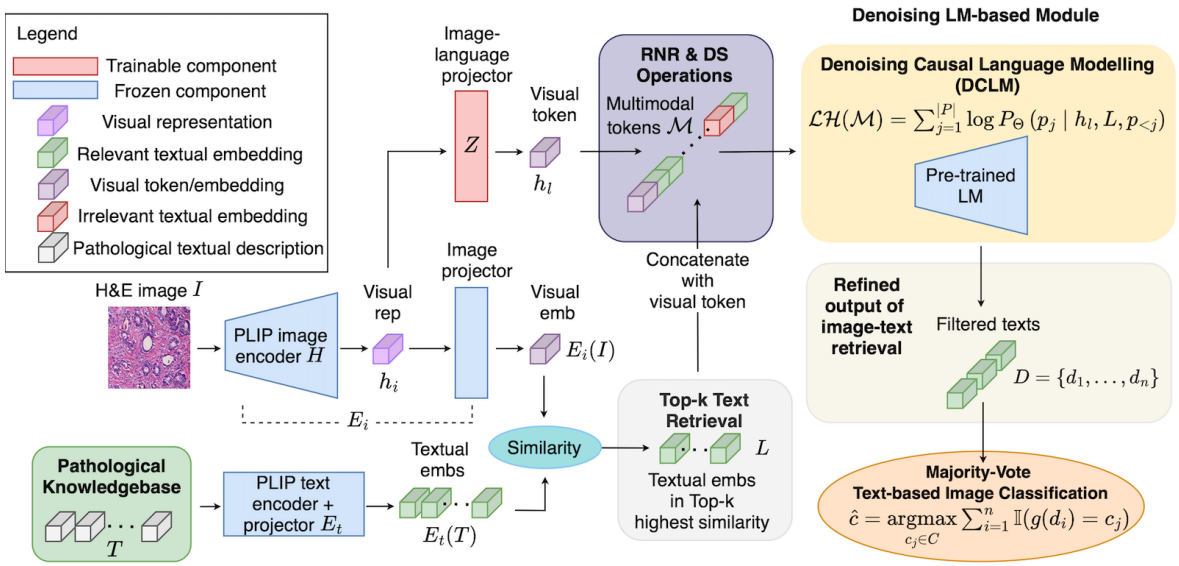


FIGURE 3.1. The overall framework of the retrieval-based de-noising causal language modelling approach.

3.1 Constructing Pathology Knowledge Base using LLM

To generate malignancy-specific, fine-grained descriptions, a large language model (Google Gemini) is prompted to produce textual descriptions of the unique visual features of benign and malignant tumour tissues in H&E-stained images. This leverages LLMs' strong performance in factual generation and simple reasoning shown in [4, 21, 22].

Since benign and malignant tissues can share some visual patterns, a second step is added to filter out non-discriminative features using the LLM. Specifically, the model is prompted to identify and remove attributes common to both classes, retaining only the descriptors unique to each.

The prompts for knowledge base generation and filtering are shown below:

(1) Generation: Generate detailed descriptions of the visual features of benign and malignant tumours in H&E images.

(2) Filtering: Filter out features that appear in both benign and malignant classes. Retain only the features unique to each class.

This two-step process first yields an original LLM-generated knowledge base with 150 total class-balanced descriptions across the benign and malignant classes. A shortened version of the same descriptions is then generated, and the final hybrid setting combines the original and shortened sets, giving approximately 300 descriptions in total. The use of this scale is an empirical design choice rather than a claim of optimal knowledge base size; knowledge bases larger than 300 descriptions were not tested in this thesis. The effect of knowledge base composition is evaluated in the ablation study in Chapter 5.

3.2 Pseudo Caption Generation

To address the lack of captions for pathology H&E images when using VLFM-based approaches, a retrieval-based method is employed. This method leverages a pathology VLFM (PLIP [1]) to retrieve textual descriptions from our customised knowledge base. PLIP was chosen as the retrieval backbone because it is a pathology-specific CLIP-style vision–language model, making it well suited for histopathology image–text alignment. Moreover, the aim of this study is to evaluate the benefit of retrieval de-noising on top of a strong pathology-oriented retriever, rather than to compare different retrieval backbones exhaustively.

In the pseudo caption generation process, the descriptions $T = \{t_1, t_2, \ldots, t_n\}$ from the knowledge base are encoded and projected into textual embeddings $E_t(T)$ using the text projector E_t of the PLIP model [1]. Similarly, the image embedding $E_i(I)$ of a query image I is projected using the image projector E_i of the PLIP model.

Next, the cosine similarities $Sim(.)$ between the query image embedding and all knowledge base textual embeddings are calculated.

Finally, the texts whose textual embeddings have the top-k highest similarity to the query image embedding are selected, as shown in Equation 3.1.

$$R = \text{top-}k\,\{t_i \mid Sim(E_i(I), E_t(t_i)) \text{ for } t_i \in T\} \tag{3.1}$$

$$\text{Sim}(\mathbf{u}, \mathbf{v}) = \frac{\sum_{i=1}^{d} u_i v_i}{\sqrt{\sum_{i=1}^{d} u_i^2}\sqrt{\sum_{i=1}^{d} v_i^2}} \tag{3.2}$$

where k is an adjustable parameter representing the number of retrieved texts R, $\mathbf{u} = (u_1, u_2, \ldots, u_d)$ and $\mathbf{v} = (v_1, v_2, \ldots, v_d)$ are embedded visual and textual vectors, and d is the dimension of PLIP's projected outputs.

3.3 Retrieval Negatives Replacement and Description-wise Shuffling (RNR&DS)

Although the retrieval process approximates the query image's visual semantics, it can suffer from retrieval redundancy, where similar or positionally fixed descriptions are repeatedly selected.

These limitations can be understood through two problems:

- Retrieval Negative Redundancy (the "familiar negative" problem): As illustrated in Figure 3.2, a small number of negative (irrelevant) text descriptions are retrieved very frequently (large, bright circles) while most others are rarely seen (small dots). This causes the model to overfit on this small set of common "easy negatives" and struggle with more challenging, less frequent, or novel irrelevant descriptions during testing. Retrieval Negatives Replacement (RNR) addresses this by swapping out these frequent, familiar negatives with randomly sampled, less frequent negative texts from the knowledge base, as demonstrated by the more spread-out, uniform frequency distribution in Figure 3.3. This forces the model to learn a more robust decision boundary against a wider variety of distracting texts.
- Positional Bias (the "fixed order" problem): If the retrieval mechanism consistently returns descriptions in a similar ranked order—for instance, placing the best retrieved positive text consistently at a fixed rank, showing imbalanced distributions of retrieval ranking—the downstream Causal Language Model (CLM) can learn to rely on this positional cue rather than the actual semantic content. This prevents the CLM from robustly attending to text features regardless of their location. Description-wise Shuffling (DS) mitigates this by randomly permuting the order of the retrieved texts before they enter the CLM. Figure 3.4 illustrates its effect: shuffling makes

the density distribution of positive texts smoother and less peaked at a fixed rank, indicating that the model is forced to consider descriptions across all positions.

These issues can significantly impair downstream reasoning, as shown in Figures 3.2, 3.3, and 3.4. To mitigate these limitations, two retrieval augmentation techniques are proposed: retrieval negatives replacement and description-wise shuffling.

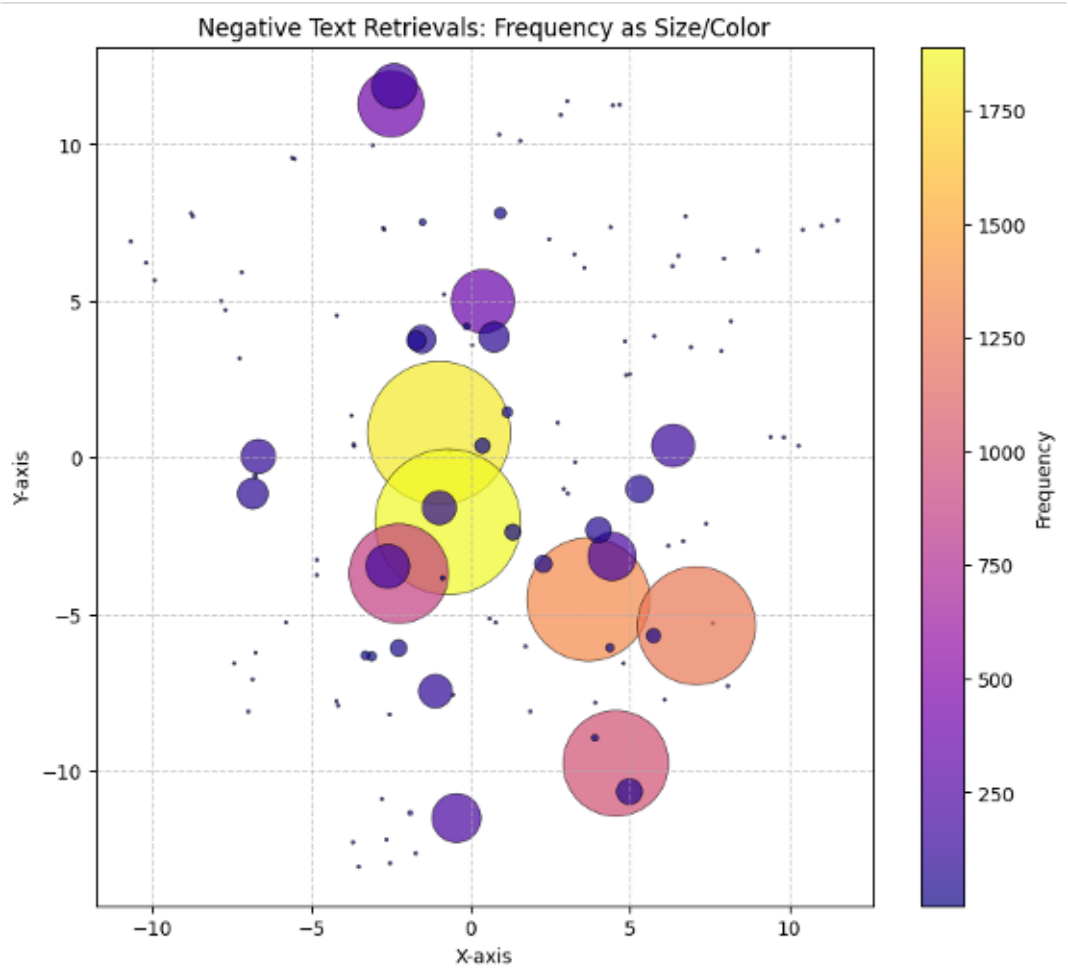


FIGURE 3.2. Retrieval frequency of the texts without RNR. Larger circles represent texts that have been retrieved more frequently. The X-axis and Y-axis represent the coordinates in a reduced-dimensionality embedding space of the retrieved negative texts.

