Multimodal Representation Learning in Medical Vision–Language Models

Negin Baghbanzadeh

A Thesis Submitted to the Faculty of Graduate Studies
In Partial Fulfillment of the Requirements
for the Degree of Master of Science

Graduate Program in Computer Science

York University
Toronto, Ontario

September 2025

Abstract

Recent advances in multimodal models such as LLaVA and InstructBLIP highlight the importance of high-quality image encoders, particularly in the biomedical domain where figures and captions are complex. Existing medical vision–language datasets primarily emphasize scale, often overlooking data quality. In this thesis, we introduce **OPEN-PMC**, a carefully curated collection of biomedical image–text pairs derived from PubMed Central. OPEN-PMC incorporates multiple refinement steps, including compound figure decomposition with a modality-robust detector, subcaption segmentation, in-text reference extraction and summarization, and modality-aware classification. Using OPEN-PMC-2M, we conduct controlled experiments to quantify the effect of each processing step on contrastive pretraining. Our findings show that subfigure decomposition and enriched captions substantially improve retrieval, zero-shot classification, and robustness, outperforming larger but noisier datasets. Scaling to **OPEN-PMC-18M**, one of the largest curated biomedical VL datasets to date, we demonstrate state-of-the-art performance while discussing remaining limitations in large-scale contextual augmentation and clinical validation.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Elham Dolatabadi, for her continuous guidance, encouragement, and support throughout the course of this thesis. I am also sincerely thankful to my co-supervisor, Dr. Ali Etemad, whose insightful feedback and expertise greatly improved the quality of my work. In addition, I extend my appreciation to Dr. Arash Afkanpour from the Vector Institute for his valuable contributions that enriched this research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

In recent years, large-scale multimodal models such as LLaVA (1) and InstructBLIP (2) have demonstrated remarkable capabilities in vision–language understanding and reasoning. A unifying characteristic of these systems is that they perceive images exclusively through an image encoder, which is then aligned with a language encoder to produce a joint representation space. As such, the quality of the image encoder becomes a central determinant of downstream performance, not only for general-domain tasks but also for domain-specific applications such as medicine. Designing or pretraining a strong image encoder is therefore crucial, since all subsequent reasoning stages depend on the fidelity of the learned visual features.

Existing biomedical vision–language (VL) datasets have largely prioritized expansion in size, often under the assumption that scale alone is the dominant driver of representation quality. For example, PMC-15M (3) and BIOMEDICA (4) have assembled tens of millions of image–caption pairs from PubMed Central articles, while other resources such as PMC-OA (5) and ROCO (6) provide smaller but still substantial corpora. Although these datasets have significantly advanced the field, their pipelines focus primarily on maximizing coverage, with less attention to curation steps that influence semantic alignment and quality. However, noisy captions, compound figures and missing in-text references can all introduce misalignments between images and text. As a result, the full potential of contrastive training may not be realized. This observation motivates a key question: *to what extent does dataset quality, rather than scale alone, influence the performance of medical vision–language models?*

To address this question, in this thesis we introduce **OPEN-PMC**, a carefully curated collection of biomedical image–text pairs derived from PubMed Central articles. Unlike previous efforts, OPEN-PMC explicitly emphasizes quality by incorporating multiple refinement steps, including (i) subfigure decomposition to avoid training on noisy compound figures, (ii) segmentation of captions into subcaptions, (iii) alignment of text with visual markers, (iv) incorporation of in-text references and their summarization using large language models, and (v) modality-aware classification of subfigures.

In particular, a major source of noise in biomedical corpora arises from compound figures, where a single panel may contain multiple subfigures spanning different modalities (for example, radiology, histology, and flow charts presented together). Treating these panels as atomic units introduces significant misalignment between the image and its caption, since the caption often describes each subfigure separately. To address this, we trained and deployed a *compound figure decomposition model* using a DAB-DETR model that was trained on a large set of synthetically generated compound figures. This model is capable of robustly identifying subfigure boundaries across a wide range of biomedical layouts and modalities, and then cropping and exporting each subfigure as an independent image. By systematically applying this decomposition step, OPEN-PMC ensures that every image–text pair corresponds to a semantically coherent unit, thereby eliminating much of the ambiguity present in previous datasets.