During training, the m negative (irrelevant) texts $N = \{n_1, n_2, \ldots, n_m\}$ are replaced within the k retrieved texts $R = \{r_1, r_2, \ldots, r_k\}$ with other randomly sampled negative texts from the knowledge base. This simulates scenarios where the VLFM, during inference on unseen

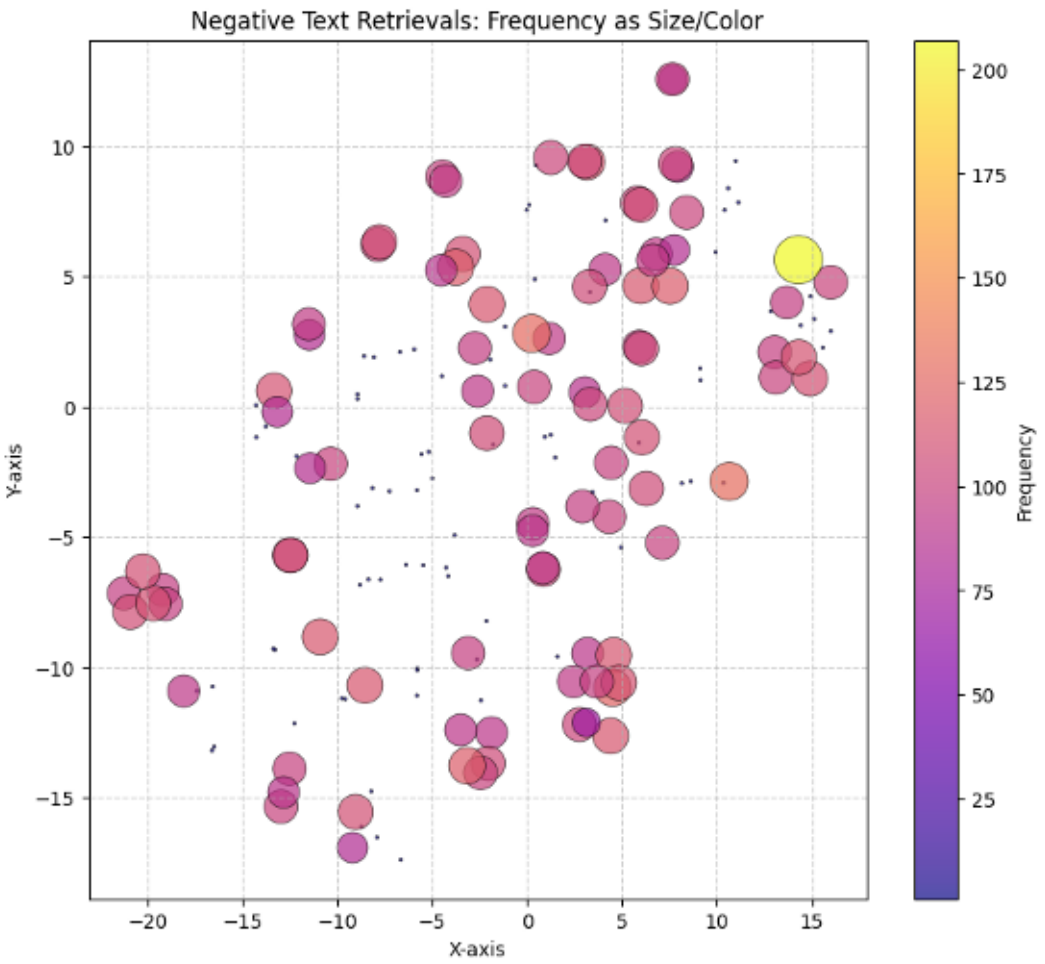


FIGURE 3.3. Retrieval frequency of the texts with RNR. The post-RNR distribution shows smaller, more evenly sized circles, indicating that the technique has successfully spread the retrieval frequency more uniformly across the knowledge base, reducing the model's reliance on a few dominant negative texts.

tumour tissues, might retrieve descriptions that are underrepresented or absent during training. Injecting these less frequent or unseen texts encourages the model to develop broader semantic awareness across the knowledge base, reducing its reliance on frequently retrieved texts biased toward the training data distribution. RNR is used only as a training-time regularisation mechanism; at inference, the model receives the unmodified top-k retrieved texts.

Subsequently, description-wise shuffling is introduced, inspired by [41]. Instead of shuffling individual words, it randomly permutes the order of entire textual descriptions within the

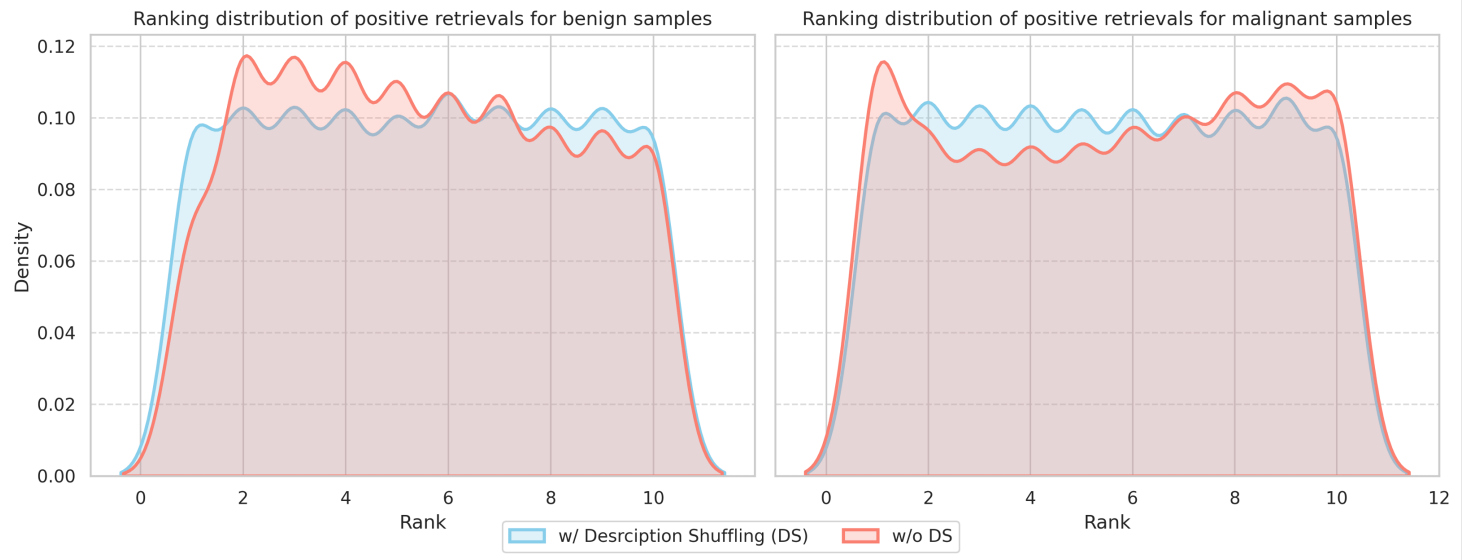


FIGURE 3.4. Comparison of ranking distribution for positive retrieved texts for benign and malignant images, with DS in blue and without DS in red.

retrieved set. This reduces the model's reliance on positional cues, forcing it to learn semantic dependencies between each description, the image content, and other retrieved texts, regardless of their order.

The texts R' after retrieval negatives replacement are represented as:

$$R' = \{r'_1, r'_2, \ldots, r'_k\} \tag{3.3}$$

$$r'_i = \begin{cases} f(r_i) & \text{if } r_i \in N \\ r_i & \text{otherwise} \end{cases} \tag{3.4}$$

where $f : R \to N$ is a function that maps a text $r_i \in R$ to a randomly sampled negative text $n \in N$.

Let $S = \{1, 2, \ldots, k\}$ be the set of indices, and $\sigma : S \to S$ be a random permutation mapping. The shuffled retrieved texts $R'' = \{r''_1, r''_2, \ldots, r''_k\}$ resulting from description-wise shuffling are represented as:

$$r''_i = r'_{\sigma(i)} \quad \text{for } i = 1, 2, \ldots, k \tag{3.5}$$

or equivalently,

$$R'' = \{r'_{\sigma(1)}, r'_{\sigma(2)}, \ldots, r'_{\sigma(k)}\} \tag{3.6}$$

This effectively reorders the retrieved texts according to the random permutation σ.

3.4 Retrieval-based De-noising Causal Language Modelling

Rather than relying on conventional image-to-text alignment that directly maps input images to target captions, our model captures the semantic dependencies between a query image, relevant (positive) texts, and irrelevant (negative) texts. This retrieval-and-de-noising design is chosen because malignancy recognition requires grounded and discriminative textual evidence: the model refines candidate descriptions from a curated pathology knowledge base instead of generating unrestricted captions from scratch. By incorporating negative samples during training, the model learns to differentiate between meaningful and misleading content, enhancing its robustness through a de-noising mechanism. To fully use the sequential reasoning capabilities of a pre-trained language model (LM), particularly for capturing the subtle visual cues that distinguish benign from malignant tumour tissues, the input is extended beyond purely textual sequences. Specifically, visual features are integrated directly into the language model's input sequence. This results in the de-noising LM module, which jointly processes visual and textual signals to suppress semantic noise and retain the most relevant information from the retrieved descriptions.

First, the image representation $h_i = H(I)$ from the output of PLIP's image encoder H is projected into the LM's language space using a trainable image-language projector Z. This projection yields an image embedding $h_l \in \mathbb{R}^{d_l}$ within the language space. The trainable image-language projector acts as a bridge between visual embeddings and the language processing space. It takes an input image representation $h_i \in \mathbb{R}^{d_i}$ and applies dropout for regularisation. Then, it passes the embedding through two linear layers with GELU activation and intermediate layer normalisation. Specifically, the projector expands the embedding to a higher dimensional space (e.g., 4 times larger than the input dimension in the experiment),

applies layer normalisation and a GELU non-linearity, and then reduces it to the desired language space dimension of the LM, again with layer normalisation. The final output h_l is the projected h_i within the LM's embedding space, ready for input to a frozen language model. The image-language projector is defined as:

$$h_l = LN\Big(Linear'\big(Dr\big(GELU\big(LN\big(Linear(Dr(h_i))\big)\big)\big)\big)\Big) \tag{3.7}$$

where $Dr(.)$ is the dropout layer, $Linear(.)$ is the first hidden linear layer with output dimension d_h, $GELU(.)$ is the GELU activation function, $LN(.)$ is the layer normalisation layer, and $Linear'(.)$ is the second linear layer that projects the hidden representation into the language space with output dimension d_l.

A multimodal token sequence $\mathcal{M}$ is constructed by concatenating the image embedding h_l with the textual tokens $\mathcal{L} = \{l_1, l_2, \ldots, l_{|L|}\} \in \mathbb{R}^{d_l}$ of the retrieved texts R'' (embedded by the LM). Here, l_i tokens can be either relevant (positive) or irrelevant (negative), and $|L|$ is the number of textual tokens. The multimodal token sequence is defined as $\mathcal{M} = [h_l, \mathcal{L}]$.

Relevance is defined at the description level. A retrieved description is treated as relevant when its benign or malignant class semantics are consistent with the query image's target class and therefore support the binary malignancy decision. The positive tokens used in the DCLM objective are the tokens originating from these relevant descriptions. This class-level labelling rule is practical for zero-shot malignancy recognition, but edge cases remain when descriptions are generic or partially applicable to both benign and malignant tissue. The knowledge base filtering step reduces such ambiguity by removing non-discriminative descriptions where possible.

Given the multimodal tokens $\mathcal{M}$, the de-noising causal language modelling objective predicts the positive tokens of relevant texts for the query image. Let P denote these positive tokens:

$$P = \{p_1, p_2, \ldots, p_{|P|}\} \in L.$$

The objective maximises the following likelihood:

$$\mathcal{LH}(\mathcal{M}) \quad = \quad \sum_{j=1}^{|P|} \log P_{\Theta}\left(p_j \mid h_l, L, p_{<j}\right) \tag{3.8}$$

where p_j is the j^{th} token of the positive (relevant) texts P, $h_l \in \mathbb{R}^{d_l}$ is the projected image token in the language space (with dimension d_l), and P_θ is the conditional probability of the frozen LM. This objective encourages the de-noising LM-based module to filter out negative context by attending to both relevant and irrelevant information, generating image-related outputs even with irrelevant or misleading retrieved text.

3.5 Image-text Retrieval and Image Classification on De-noised Texts

The refined texts resulting from the de-noising LM-based module serve as the final output for image-text retrieval. Furthermore, these de-noised texts are directly used for text-based classification of the input images. In this study, a simple majority-vote classifier is employed for predictions. Each retained de-noised description contributes a benign or malignant vote according to the class semantics it expresses, and no additional complex classifier is trained at this stage. This keeps the final decision rule transparent and ensures that the main methodological contribution remains the improvement of retrieved evidence quality. Given a set of n de-noised texts $D = \{d_1, \ldots, d_n\}$ and a set of malignancy classes $C = \{c_1, c_2\}$, let $g(d_i) \in C$ be the class assigned to each de-noised description d_i. The predicted class $\hat{c}$ for D is determined by:

$$\hat{c} = \underset{c_j \in C}{\operatorname{argmax}} \sum_{i=1}^{n} \mathbb{I}(g(d_i) = c_j) \tag{3.9}$$

where $\mathbb{I}(\cdot)$ is an indicator function. The predicted class is the class most frequently assigned to the de-noised texts by the base classifier. This choice is simple and interpretable, although borderline cases with nearly balanced retained descriptions may require confidence-weighted or learned aggregation in future work.

3.6 Inter- and Intra-Branch Min-Max Mutual Information Optimisation

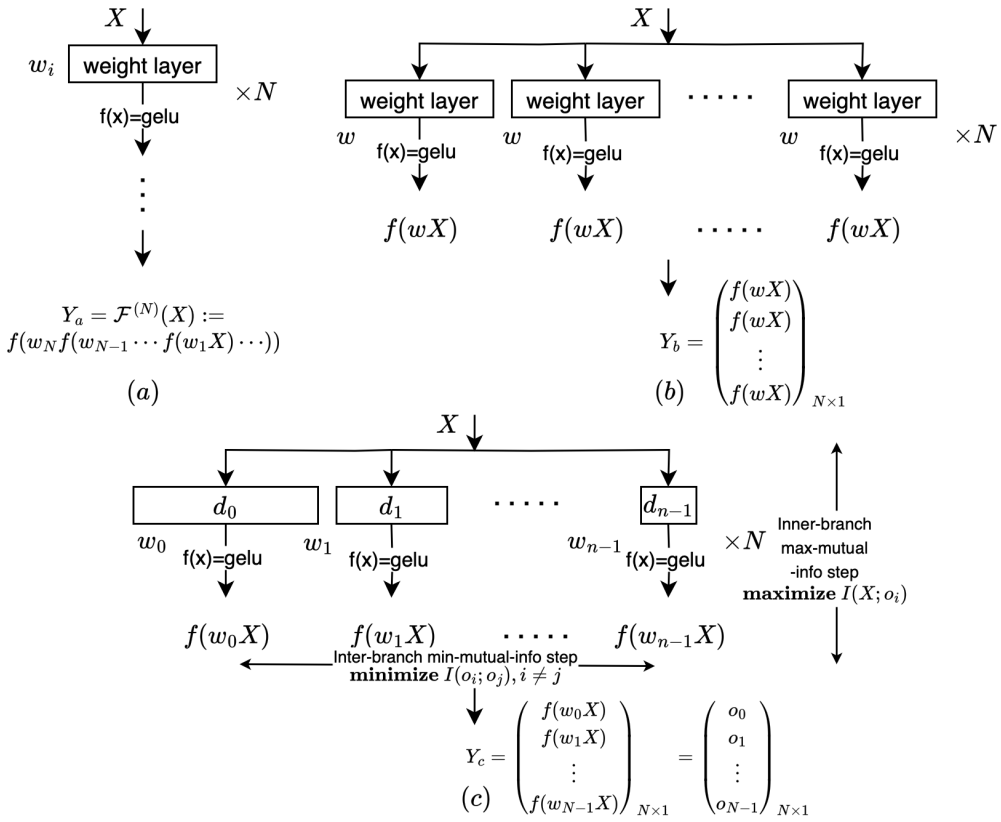


FIGURE 3.5. The comparison of the projection layers for zero-shot learning on limited training samples. (a) shows our image-text projection architecture with the proposed auxiliary branch-wise mutual information loss.

Current vision-language architectures [42, 43, 44] typically employ a shallow and simple projection mechanism for cross-modal mapping, that is, projecting image features into the language embedding space. However, such limited projection depth may reduce the model's capacity for zero-shot transfer. A common strategy to address this limitation is to increase model complexity by stacking additional non-linear layers, thereby enhancing its generalisation ability [6, 7, 45]. As shown in Figure 3.5 (a), stacking N such blocks yields the sequential transformation

$$Y_a = f\Big(w_N f\big(w_{N-1} \cdots f(w_1 X)\big)\Big) \tag{3.10}$$

The output of a single block is denoted as

$$F_i(X) = f(w_i X) \tag{3.11}$$

where w_i is the weight parameter of the i-th block and $f(\cdot)$ is the activation function and X is the output image feature from the pre-trained image encoder H.

$$Y_a = (F_N \circ F_{N-1} \circ \cdots \circ F_1)(X) \tag{3.12}$$

For brevity, we introduce the notation

$$Y_a = \mathcal{F}^{(N)}(X) \tag{3.13}$$

where $\mathcal{F}^{(N)}$ denotes the N-fold composition of weight–activation blocks $(F_N, \ldots, F_1)$. This notation emphasises that as the depth N increases, the transformation $\mathcal{F}^{(N)}$ becomes progressively more complex, thereby enabling stronger pattern extraction and enhancing the model's generalisation capability.

However, this may not hold when only small samples and few classes are available for training. A complex model architecture with a large N of weight–activation layers $F_i(.)$ can overfit easily on small samples and harm the model's zero-shot performance. Thus, deep and complex networks should be avoided for small-sample fine-tuning.

One simple solution to the complexity problem is to place the weight–activation blocks $F_i(.)$ at the same level rather than stacking them, yielding the parallel transformation shown in Figure 3.5 (b):

$$Y_b = \begin{pmatrix} f(wX) \\ f(wX) \\ \vdots \\ f(wX) \end{pmatrix}_{N \times 1} \tag{3.14}$$

where w is the shared weight used by the blocks when all branches have the same architecture. For simplicity, we assume the blocks share the same initial weights. Y_b is the concatenation of the block outputs. This transformation is evidently much less complex than that in

Equation 3.13 when the number of blocks N becomes large. Specifically, the complexity of the one-layer parallel transformation remains constant as N increases, and it is equivalent to that of the stacked transformation $\mathcal{F}^{(N)}(\cdot)$ when $N = 1$.

However, the parallel transformation may capture limited semantics due to the identical architecture of each branch, which leads to identical outputs as shown in Equation 3.14. This limitation can be mitigated by varying the layer dimensions across branches, as illustrated in Figure 3.5(c). The branch dimensions are chosen empirically to keep the projection module lightweight for limited pathology training data while encouraging complementary representations across branches; they are not claimed to be uniquely optimal.

$$Y_c = \begin{pmatrix} f(w_0 X) \\ f(w_1 X) \\ \vdots \\ f(w_{N-1} X) \end{pmatrix}_{N \times 1} \tag{3.15}$$

where w_i is the weight of the i^{th} branch. Y_c is the concatenation of the block outputs, and d_i is the dimension of the i^{th} branch. The branch dimensions are all distinct.

This design not only differentiates the outputs of individual branches but also encourages the compressed input image features h_i to be further learned in different dimensional spaces. A practical intuition is that a single projector may concentrate on dominant visual-textual patterns while overlooking weaker but still informative malignancy cues. By contrast, diverse branches encourage the model to use multiple alignment subspaces, reducing reliance on a single narrow mapping. However, employing a diverse-dimension branch architecture alone may not be sufficient for effectively learning distinct semantics. It is hypothesised that enabling each branch to extract distinct semantic representations from the prior compressed image feature h_i, thereby enhancing the subsequent projected image embedding h_l, can be beneficial for zero-shot learning.

To this end, the concept of mutual information preservation from [12] is adopted, and a *multi-branch diverse-dimension projector* equipped with *Inter- and Intra-Branch Min-Max Mutual Information Optimisation* is proposed for fine-grained image-text semantic alignment. This

mechanism encourages each branch to learn distinct semantic information from the shared compressed image feature, while maintaining consistency with the original auto-regressive objective.

The optimisation process comprises two complementary steps: a *Min-Mutual-Information* step and a *Max-Mutual-Information* step. To achieve the goal of learning diverse semantics, the Min-Mutual-Information step minimises the mutual information between different branch outputs:

$$I(o_i; o_j) = I\big(f(w_i X); f(w_j X)\big), \quad i \neq j, \tag{3.16}$$

where $o_i = f(w_i X)$ denotes the output of the i-th branch. Conversely, to prevent learning uninformative or noisy features during the Min step, the Max-Mutual-Information step maximises the mutual information between the input X and each branch output:

$$I(X; o_i) = I\big(X; f(w_i X)\big). \tag{3.17}$$

Directly computing these mutual information terms is intractable in high-dimensional spaces. Following [12, 15], a contrastive formulation is adopted that provides a tractable lower bound. Let $\widetilde{x} = \mathrm{unit}(X)$ denote the normalised input embedding, and let $\widetilde{o}_i = \mathrm{unit}(o_i)$ be the normalised output of the i-th branch. For each branch i, we treat its correspondence with the input $\widetilde{x}$ as the positive pair, while the outputs from all other branches $\{\widetilde{o}_j : j \neq i\}$ serve as negatives. The proposed branch-wise mutual information optimisation loss is then formulated as:

$$\mathcal{L}_{\text{mutual}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\big(\mathrm{sim}(\widetilde{o}_i, \widetilde{x})\big)}{\exp\big(\mathrm{sim}(\widetilde{o}_i, \widetilde{x})\big) + \sum_{j \neq i} \exp\big(\mathrm{sim}(\widetilde{o}_i, \widetilde{o}_j)\big)}, \tag{3.18}$$

where $\mathrm{sim}(.)$ denotes the similarity function. This loss simultaneously maximises the mutual information between each branch output and the input (Max step) and minimises the mutual information between different branch outputs (Min step), following the standard contrastive learning formulation.

Finally, L_{mutual} serves as an auxiliary loss for the main auto-regressive loss L_{dclm} (de-noising causal language modelling loss) in Section 3.4, and a weighted sum is employed over the two

losses.

$$L = (1 - \alpha)L_{dclm} + \alpha L_{mutual} \tag{3.19}$$

3.7 De-noising Generated Descriptions via Auxiliary Precision Reward Loss (PRL)

Despite its ability to reduce noise and irrelevant descriptions in the retrieved text, the proposed method—which adopts a causal language modelling objective—still struggles to completely eliminate such noise. As an auxiliary refinement rather than the central methodological contribution, this section evaluates a precision-oriented reward term inspired by reinforcement learning (RL), where a reward function evaluates the quality of the generated text to guide gradient optimisation.

To further refine the textual outputs, a *precision reward loss* is introduced as an auxiliary objective. This loss function aims to enhance the de-noising process by explicitly considering the ratio between the number of relevant generated samples and the total number of generated samples for each query image.

The precision reward function f_{pr} is computed as:

$$f_{pr} = \frac{N_{rel}}{N_{total}} \tag{3.20}$$

where N_{total} is the total number of retrieved texts and N_{rel} is the number of relevant texts among the retrieved texts.