Another critical refinement is the handling of *in-text references*. In many scientific articles, the figure caption alone does not capture the full meaning of a biomedical image. Authors often refer to figures in the main body of the text, providing additional explanation, context, or interpretation. Simply discarding these references leads to a loss of valuable semantic information, while including them verbatim results in exces-

sively long inputs that exceed model context windows. To strike a balance, we used large language models to summarize in-text references into concise, context-rich statements that preserve the essential information without introducing redundancy. These summaries were then merged with the original captions, producing a more complete and contextually faithful description for each subfigure. This process allows the dataset to preserve the interpretive richness of biomedical literature while maintaining compatibility with standard contrastive pretraining frameworks.

Our first dataset, **OPEN-PMC-2M**, was curated with the explicit goal of evaluating how each processing step contributes to downstream performance. By conducting a series of ablation studies, we quantified the impact of subfigure decomposition, caption granularity, and contextual augmentation on retrieval, zero-shot classification, and robustness benchmarks. The experiments revealed several key findings. First, subfigure decomposition consistently improved cross-modal alignment compared to treating compound figures as atomic units, yielding particularly large gains in radiology benchmarks. Second, although subcaptions provide finer alignment, full captions preserved more contextual richness and resulted in stronger overall performance. Third, combining subcaptions with in-text references emerged as the best setting, striking a balance between specificity and context. Together, these results demonstrate that high-quality dataset design can outperform larger but noisier alternatives, highlighting that careful preprocessing directly shapes model performance.

Building on these insights, we scaled our pipeline to create **OPEN-PMC-18M**, one of the largest curated biomedical VL datasets to date. OPEN-PMC-18M benefits from large-scale subfigure decomposition using a modality-robust DAB-DETR model trained on synthetic compound figures, thereby improving coverage across diverse imaging modalities. However, due to computational and resource constraints, we were unable to apply subcaption extraction or in-text summarization at this scale. As a result, OPEN-PMC-18M represents a trade-off: it combines unprecedented size with curated subfigure-level quality, but does not yet fully incorporate the richer textual augmentations shown to be beneficial in OPEN-PMC-2M. Despite this limitation, models trained on OPEN-PMC-18M achieve state-of-the-art performance across retrieval and zero-shot classification benchmarks, as well as enhanced robustness to visual perturbations, confirming that scale combined with subfigure-level curation produces strong and generalizable encoders.

In addition to evaluating retrieval and classification performance, this thesis also investigates the nature of the learned representations through systematic similarity analysis. Using techniques such as Maximum Mean Discrepancy (MMD), we probe how different preprocessing strategies influence the structure of embedding spaces. These analyses reveal that models trained on high-quality datasets not only achieve stronger downstream results but also produce more stable and semantically meaningful embedding distributions. For instance, incorporating subfigure decomposition and in-text references yields representations that cluster more coherently by modality and clinical context, while noisy or minimally curated datasets lead to fragmented neighborhoods and inflated similarity measures. By complementing benchmark results with representation-level insights, this work provides a deeper understanding of why dataset quality matters, showing that careful curation improves both external task performance and the internal geometry of learned multimodal spaces.

In summary, this thesis contributes three key elements to the field of medical vision–language modeling. First, it demonstrates through systematic experiments that data quality, achieved via subfigure decompo-

sition, caption contextualization, and semantic filtering, substantially improves model performance, often outweighing the effect of raw dataset size. Second, it introduces OPEN-PMC-2M as a high-quality benchmark for controlled evaluation of preprocessing strategies, yielding concrete guidelines for optimal dataset construction. Third, it delivers OPEN-PMC-18M, a large-scale, publicly available resource that sets new standards for medical vision–language pretraining. Taken together, these contributions highlight that careful dataset curation is not only a methodological choice but a fundamental requirement for building reliable, high-performing multimodal encoders for biomedical applications.

The remainder of this thesis is organized as follows. Chapter 2 provides a detailed review of the literature on medical vision-language modeling, including prior datasets, model architectures, and evaluation methodologies. Chapter 3 introduces the overall methodology and describes the curation pipeline used to construct the OPEN-PMC datasets, with particular attention to compound figure decomposition, subcaption segmentation, in-text reference summarization, and modality classification. Chapter 4 presents the experimental design, outlining the training setup, evaluation metrics, ablation studies conducted to assess the contribution of each curation step, and the main results, including performance comparisons of models trained on OPEN-PMC-2M and OPEN-PMC-18M against prior state-of-the-art baselines, as well as a detailed representation analysis of the learned embeddings. Finally, Chapter 5 concludes the thesis by summarizing the key findings, highlighting the contributions of OPEN-PMC, and discussing limitations and future research directions.