Our objective is to maximise the number of relevant samples within the generated textual descriptions, and the precision reward function follows the same trend. To facilitate gradient-based optimisation, the negative logarithm is applied to the precision reward, yielding the final loss formulation. Accordingly, the precision reward loss L_{pr} is defined as:

$$L_{pr} = -\log(f_{pr}) \tag{3.21}$$

A weighted sum is then employed over the main loss L_{dclm} (de-noising causal language modelling loss) in Section 3.4 and the proposed auxiliary loss L_{pr}.

$$L = (1 - \beta)L_{dclm} + \beta L_{pr} \tag{3.22}$$

Algorithm 1 Retrieval-based De-noising Causal Language Modelling Framework

Require: Vision-Language Foundation Model $\mathcal{V}$ (image encoder H, image/text projectors E_i, E_t),
Frozen language model $\mathcal{LM}$, large language model for knowledge base generation $\mathcal{LLM}$,
Training image set $\mathcal{I}_{train}$, pathology class set $\mathcal{C} = \{\text{benign, malignant}\}$,
hyperparameters: top-k retrieval range K, RNR replacement count m, training epochs T.
Ensure: Trainable image-language projector Z.
Stage 1: Construct pathology knowledge base
Prompt $\mathcal{LLM}$ to generate descriptive sentences for each class $c \in \mathcal{C}$.
Filter knowledge base to remove descriptions that are non-discriminative across classes.
$\mathcal{KB} \leftarrow \{t_1, \ldots, t_{|\mathcal{KB}|}\}$ (each t is a short fine-grained description)

Stage 2: Precompute knowledge base embeddings
for all $t \in \mathcal{KB}$ **do**
    $e_t \leftarrow E_t(t)$ ▷ project text via VLFM text projector
end for

Stage 3: Train retrieval de-noising module
for $epoch \leftarrow 1$ to T **do**
    for all I in minibatches of $\mathcal{I}_{train}$ **do**
        $h_i \leftarrow H(I)$ ▷ image feature from VLFM
        $q \leftarrow E_i(h_i)$ ▷ VLFM image projector (or image embedding used for retrieval)
        $\{r_1, \ldots, r_k\} \leftarrow \text{Top-}k\_by\_cosinesim(q, \{e_t\}_{t \in \mathcal{KB}}, k \in K)$

        Apply Retrieval Negatives Replacement (RNR)
        Select m indices $\mathcal{M} \subset \{1, \ldots, k\}$ as negatives
        for $j \in \mathcal{M}$ **do**
            $r_j \leftarrow$ random negative embedding sampled from $\mathcal{KB} \setminus \{r_1, \ldots, r_k\}$
        end for

        Apply Description-wise Shuffling (DS)
        Permute $\{r_1, \ldots, r_k\}$ by random permutation σ to obtain $\{r'_1, \ldots, r'_k\}$

        Project image and construct LM input
        $h_\ell \leftarrow Z(h_i)$ ▷ trainable image-language projector
        Convert each retrieved embedding r'_j back (or map) to text tokens L_j for $\mathcal{LM}$ input
        Construct multimodal token sequence $M \leftarrow [h_\ell; L_1; \ldots; L_k]$

        Optimise de-noising causal LM objective (DCLM)
        $\mathcal{L}_{\text{dclm}} \leftarrow -\sum_{t \in P} \log P_{\mathcal{LM}}(t \mid M, t_<)$ ▷ maximise likelihood of positive (relevant) tokens given M
        Update trainable parameters (e.g., Z) via AdamW on $\mathcal{L}$
    end for
end for

Stage 4: Inference (zero-shot retrieval & classification)
Require: Test image I^*
$q^* \leftarrow E_i(H(I^*))$
Retrieve top-k descriptions $\{r_1^*, \ldots, r_k^*\}$ from $\mathcal{KB}$
No RNR&DS is applied at inference time
$M^* \leftarrow [Z(H(I^*)); \text{tokens}(r_1^*), \ldots]$
Use $\mathcal{LM}$ to produce de-noised texts $\{d_1, \ldots, d_n\}$
Classify via majority vote over base classifier outputs on $\{d_i\}$
return predicted class

CHAPTER 4

Experimental Setup and Results

This chapter presents the experimental design, implementation details, and quantitative as well as qualitative evaluations of the proposed Retrieval-based De-noising Causal Language Modelling (RDCLM) framework. The experiments are conducted to validate the effectiveness of RDCLM in enhancing zero-shot image-text retrieval and image classification performance within the domain of histopathology. Specifically, this chapter outlines the datasets used for training and evaluation, the evaluation metrics adopted to assess both retrieval and classification performance, and the implementation settings employed to ensure fair and reproducible comparisons with baseline methods. Comprehensive quantitative results are provided on multiple histopathology datasets, comparing RDCLM against a range of state-of-the-art vision–language and retrieval-based models. In addition, ablation studies are performed to examine the contribution of each key component, including the retrieval negatives replacement and description-wise shuffling (RNR&DS) operations, the de-noising causal language modelling objective, the multi-branch diverse-dimension projection architecture, and the auxiliary precision reward loss. Finally, qualitative analyses are included to illustrate how RDCLM effectively refines noisy retrieved texts and improves malignancy recognition in challenging zero-shot scenarios.

4.1 Dataset and Evaluation Metrics

This study evaluates the methods using five histopathological cancer datasets:

- BreakHis [20] (breast, >9K H&E images, 40-400X magnifications, 8 subtypes, 82 patients)
- LC25000 [18] (lung/colon, 5 subclasses, 5K images/subclass)
- CRC-HE-7K [17] (colon, 9 subclasses, 7180 patches, 50 patients)
- EBHI-Seg [19] (colorectal, 5170 H&E images, 6 stages)
- Camelyon17 [46] (breast, 1,399 H&E whole-slide images (WSIs) of sentinel lymph nodes, for classification of metastases (No metastases, Isolated tumour cells (ITC), Micro-metastases, Macro-metastases))

Training Dataset	
Dataset (Category)	**Classes**
BreakHis [20] (Breast cancer)	Fibroadenoma, Phyllodes tumour, Tubular adenoma, Ductal carcinoma, Lobular carcinoma, Mucinous carcinoma
Testing Datasets	
BreakHis [20] (Breast cancer)	Adenosis, Papillary carcinoma
LC25000 [18] (Lung/Colon cancer)	Colon adenocarcinoma, Benign colonic tissue, Lung adenocarcinoma, Lung squamous cell carcinoma, Benign lung tissue
CRC-HE-7K [17] (Colon cancer)	Smooth muscle, Normal colon mucosa, Colorectal adenocarcinoma epithelium
EBHI-Seg [19] (Colon cancer)	Low-grade intraepithelial neoplasia, Serrated adenoma, Adenocarcinoma
Camelyon17 [46] (Breast cancer)	Present tumour and no-tumour

TABLE 4.1. Datasets Used for Training and Evaluation

The models were fine-tuned on a data-augmented subset (20K samples, 3 benign/3 malignant subtypes) of BreakHis and then tested via zero-shot image-text retrieval and image classification on unseen subtypes from LC25000, EBHI-Seg, Camelyon17, BreakHis, and CRC-HE-7K (see Table 4.1). The 20K BreakHis subset was selected as a practical fine-tuning set that preserved benign/malignant balance while retaining morphological diversity. It included three

benign and three malignant BreakHis subtypes across multiple magnifications (40X–400X), exposing the model to variation in both subtype appearance and image scale. To reduce subtype-specific and magnification-related confounding, the BreakHis training and testing subtype sets were kept disjoint: RDCLM was trained on six BreakHis subtypes and evaluated on the remaining unseen BreakHis subtypes, as well as on external datasets.

The cross-dataset zero-shot evaluations are source-disjoint by construction, since the model is trained on BreakHis and evaluated on separate external datasets. For the BreakHis setting, the split is subtype-disjoint rather than a random within-subtype split. Patient-level or slide-level separation was not explicitly enforced within BreakHis; therefore, the BreakHis same-dataset result should not be interpreted as a formal patient-level leakage-controlled evaluation.

The Precision@k metric is used from information retrieval to assess image-text retrieval performance. It is calculated as:

$$\text{Precision@k} = \frac{|\text{Relevant} \cap \text{Retrieved}|}{|\text{Retrieved}|} \equiv \frac{N_{rel}}{k} \tag{4.1}$$

where k is the total number of retrieved texts and N_{rel} is the number of relevant texts among the retrieved texts.

The recall metric was not used, because our approach aims to reduce irrelevant texts while maintaining the number of relevant texts. Recall, which measures the proportion of retrieved relevant texts relative to the total number of relevant texts in the dataset, is unsuitable for this evaluation. Since our method does not change the number of relevant texts retrieved, recall remains unchanged and provides no additional insight into retrieval performance.

For classification, F1-score and accuracy are used as the evaluation metrics.

4.2 Implementation Setup

Implementation. PLIP [1] is used as the base VLFM and the backbone for other baseline approaches. PLIP was selected because it is a pathology-specific CLIP-style model and therefore provides a strong foundation for histopathology image-text retrieval. The aim of this

thesis is not to benchmark all possible retrieval backbones, but to evaluate whether de-noising retrieved textual evidence improves zero-shot malignancy recognition when built on top of a competent pathology retriever. Broader evaluation with additional pathology foundation models remains an important direction for future work. FLAN-T5 [47] served as the LM in our de-noising LM-based module. The image-language projector is trained, freezing all other components, including the LM and VLFM. The model was trained using the AdamW optimiser with a learning rate of 0.05, betas of (0.9, 0.98), epsilon of 1e-6, and a weight decay of 1e-06. A Cosine Annealing Warm Restarts scheduler was employed with an initial cycle length of 5 epochs and a cycle multiplier of 1, resetting the learning rate after each 5-epoch cycle. Training spanned 50 epochs, with a batch size of 5 and gradient accumulation over 3 iterations to simulate a larger effective batch size. A dropout rate of 0.39 was applied to mitigate overfitting. For top-k retrieval, k was randomly selected from 5 to 10 in each training iteration.

Baselines. To evaluate our method's effectiveness, the proposed method is compared against established approaches from the literature, covering CLIP-based, retrieval-based, and unimodal (text-only and image-only) methodologies. Specifically, the evaluations include:

- **CLIP-A-self** [21]: A CLIP-based method using a novel self-attention adapter, fine-tuned on visually descriptive captions generated by GPT-4.
- **CLIP$^{\text{FT}}$+A** [22]: Fine-tunes CLIP using LLM-refined attribute-rich captions and introduces a contrastive loss to mitigate limited class diversity in mini-batches.
- **CaSED** [23]: A retrieval-based method performing training-free prediction by computing semantic similarity between candidate labels and a joint representation of the input image and retrieved captions.
- **PLIP** [1]: The base model used in our study, a CLIP-derived model fine-tuned using a ViT-based image encoder and text encoder on a large dataset of pathology image-text pairs from medical Twitter.
- **PLIP$^{\text{FT}}$**: Fine-tuned projection layers of PLIP with frozen encoders.

- **ViT** [7]: A unimodal vision Transformer, fine-tuned solely on the pathology images. PLIP's ViT-based image encoder is used as the vision backbone, and a classifier is fine-tuned following the original ViT fine-tuning procedure.
- **FLAN-T5** [47]: An instruction fine-tuned Transformer-based language model. The model's weights are frozen, and LoRA adapters [48] are fine-tuned on retrieved text to classify textual class labels for input images.

The baseline set was chosen according to three principles: relevance to zero-shot pathology classification, strength and recognition in prior literature, and suitability for fair comparison under the same benign/malignant pathology evaluation setting. Based on these criteria, the comparison includes representative CLIP-style, retrieval-based, adaptation-based, and unimodal baselines. This design allows the evaluation to test whether RDCLM's gains arise specifically from improving retrieved textual evidence, rather than from an unrelated change in backbone, modality, or task formulation.

CHAPTER 5

Results

This chapter presents the quantitative and qualitative evaluations of the proposed Retrieval-based De-noising Causal Language Modelling (RDCLM) framework. The experiments are conducted to validate the effectiveness of RDCLM in enhancing zero-shot image-text retrieval and image classification performance within the domain of histopathology. Specifically, this chapter outlines the retrieval and classification performance, and the implementation settings employed to ensure fair and reproducible comparisons with baseline methods. Comprehensive quantitative results on multiple histopathology datasets compare RDCLM against a range of state-of-the-art vision–language and retrieval-based models. In addition, ablation studies are performed to examine the contribution of each key component, including the retrieval negatives replacement and description-wise shuffling (RNR&DS) operations, the de-noising causal language modelling objective, the multi-branch diverse-dimension projection architecture, and the auxiliary precision reward loss. Finally, qualitative analyses are included to illustrate how RDCLM effectively refines noisy retrieved texts and improves malignancy recognition in challenging zero-shot scenarios.

5.1 Results for Zero-Shot Image-Text Retrieval

Image-text retrieval performance was evaluated using precision at 5, 10, and 15, focusing on CLIP-based methods due to the cross-modal nature of retrieval. For image classification, the proposed method is compared against CLIP-based, retrieval-based, and unimodal methods. To account for the sensitivity of vision-language models to class prompt selection and the number k of retrieved texts in the classification, the results of CLIP-based methods across two prompting techniques were averaged: fixed prompt templates [1] and ensemble prompting [3] (mean embedding of class descriptions). For retrieval-based classification approaches, results were averaged across k values from 5 to 10 for the top-k retrieved items.

Table 5.1 presents a comparative analysis of zero-shot image-text retrieval performance. Our proposed RDCLM method achieved competitive performance. Notably, RDCLM demonstrates a substantial improvement in precision, especially at lower retrieval ranks (P@5 and P@10), indicating more effective retrieval of relevant items within the top results. However, the de-noising performance slightly degraded when the number of retrieved texts is large.

Most notably, RDCLM achieved the highest average precision across all datasets, with a 3.1% average improvement over the second-best results for each retrieval rank. This highlights the robustness and generalisability of our de-noising LM-based approach to unseen characteristics of pathology data, compared to conventional methods that rely on learning contrastive embedding spaces.

Furthermore, RDCLM significantly enhanced the baseline PLIP model's retrieval capacity. In contrast, fine-tuning PLIP$^{\text{FT}}$ directly on the same dataset led to performance degradation in the zero-shot scenario, likely due to over-fitting. This demonstrates RDCLM's ability to effectively adapt to unseen patterns and generalise well by leveraging the strong sequential dependency learning capacity of pre-trained LMs, even with limited fine-tuning data (over 10 times smaller than PLIP's training dataset size).

Methods	CRC-HE-7K			EBHI-Seg			LC25000			BreakHis			Camelyon			Average		
	P@5	P@10	P@15	P@5	P@10	P@15	P@5	P@10	P@15	P@5	P@10	P@15	P@5	P@10	P@15	P@5	P@10	P@15
PLIP	0.577	0.566	0.539	0.499	0.484	0.509	0.503	0.520	0.544	0.545	0.544	0.540	0.478	0.484	0.469	0.520	0.520	0.520
PLIP^{FT}	0.454	0.475	0.461	0.425	0.436	0.451	0.484	0.493	0.502	0.544	0.553	**0.555**	0.508	0.455	0.461	0.483	0.482	0.486
CLIP-A-self	0.580	0.589	0.566	0.541	0.499	0.484	0.504	0.481	0.510	**0.603**	**0.570**	0.555	0.478	0.484	0.469	0.541	0.523	0.517
CLIP^{FT}+A	0.563	0.543	0.520	0.492	0.481	0.508	0.520	0.506	0.541	0.517	0.518	0.521	0.449	0.489	**0.481**	0.508	0.507	0.515
RDCLM (proposed)	0.638	0.632	0.637	0.594	0.607	0.587	0.495	0.570	0.577	0.540	0.518	0.519	0.508	0.475	0.462	0.543	0.560	0.556
RDCLM + PRL	**0.639**	**0.642**	**0.638**	**0.642**	**0.639**	**0.615**	**0.577**	**0.610**	**0.608**	0.554	0.543	0.540	**0.509**	**0.490**	**0.481**	**0.578**	**0.601**	**0.573**

TABLE 5.1. Zero-shot image-text retrieval on the histopathology datasets with unseen subclasses, evaluated by the Precision@k metric (k=5, 10, 15). The **best** and second-best results are highlighted in bold and underlining.

Methods	CRC-HE-7K		EBHI-Seg		LC25000		BreakHis		Camelyon		Average	
	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy
PLIP	0.623	0.491	0.548	0.455	0.670	0.523	0.494	0.514	**0.646**	**0.593**	0.596	0.515
ViT (PLIP backbone)	0.711	0.573	0.212	0.487	0.465	0.516	0.536	0.504	0.519	0.357	0.488	0.487
PLIP^{FT}	0.382	0.327	0.424	0.378	0.627	0.502	0.517	0.563	0.369	0.447	0.463	0.443
FLAN-T5-large	0.557	0.497	0.516	0.490	0.561	0.523	0.496	0.501	0.516	0.503	0.530	0.503
CLIP-A-self	0.653	0.499	0.430	0.385	0.665	0.512	0.577	0.489	0.617	0.479	0.588	0.473
CLIP^{FT}+A	0.738	0.603	0.475	0.527	0.690	0.623	0.538	0.490	0.540	0.489	0.596	0.547
CaSED	0.286	0.448	0.226	0.387	0.178	0.343	0.496	**0.588**	0.361	0.321	0.310	0.418
RDCLM (proposed)	**0.761**	**0.620**	**0.695**	**0.739**	**0.756**	**0.673**	**0.608**	0.569	0.636	0.469	**0.691**	**0.614**

TABLE 5.2. Zero-shot image classification on the histopathology datasets assessed by F1-score and accuracy metrics. Additional retrieval-based and unimodal methods are compared. Bold indicates best and underlined indicates second-best results per metric.

5.2 Results for Zero-Shot Image Classification

Table 5.2 shows the zero-shot image classification results, comparing our approach with other state-of-the-art (SOTA) methods.

RDCLM consistently outperformed other methods across most datasets and evaluation metrics, achieving the highest average scores. Specifically, RDCLM achieved a significant improvement in F1-score and accuracy over the second-best competitor.

This confirms RDCLM's effectiveness against unimodal, multimodal retrieval-based, and CLIP-based methods. The superior performance is attributed to: (1) robust learning of cross-modal semantic alignment through negative semantics filtering, and (2) the efficacy of using pre-trained LMs to de-noise retrieved texts, yielding refined textual information that significantly improves image classification accuracy.

5.3 Discussion

1. Comparison of $CLIP$-based Methods to $PLIP$

The observation that certain $CLIP$-based methods, such as **CLIP^{FT}+A** and **CLIP-A-self**, can outperform the base $PLIP$ model—despite $PLIP$ being pre-trained on pathology data—suggests that textual alignment and knowledge refinement can be more influential than the domain-specificity or scale of pre-training data in zero-shot settings. The base $PLIP$ model is fine-tuned on pathology image–text pairs that are often collected from less-controlled sources (e.g., online medical forums), where descriptions can be coarse-grained and omit fine-grained visual cues needed to distinguish subtle benign versus malignant tissue patterns. As a result, the learned image–text embedding space may contain substantial semantic noise. In contrast, **CLIP^{FT}+A** and **CLIP-A-self** leverage retrieval pseudo-captions and optimised contrastive loss, improving the semantic specificity of the text and strengthening cross-modal alignment, which benefits both retrieval (e.g., P@k) and classification (e.g., F1-score).

2. Efficacy of $RDCLM$ and Variation Across Datasets

How $RDCLM$ Improved Performance

The substantial improvement achieved by $RDCLM$ (e.g., an average 9.5% F1-score gain and 6.7% accuracy gain over the second-best competitor in classification) is primarily attributed to its Retrieval-based De-noising Causal Language Modelling strategy, which targets a key bottleneck of pathology VLFMs: semantic noise in retrieved texts. The most common noise patterns observed in the retrieved evidence pool were wrong-class morphology descriptions, generic cancer-related text with weak discriminative value, and repeated negative retrieval patterns from the base retriever. Rather than only refining prompts or visual features, $RDCLM$ trains a LM-based de-noising module to identify and suppress irrelevant or non-discriminative tokens while preserving task-relevant ones within the retrieved descriptions. This token-level de-noising yields a cleaner, more semantically aligned set of de-noised texts for subsequent text-based classification. The approach is supported by a pathology-specific knowledge base containing LLM-generated, discriminative descriptions of tumour features, and further strengthened by augmentations such as RNR and DS to encourage robustness and reduce reliance on dataset-specific biases and positional cues.

Variation of Results Across the Datasets

Performance varies across datasets, with EBHI-Seg consistently achieving the strongest results (e.g., Avg P@k ≈ 0.63 and F1/Acc ≈ 0.72) and BreakHis and Camelyon the lowest (e.g., Avg P@k ≈ 0.49 and F1/Acc ≈ 0.55). This heterogeneity likely reflects differences in visual complexity, intra-class variability, and how well each dataset aligns with the LLM-generated knowledge base used by $RDCLM$. EBHI-Seg focuses on well-defined colorectal cancer stages (intraepithelial neoplasia, adenoma, adenocarcinoma), for which generic malignancy-related descriptors appear to align well with the distinct morphological cues, making zero-shot discrimination more tractable. In contrast, BreakHis contains many unseen histological subtypes across a wide magnification range (40X–400X), introducing substantial appearance variation and requiring generalisation to out-of-distribution subtype patterns, which lowers

performance across metrics. One encouraging result is the strong performance on LC25000 despite its organ-level difference from the BreakHis training data; however, this should be interpreted cautiously because dataset composition, staining variation, and organ-specific appearance can still contribute to the remaining correction errors.

Why Binary Malignancy Recognition Was Chosen as the Primary Task

This thesis focuses on binary benign-versus-malignant recognition rather than fine-grained subtype classification because malignancy status provides a more consistent and clinically meaningful target across heterogeneous pathology datasets. The benign-versus-malignant distinction is shared across BreakHis, LC25000, CRC-HE-7K, EBHI-Seg, and Camelyon17, even though these datasets differ in organ type, subtype definitions, staining, and image source. Subtype labels are often dataset-specific, unevenly distributed, and less directly comparable across organs, which makes them less suitable for evaluating whether RDCLM improves zero-shot semantic alignment under distribution shift. Establishing effectiveness on a robust binary task is therefore a necessary first step before extending the framework to finer-grained subtype prediction.

3. Differences Between P@k Metrics and Their Correlation with Classification

Differences Between P@k and Underlying Causes

Analysing Precision@k (P@5, P@10, P@15) for $RDCLM$ relative to the baseline $PLIP$ provides insight into the DCLM's filtering dynamics. While standard retrieval models typically show decreasing precision as k increases (P@5 $>$ P@10 $>$ P@15), $RDCLM$ sometimes exhibits stability or a mild anomaly where P@10 is comparable to or marginally higher than P@5 on certain datasets (e.g., CRC-HE-7K, LC25000). This suggests that, when the top-5 pool contains noisy or partially relevant items, expanding to $k = 10$ can introduce additional descriptions that the de-noising module can selectively retain, thereby improving the effective relevance of the aggregated set. The eventual degradation at P@15 indicates a

limit to the de-noising capacity: as k grows, the proportion of irrelevant content increases and can overwhelm the de-noising module's ability to reliably separate relevant from noisy tokens.

P@k Validity for Classification

P@k is a meaningful proxy for classification quality in this framework because classification is performed via majority vote over the (de-noised) retrieved texts. Therefore, higher P@k implies a cleaner and more relevant set of de-noised texts, which yields a more reliable voting signal and tends to improve final classification accuracy and F1-score.

4. Cross-organ Generalisation and Dataset Bias

The cross-organ transfer observed in this thesis should be interpreted cautiously. The model is not assumed to transfer breast-specific appearance directly to colon or lung tissue. Rather, the intended transferable signal is the binary benign-versus-malignant semantic distinction encoded in the retrieved descriptions and reinforced by de-noising. Empirically, RDCLM is trained on selected BreakHis breast subtypes and evaluated both on unseen BreakHis subtypes and on external colon, lung, and lymph-node datasets, which reduces the likelihood that the results are explained only by memorisation of the training subtypes. Mechanistically, RDCLM is trained to prefer class-relevant textual descriptions associated with malignancy-related morphology, not to memorise organ-specific subtype templates. This mechanism can support transfer when malignancy cues are expressed in semantically compatible ways across datasets, but it is likely to weaken when organ-specific morphology dominates, when the knowledge base lacks sufficient semantic coverage for the target organ, or when retrieval is biased toward breast-centric descriptions. For this reason, the current results support cross-dataset generalisation as an encouraging empirical observation, but not as proof of organ-invariant pathology understanding.

5.4 Ablation Study

Main Components. An ablation study is conducted to evaluate the contribution of each component of the proposed RDCLM approach in zero-shot image classification tasks. The main components are: VLFM (PLIP), image-language projector, retrieval negatives replacement & description-wise shuffling (RNR&DS), de-noising causal language modelling (DCLM), and pre-trained LM (FLAN-T5-large).

Table 5.3 shows that all components contributed to enhancing the zero-shot classification performance. The RDCLM+PRL variant achieved the best results, suggesting that the auxiliary precision reward loss can be beneficial, while the core RDCLM components provide the main performance gains. The **PLIP-I2L projector** and **PLIP image branch** methods, using either fine-tuned or pre-trained projectors, performed comparatively poorly in malignancy determination on unseen tumour tissues, sometimes with accuracy below random chance. Ablating DCLM (**RDCLM w/o DCLM**), and using conventional causal language modelling (CLM) to generate relevant texts, yielded results below 20% in both metrics. This supports the choice of a retrieval-and-de-noising design over direct image-caption generation: unrestricted generation is less grounded in the curated pathology knowledge base and is less effective for retaining discriminative benign-versus-malignant evidence. Conversely, introducing noisy retrieved texts into the cross-modal feature alignment process, enabling the LM to de-noise the retrieved text, was able to effectively map query images and corresponding textual information with reduced noise (irrelevant descriptions) through the proposed **RDCLM** approach. Removing RNR&DS (**RDCLM w/o RNR&DS**) dramatically reduced performance, highlighting the importance of replacing irrelevant textual semantics and description-wise shuffling for good generalisation on out-of-distribution pathology characteristics.

Method	CRC-HE-7K		EBHI-Seg		LC25000		BreakHis		Camelyon		Average	
	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy
PLIP image branch	0.711	0.573	0.212	0.487	0.465	0.516	0.536	0.504	0.519	0.357	0.488	0.487
PLIP-I2L projector	0.382	0.327	0.424	0.378	0.627	0.502	0.517	0.563	0.369	0.447	0.463	0.443
RDCLM w/o DCLM	0.632	0.473	0.323	0.529	0.695	0.609	0.532	0.484	0.489	0.324	0.534	0.484
RDCLM w/o RNR&DS	0.057	0.036	0.058	0.033	0.053	0.030	0.049	0.038	0.023	0.012	0.054	0.034
RDCLM	0.761	**0.640**	0.553	0.628	0.674	0.583	0.601	0.529	0.646	**0.482**	0.647	0.572
RDCLM+PRL	**0.770**	0.631	**0.684**	**0.647**	**0.729**	**0.618**	**0.641**	**0.548**	**0.648**	0.480	**0.695**	**0.585**

TABLE 5.3. Ablation study on components of the proposed RDCLM in zero-shot image classification. **Bold** indicates the best and underlined the second-best performance per dataset and metric.

LM components	CRC-HE-7K		EBHI-Seg		LC25000		BreakHis		Camelyon		Average	
	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy
T5-base	0.743	0.595	0.444	0.590	0.728	0.623	**0.610**	**0.590**	**0.614**	0.446	0.628	0.569
T5-large	**0.761**	**0.640**	0.553	0.628	0.674	0.583	0.601	0.529	0.646	**0.482**	0.647	0.572
T5-xl	**0.761**	0.620	**0.695**	**0.739**	**0.756**	**0.673**	0.608	0.569	0.636	0.469	**0.691**	**0.614**

TABLE 5.4. Ablation study on FLAN-T5 language models of the proposed RDCLM in zero-shot image classification.

Knowledge base setting	CRC-HE-7K		EBHI-Seg		LC25000		BreakHis		Camelyon		Average	
	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy
original_descrip	0.733	0.579	0.610	0.632	0.703	0.585	0.586	0.530	0.625	0.456	0.651	0.556
shorten_descrip	0.756	0.620	0.600	0.637	0.689	0.596	**0.626**	**0.539**	0.648	0.479	0.664	0.574
orig_and_shorten_descrip	**0.765**	**0.622**	**0.700**	**0.690**	**0.731**	**0.620**	0.617	0.522	**0.672**	**0.506**	**0.697**	**0.592**

TABLE 5.5. Ablation study on knowledge base design choices in zero-shot image classification. We compare the original LLM-generated descriptions, shortened descriptions, and a hybrid knowledge base combining both.

RNR Ratio	CRC-HE-7K		EBHI-Seg		LC25000		BreakHis		Camelyon		Average	
	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy
ratio = 0.0 (w/o RNR)	0.752	**0.631**	0.577	0.666	0.697	0.608	0.604	0.532	0.637	0.473	0.653	0.582
ratio = 0.2	0.758	0.619	0.694	**0.719**	0.704	0.598	0.626	0.542	0.639	0.471	0.684	0.590
ratio = 0.4	0.759	0.621	0.609	0.673	0.733	0.625	0.597	0.523	0.665	0.499	0.673	0.588
ratio = 0.6	0.758	0.626	0.632	0.673	0.710	0.622	**0.627**	**0.554**	0.671	**0.513**	0.680	**0.598**
ratio = 0.8	0.754	0.607	0.495	0.619	**0.752**	**0.659**	0.607	0.547	0.646	0.478	0.651	0.582
ratio = 1.0 (w/ RNR)	**0.765**	0.622	**0.700**	0.690	0.731	0.620	0.617	0.522	**0.672**	0.506	**0.697**	0.592

TABLE 5.6. Evaluation of the effect of replacement ratio on performance under Retrieval Negatives Replacement (RNR).

FLAN-T5 Language Models. Table 5.4 shows the impact of different FLAN-T5 language models. Deeper pre-trained LMs enhance the model's ability to determine the malignancy of unseen tumour tissues, with the largest LM, FLAN-T5-xl, achieving the highest scores.

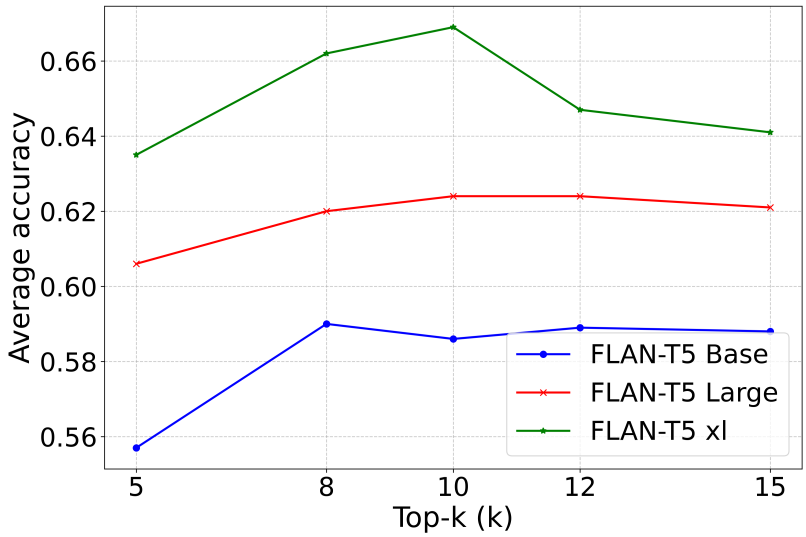


FIGURE 5.1. Ablations on the number of retrieval texts k with different sizes of FLAN-T5 models.

Value of k in Top-k retrieval. Figure 5.1 illustrates the effect of the number of retrieved texts (k) on performance. The zero-shot image classification (ZSIC) performance generally improved as k increased from 10 to 15 (the range used during training) for models with various LM sizes. Performance shows slight variations after $k = 8$. This trend is reasonable because retrieving more texts (larger k) increases the likelihood of including more relevant texts, which aids in aligning correct textual content with the visual semantics of the query images.

Knowledge base design choices. The structure of the pathology knowledge base was examined through the variant study shown in Table 5.5. Knowledge bases larger than 300 descriptions were not tested in this thesis, so the results should not be interpreted as measuring how performance scales with larger knowledge bases. Instead, three knowledge base settings were compared to choose the final setting empirically: (1) an original LLM-generated knowledge base with 150 total descriptions across the benign and malignant classes,

(2) a shortened-description knowledge base with the same number of descriptions, and (3) a hybrid knowledge base combining both sets, giving a final size of approximately 300 descriptions.

As shown in Table 5.5, the original 150-description knowledge base achieved lower average performance (F1-score = 0.651, accuracy = 0.556), while the shortened 150-description knowledge base improved the results (F1-score = 0.664, accuracy = 0.574). The hybrid 300-description knowledge base achieved the best overall performance (F1-score = 0.697, accuracy = 0.592). Therefore, the choice of approximately 300 descriptions was not arbitrary, but was the best-performing setting among the knowledge base variants tested in this thesis.

The aim of the knowledge base design was to include informative and distinct malignancy-related descriptions that support learning discriminative morphological differences between benign and malignant tumour tissues. The knowledge base was not enlarged further because additional descriptions were expected to introduce more redundancy or weakly discriminative text, which could distract the model and increase retrieval cost. A controlled size-sensitivity study with several larger knowledge bases remains future work.

Effect of replacement ratio under RNR. We further evaluated whether the degree of negative replacement affects the model's performance. As shown in Table 5.6, introducing negative replacement during training generally improves performance over the no-replacement setting (ratio $= 0.0$). In particular, all non-zero ratios yield higher average F1-scores than ratio $= 0.0$, improving from 0.653 to a range of 0.673–0.697. The best average F1-score is achieved at ratio $= 1.0$, whereas the best average accuracy is obtained at ratio $= 0.6$ (0.598). These results indicate that the benefit of RNR is not tied to a single highly sensitive ratio choice; rather, exposing the model to perturbed negative retrieval patterns during training consistently improves robustness. At the same time, the trend is not strictly monotonic across ratios or datasets, suggesting that stronger replacement does not uniformly optimise all metrics. Overall, the results support the view that the train-time perturbation introduced by RNR acts as beneficial regularisation, helping the model generalise better to unseen retrieval noise instead of causing harmful train–test mismatch.

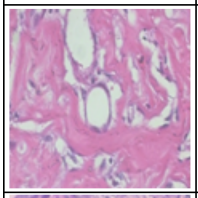
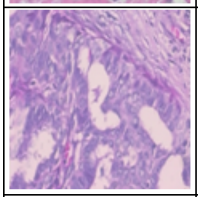
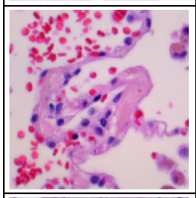
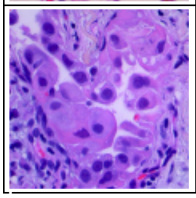
Image	Original retrieved texts	De-noised texts	GT	$P_{retrieved}$	$P_{denoised}$
	'Immune cells are present', 'No tissue death', 'non-cancerous cells', 'Increased epithelial cells', 'Cells are disorganized'	'Immune cells are present', 'No tissue death', 'non-cancerous cells', 'Increased epithelial cells'	0	0	0
	'Increased epithelial cells', 'cancerous tumour', 'Increased number of epithelial cells', 'No tissue death', 'Visible Nucleoli'	'cancerous tumour'	1	0	1
	'Immune cells are present', 'No tissue death', 'Invasive Growth', 'Tissue death is present', 'Cells are disorganized'	'No tissue death'	0	1	0
	'Immune cells are present', 'No tissue death', 'Tissue death is present', 'non-cancerous cells', 'malignant tissues'	'Immune cells are present', 'Tissue death is present', 'malignant tissues'	1	0	1

FIGURE 5.2. The qualitative results on unseen pathology images. GT is the ground truth; $P_{retrieved}$ is the prediction on retrieved text and $P_{denoised}$ is the prediction on de-noised text; 0 and 1 denote benign and malignant labels, respectively. Irrelevant texts and incorrect predictions are highlighted in red.

Qualitative comparison. Figure 5.2 presents qualitative results on the zero-shot pathology image datasets. In the first row, the de-noised texts generated by RDCLM successfully filtered out undesired red-highlighted text from the benign images. Text-based classification using both original and de-noised retrieved texts is correct in this case, likely due to the low proportion of noise in both text groups. However, predictions for the remaining images are incorrect, coinciding with an increased proportion of irrelevant texts in the retrieved texts. The most visible noise patterns include irrelevant morphology descriptions, generic malignancy-related wording, and wrong-class descriptions that remain visually plausible for the query image. Although RDCLM does not completely eliminate noise in the last example, it still produces the correct prediction. This is because the majority-vote classification module (Section 3.5) ensures that as long as the number of relevant texts exceeds the irrelevant ones

in the de-noised output, the classification remains accurate. Hence, even with imperfect de-noising, the approach maintains robustness.

5. Failure Analysis of RDCLM De-noising

To further examine the cases in which noise is not fully removed, a sample-level error analysis was conducted using the trained RDCLM model. For each image, the retrieved texts, the RDCLM-generated de-noised texts, and the class-consistent descriptions were compared across the five zero-shot datasets and retrieval depths $k = \{5, 10, 15\}$. This analysis quantifies two complementary behaviours: *under-correction*, where wrong-class noisy text remains in the RDCLM output, and *over-correction*, where useful class-relevant text from the retrieved set is removed. The analysis also records whether errors are associated with retrieval misses, knowledge base ambiguity, stain/domain outliers, or near-balanced majority-vote results.

Table 5.7 shows that the remaining errors are not dominated by arbitrary generation failures. Instead, most failures arise when the retrieved set already contains visually plausible wrong-class descriptions, and RDCLM keeps these noisy retrieved texts in the output. Under-correction is the main reason for 4,595 out of 4,751 failed cases (96.7%), whereas pure retrieval misses account for 156 failed cases (3.3%). This indicates that RDCLM usually retrieves at least some relevant text, but can fail when wrong-class text such as “immune cells are present”, “cells are disorganised”, “tissue death is present”, or “cancerous tumour” remains more common in the de-noised output. Knowledge base ambiguity was not observed: the benign and malignant descriptions in the knowledge base are distinct, with no identical description shared by both classes. The more important issue is semantic overlap between generic descriptions, such as “no tissue death” versus “tissue death is present”, or “non-cancerous cells” versus “cancerous tumour”.

Table 5.8 further shows that RDCLM improves the precision of the generated texts overall, increasing mean precision from 0.510 in the retrieved set to 0.633 in the de-noised output. However, increasing k gives the model more noisy retrieved texts to process: the under-correction rate rises from 0.367 at $k = 5$ to 0.465 at $k = 15$, and accuracy decreases from

Error type	Cases	Rate	Main error cases	Portion of error cases	Meaning
Under-correction	5,340	0.418	4,595	0.967	Wrong-class noisy text is still present after de-noising.
Over-correction	4,792	0.375	–	–	Useful text from the true class is removed during de-noising.
Retrieval miss	320	0.025	156	0.033	The top-k retrieved texts contain too little true-class information for RDCLM to keep.
Stain/domain outlier	597	0.047	–	–	The image colour or stain pattern differs from most images in the dataset, making retrieval and de-noising harder.
Majority-vote limit	442	0.035	–	–	The final texts contain similar numbers of benign and malignant descriptions, so the vote is unstable.
Knowledge base ambiguity	0	0.000	0	0.000	Benign and malignant descriptions in the knowledge base are distinct; no identical description is shared by both classes.

TABLE 5.7. Error-source analysis for RDCLM over 12,786 evaluated image–k pairs. Error types are not mutually exclusive, while "main error cases" counts the most likely reason for each incorrect prediction.

0.653 to 0.603. This pattern explains why the qualitative examples sometimes retain noise: RDCLM is most challenged when a larger retrieval pool contains multiple visually plausible wrong-class phrases rather than a few clearly irrelevant ones.

Setting	N	Retrieval precision	Output precision	Accuracy	Under-correction	Over-correction
All	12,786	0.510	0.633	0.628	0.418	0.375
$k = 5$	4,262	0.507	0.657	0.653	0.367	0.337
$k = 10$	4,262	0.513	0.632	0.630	0.421	0.375
$k = 15$	4,262	0.510	0.609	0.603	0.465	0.412

TABLE 5.8. RDCLM de-noising analysis by retrieval depth. RDCLM improves output precision relative to the retrieved texts, but larger k increases the chance of retaining plausible noisy text.

The dataset-level analysis in Table 5.9 identifies the most difficult visual settings. EBHI-Seg has the highest accuracy and lowest correction error rates, suggesting that its visual patterns are better matched to the current knowledge base. In contrast, BreakHis is the most challenging dataset, with the highest under-correction (0.525) and over-correction (0.498) rates. This is consistent with its diverse breast tumour subtypes and magnification variation, which make wrong-class descriptions visually plausible. Camelyon and LC25000 also show higher under-correction than EBHI-Seg, indicating that stain/domain variation and organ-specific appearance can make the retained texts less stable. Overall, the remaining errors are therefore best understood as difficult retrieval cases: RDCLM generally improves output

precision, but can still be biased by strong wrong-class descriptions when the retrieved texts are noisy, visually plausible, or domain shifted.

Dataset	N	Retrieval precision	Output precision	Accuracy	Under-correction	Over-correction
CRC-HE-7K	3,000	0.539	0.640	0.635	0.416	0.369
EBHI-Seg	2,997	0.503	0.757	0.754	0.293	0.259
LC25000	3,000	0.484	0.620	0.618	0.428	0.382
BreakHis	3,000	0.522	0.514	0.509	0.525	0.498
Camelyon	789	0.477	0.632	0.616	0.447	0.342

TABLE 5.9. RDCLM de-noising analysis by dataset. BreakHis is the hardest setting because subtype and magnification diversity make wrong-class descriptions more visually plausible; EBHI-Seg is the most stable setting.

CHAPTER 6

Conclusion

This thesis investigated how to improve zero-shot image-text retrieval and image classification for malignancy recognition in histopathology by addressing the common problem of noisy retrieved texts. Retrieval-based De-noising Causal Language Modelling (RDCLM) is introduced as a pipeline that (1) builds a pathology-focused knowledge base, (2) retrieves candidate descriptions using a vision-language foundation model, (3) applies retrieval augmentations (retrieval negatives replacement and description-wise shuffling) during training, and (4) refines retrieved text via a de-noising causal language modelling module that integrates image features with retrieved textual context. To improve the robustness of the image-to-text projection, a multi-branch diverse-dimension projector with an inter- and intra-branch mutual-information objective is also proposed to encourage informative and diverse branch-wise representations.

Extensive experiments across multiple histopathology datasets demonstrate that RDCLM consistently improves top-k retrieval precision and zero-shot classification performance compared to a range of CLIP-style and retrieval-augmented baselines. Ablation studies confirm that the retrieval de-noising stage and the multi-branch projection each contribute materially to the observed gains, and that the retrieval augmentation strategies help generalisation to unseen subtypes and dataset shifts.

An auxiliary component in our training objective is a precision-inspired reward term used to mildly bias the model toward producing cleaner outputs. This auxiliary loss was included to encourage slightly higher output precision but was not central to the method's design; the primary improvements stem from the retrieval de-noising mechanism and the projection architecture. In other words, while the auxiliary term provides modest additional stabilisation in some settings, RDCLM's principal effectiveness arises from refining retrieved texts and aligning them with visual features.

Methodologically, RDCLM highlights a practical strategy for domain adaptation with large, mostly frozen models: (i) it constructs domain-specific textual resources, (ii) retrieves and refines noisy candidates rather than relying on single-shot prompts, and (iii) combines a lightweight trainable cross-modal projector with contrastive-style regularisers to preserve diversity and informativeness. This approach reduces the need for large-scale re-training while achieving strong zero-shot generalisation in a fine-grained domain.

6.1 Limitations

Although RDCLM shows strong zero-shot retrieval and classification performance, several limitations should be noted. First, the study focuses on methodological development and technical validation on public pathology datasets, rather than clinical workflow evaluation. Direct reader-study feedback from pathologists was not collected. If used in a clinical setting, the de-noised outputs should therefore be interpreted as filtered textual evidence about malignancy-related morphology, not as an autonomous diagnosis. Responsible deployment would require prospective clinical validation, calibration analysis, out-of-distribution detection, confidence reporting, bias assessment, and clinician-in-the-loop use.

Second, RDCLM was evaluated at the pathology patch level rather than as an end-to-end whole-slide-image (WSI) system. The current findings show that the retrieval-and-de-noising mechanism is effective for local pathology regions, but they do not demonstrate clinical-scale WSI deployment. For gigapixel WSIs, RDCLM would need to treat each slide as a collection

of tiles, with tissue detection, tile filtering, GPU batching, selection of informative regions, and aggregation of tile-level outputs into a slide-level decision.

Third, the evaluation design reduces, but does not eliminate, dataset-related limitations. The external zero-shot evaluations are source-disjoint by construction, and the BreakHis split is subtype-disjoint. However, patient-level or slide-level grouping was not explicitly enforced within BreakHis, so formal patient-level leakage control is not claimed for that setting. The cross-organ results should also be interpreted as encouraging empirical evidence rather than proof of organ-invariant pathology understanding. In addition, the downstream task is limited to binary benign-versus-malignant recognition; fine-grained subtype classification was not evaluated because subtype labels are less consistent across the heterogeneous datasets used in this thesis.

Fourth, several methodological choices were selected empirically but not exhaustively tuned. PLIP was used as the retrieval backbone to isolate the effect of retrieval de-noising on top of a pathology-specific VLFM, but broader comparisons across multiple pathology foundation model backbones were not conducted. The hybrid knowledge base of approximately 300 descriptions performed best among the tested description settings, but knowledge bases larger than 300 descriptions were not evaluated; therefore, the effect of further enlarging the knowledge base remains unknown. Similarly, the dimensions of the multi-branch projector were chosen to encourage complementary semantic learning while keeping the model lightweight, but exhaustive sensitivity analysis over branch number, dimension allocation, and mutual-information regularisation strength was not performed. A formal multi-seed or repeated training-subset stability analysis was also not conducted because of the computational cost of repeated end-to-end RDCLM training.

Finally, the failure analysis shows that the remaining errors are dominated by difficult retrieval cases rather than arbitrary generation failures. RDCLM substantially improves the precision of the de-noised texts, but it can still under-correct when the retrieved set contains visually plausible wrong-class descriptions, and it can over-correct by removing weaker but useful class-relevant phrases. The current majority-vote decision rule is transparent, but it may

be limited in borderline cases where the retained benign and malignant evidence is nearly balanced.

6.2 Future Outlook

Future work should address these limitations while extending RDCLM toward broader scientific and clinical utility. Clinical translation will require simulated or prospective reader studies to evaluate whether the filtered textual evidence improves pathologist efficiency, diagnostic confidence, and inter-observer agreement. WSI-scale deployment should be explored through tile-based pipelines with tissue detection, informative-region selection, efficient batching, and slide-level aggregation.

The framework should also be evaluated beyond binary tumour malignancy recognition. Potential extensions include tumour grading, molecular subtype prediction, treatment-response assessment, and non-oncological pathology tasks such as inflammatory, renal, or hepatic disease recognition. These settings would require richer and more organ-specific knowledge bases, as well as systematic studies of knowledge base size, quality, and redundancy.

Methodologically, future work should report stability across random seeds and training subsets, compare RDCLM across multiple pathology foundation model backbones, and test sensitivity to multi-branch projector design. The current majority-vote classifier could also be replaced or supplemented with confidence-weighted aggregation, uncertainty estimation, or learned slide-level aggregation to better handle borderline cases and visually plausible wrong-class retrieved descriptions.

Bibliography

[1] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine and J. Zou. 'A visual–language foundation model for pathology image analysis using medical Twitter'. In: *Nature Medicine* 29.9 (2023), pp. 2307–2316.

[2] M. Y. Lu, B. Chen, A. Zhang, D. F. Williamson, R. J. Chen, T. Ding et al. 'Visual language pretrained multiple instance zero-shot transfer for histopathology images'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 19764–19775.

[3] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, I. Liang, T. Ding et al. 'A visual-language foundation model for computational pathology'. In: *Nature Medicine* 30.3 (2024), pp. 863–874.

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

[5] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn et al. 'BiomedCLIP: A Multimodal Biomedical Foundation Model Pretrained from Fifteen Million Scientific Image-Text Pairs'. In: *arXiv preprint arXiv:2303.00915* (2023).

[6] K. He, X. Zhang, S. Ren and J. Sun. 'Deep Residual Learning for Image Recognition'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner et al. 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale'. In: *arXiv preprint arXiv:2010.11929* (2020).

[8] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian et al. 'A Survey on Deep Learning in Medical Image Analysis'. In: *Medical Image Analysis* 42 (2017), pp. 60–88. DOI: 10.1016/j.media.2017.07.005.

[9] K. Ding, M. Zhou, H. Wang, O. Gevaert, D. Metaxas and S. Zhang. 'A large-scale synthetic pathological dataset for deep learning-enabled segmentation of breast cancer'. In: *Scientific Data* 10.1 (2023), p. 231.

[10] H. Wang, Q. Jin, S. Li, S. Liu, M. Wang and Z. Song. 'A Comprehensive Survey on Deep Active Learning in Medical Image Analysis'. In: *Medical Image Analysis* (2024), p. 103201.

[11] J. A. Diao, R. J. Chen and J. C. Kvedar. 'Efficient Cellular Annotation of Histopathology Slides with Real-Time AI Augmentation'. In: *NPJ Digital Medicine* 4.1 (2021), p. 161.

[12] A. v. d. Oord, Y. Li and O. Vinyals. 'Representation Learning with Contrastive Predictive Coding'. In: *arXiv preprint arXiv:1807.03748* (2018).

[13] T. Chen, S. Kornblith, M. Norouzi and G. Hinton. 'A Simple Framework for Contrastive Learning of Visual Representations'. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1597–1607.

[14] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam et al. 'Clinical-grade computational pathology using weakly supervised deep learning on whole slide images'. In: *Nature Medicine* 25.8 (2019), pp. 1301–1309. DOI: 10.1038/s41591-019-0508-1.

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal et al. 'Learning transferable visual models from natural language supervision'. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8748–8763.

[16] Z. Chen, S. Diao, B. Wang, G. Li and X. Wan. 'Towards unifying medical vision-and-language pre-training via soft prompts'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 23403–23413.

[17] J. N. Kather, N. Halama and A. Marx. *100,000 histological images of human colorectal cancer and healthy tissue*. Version v0.1. Zenodo, Apr. 2018. DOI: 10.5281/zenodo.1214456. URL: https://doi.org/10.5281/zenodo.1214456.

[18] A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand and S. M. Mastorides. 'Lung and Colon Cancer Histopathological Image Dataset (LC25000)'. In: *arXiv preprint arXiv:1912.12142* (2019).

[19] L. Shi, X. Li, W. Hu, H. Chen, J. Chen, Z. Fan et al. 'EBHI-Seg: A Novel Enteroscope Biopsy Histopathological Hematoxylin and Eosin Image Dataset for Image Segmentation Tasks'. In: *Frontiers in Medicine* 10 (2023), p. 1114673.

[20] F. A. Spanhol, L. S. Oliveira, C. Petitjean and L. Heutte. 'A dataset for breast cancer histopathological image classification'. In: *IEEE Transactions on Biomedical Engineering* 63.7 (2016), pp. 1455–1462. DOI: 10.1109/TBME.2015.2496264.

[21] M. Maniparambil, C. Vorster, D. Molloy, N. Murphy, K. McGuinness and N. E. O'Connor. 'Enhancing CLIP with GPT-4: Harnessing Visual Descriptions as Prompts'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 262–271.

[22] O. Saha, G. Van Horn and S. Maji. 'Improved Zero-Shot Classification by Adapting VLMs with Text Descriptions'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 17542–17552.

[23] A. Conti, E. Fini, M. Mancini, P. Rota, Y. Wang and E. Ricci. 'Vocabulary-Free Image Classification'. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 30662–30680.

[24] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot and B. Yener. 'Histopathological image analysis: A review'. In: *IEEE Reviews in Biomedical Engineering* 2 (2009), pp. 147–171. DOI: 10.1109/RBME.2009.2034865.

[25] D. Komura and S. Ishikawa. 'Machine learning methods for histopathological image analysis'. In: *Computational and Structural Biotechnology Journal* 16 (2018), pp. 34–42. DOI: 10.1016/j.csbj.2018.01.001.

[26] A. Madabhushi and G. Lee. ‘Image analysis and machine learning in digital pathology: Challenges and opportunities’. In: *Medical Image Analysis* 33 (2016), pp. 170–175. DOI: 10.1016/j.media.2016.06.037.

[27] A. Janowczyk and A. Madabhushi. ‘Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases’. In: *Journal of Pathology Informatics* 7.1 (2016), p. 29. DOI: 10.4103/2153-3539.186902.

[28] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens et al. ‘Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer’. In: *JAMA* 318.22 (2017), pp. 2199–2210. DOI: 10.1001/jama.2017.14585.

[29] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri and F. Mahmood. ‘Data-efficient and weakly supervised computational pathology on whole-slide images’. In: *Nature Biomedical Engineering* 5 (2021), pp. 555–570. DOI: 10.1038/s41551-021-00762-4.

[30] R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan et al. ‘Scaling vision transformers to gigapixel images via hierarchical self-supervised learning’. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 16144–16155.

[31] S. Wang, Y. Zhu, P. Yu et al. ‘RetCCL: A Self-Supervised Framework for Histopathology Representation Learning via Clustering-Enhanced Contrastive Learning’. In: *Medical Image Analysis* 84 (2023), p. 102699. DOI: 10.1016/j.media.2023.102699.

[32] W. O. Ikezogwo, S. Seyfioglu, F. Ghezloo, D. Geva, F. Sheikh Mohammed, P. K. Anand et al. ‘Quilt-1M: One Million Image-Text Pairs for Histopathology’. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 37995–38017.

[33] V. Udandarao, A. Gupta and S. Albanie. ‘SuS-X: Training-Free Name-Only Transfer of Vision-Language Models’. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 2725–2736.

[34] M. Ilse, J. M. Tomczak and M. Welling. ‘Attention-based Deep Multiple Instance Learning’. In: *Proceedings of the 35th International Conference on Machine Learning*

(ICML). Vol. 80. Proceedings of Machine Learning Research. 2018, pp. 2127–2136. URL: https://proceedings.mlr.press/v80/ilse18a/ilse18a.pdf.

[35] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji et al. 'TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification'. In: *arXiv preprint arXiv:2106.00908* (2021). URL: https://arxiv.org/abs/2106.00908.

[36] P. Tourniaire et al. 'MS-CLAM: Mixed supervision for the classification and interpretation of whole-slide images'. In: *Computer Methods and Programs in Biomedicine* (2023). DOI: 10.1016/j.cmpb.2023.107406. URL: https://www.sciencedirect.com/science/article/pii/S0169260723002462.

[37] Z. Wang, X. Jiang, Y. Li, Y. Xie et al. 'MedCLIP: Contrastive Learning from Unpaired Medical Images and Text'. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2022. URL: https://aclanthology.org/2022.emnlp-main.256/.

[38] P. F. (authors). *PLIP: Pathology Language and Image Pre-Training (code and model repository)*. Accessed: 2025-10-31. 2024. URL: https://github.com/PathologyFoundation/plip.

[39] J. Xiang, X. Wang, X. Zhang, Y. Xi, F. Eweje, Y. Chen et al. 'A vision–language foundation model for precision oncology'. In: *Nature* (2025), pp. 1–10.

[40] X. Zhou, X. Zhang, C. Wu, Y. Zhang, W. Xie and Y. Wang. 'Knowledge-Enhanced Visual-Language Pretraining for Computational Pathology'. In: *European Conference on Computer Vision*. Springer. 2024, pp. 345–362.

[41] R. Taware, S. Varat, G. Salunke, C. Gawande, G. Kale, R. Khengare et al. 'ShufText: A Simple Black-Box Approach to Evaluate the Fragility of Text Classification Models'. In: *International Conference on Machine Learning, Optimization, and Data Science*. Springer. 2021, pp. 235–249.

[42] H. Liu, C. Li, Q. Wu and Y. J. Lee. 'Visual Instruction Tuning'. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 34892–34916.

[43] J. Li, D. Li, S. Savarese and S. Hoi. ‘BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models’. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 19730–19742.

[44] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar et al. ‘Phi-4 Technical Report’. In: *arXiv preprint arXiv:2412.08905* (2024).

[45] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal et al. ‘Language Models are Few-Shot Learners’. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1877–1901.

[46] G. Litjens, P. Bandi, B. Ehteshami Bejnordi, O. Geessink, M. Balkenhol, P. Bult et al. ‘1399 H&E-Stained Sentinel Lymph Node Sections of Breast Cancer Patients: The CAMELYON Dataset’. In: *GigaScience* 7.6 (2018), giy065.

[47] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus et al. ‘Scaling Instruction-Finetuned Language Models’. In: *Journal of Machine Learning Research* 25.70 (2024), pp. 1–53.

[48] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang et al. ‘LoRA: Low-Rank Adaptation of Large Language Models’. In: *International Conference on Learning Representations*. 2022.