The main contributions of this thesis are as follows:

- **Introduction of OPEN-PMC datasets.** We curated a new family of biomedical vision–language datasets derived from PubMed Central articles, emphasizing quality through subfigure decomposition, caption segmentation, in-text reference summarization, and modality-aware filtering.
- **OPEN-PMC-2M for controlled studies.** We constructed a high-quality dataset of two million image–text pairs designed for ablation studies. This dataset enabled systematic evaluation of preprocessing strategies, including compound figure decomposition, subcaption extraction, and in-text reference augmentation.
- **Empirical analysis of dataset quality.** Through extensive experiments, we demonstrated that dataset quality, especially subfigure decomposition and in-text contextualization, substantially improves performance across retrieval and zero-shot classification benchmarks, often outweighing the benefits of larger but noisier datasets.
- **Identification of the best training setting.** We showed that the combination of subfigures, subcaptions with in-text reference summaries produced the most effective representation learning, balancing fine-grained alignment with contextual richness.
- **Compound figure decomposition.** We synthesized a dataset of compound biomedical figures to train a DAB-DETR–based decomposition model, enabling large-scale separation of figures into subfigures and improving image–text alignment in the OPEN-PMC datasets.
- **Scaling to OPEN-PMC-18M.** We developed one of the largest curated biomedical VL datasets to

date, containing 18 million subfigure-level image–text pairs. OPEN-PMC-18M achieves state-of-the-art performance in retrieval, classification, and robustness tasks.

- **Guidelines for dataset curation.** Based on our findings, we establish practical recommendations for designing biomedical vision–language datasets, highlighting that fidelity and contextual richness are critical factors in addition to dataset size.

2 Background

2.1 Pretraining Image Encoders: From Supervised to Self-Supervised

The success of vision models has been closely tied to the methods used to pretrain image encoders. Over time, the field has evolved from purely supervised pretraining to a variety of self-supervised approaches that aim to reduce reliance on human annotations. This section provides an overview of supervised methods, early self-supervised pretext tasks, and generative representation learning, before introducing contrastive learning.

2.1.1 Supervised Pretraining on Labeled Data

The earliest breakthroughs in deep learning for vision were achieved through supervised training on large labeled datasets. AlexNet (7), trained on the ImageNet dataset (8), demonstrated the power of convolutional neural networks (CNNs) for large-scale image classification. Subsequent architectures such as VGG (9), ResNet (10), and EfficientNet (11) further advanced performance through deeper networks, residual connections, and architecture search.

Supervised ImageNet pretraining became the standard initialization for image encoders across a wide range of computer vision tasks, including object detection, semantic segmentation, and image retrieval. By transferring the learned representations to downstream tasks, researchers were able to achieve strong results with limited task-specific data. However, this approach is fundamentally limited by the need for large-scale human annotations. In addition, models pretrained in this way often exhibit reduced generalization to domains that differ substantially from ImageNet, such as medical or satellite imagery.

2.1.2 Early Self-Supervised Pretext Tasks

To reduce dependence on manual labels, early self-supervised methods introduced *pretext tasks*, which generate supervisory signals directly from the input data. The goal was to force the model to learn semantically meaningful features by solving proxy problems that do not require labels.

Context Prediction (12) proposed predicting the relative position of image patches within an image. By requiring the model to infer spatial relationships, the network was encouraged to capture semantic structure. Jigsaw Puzzles (13) introduced the jigsaw puzzle task, where an image is divided into shuffled patches and the model must predict the correct ordering. This task encouraged reasoning about object parts and global context. Colorization (14) explored automatic image colorization, where a model predicts the color channels from grayscale input. This formulation leverages the natural correlation between luminance and chrominance as a supervisory signal. Rotation Prediction (15) introduced a simple yet effective task of predicting the rotation angle (0°, 90°, 180°, 270°) applied to an image. The model is forced to recognize object orientation and semantics in order to succeed.

Although these approaches demonstrated that useful features could be learned without labels, their performance often lagged behind supervised pretraining. The features tended to overfit to the artificial pretext task rather than generalize to downstream applications.

2.1.3 Generative Self-Supervised Learning

Another major direction in self-supervised representation learning involved generative modeling. Here, the objective is to reconstruct input data or generate realistic samples, with the expectation that strong generative models will capture meaningful latent representations. Autoencoders (16) learn compressed latent representations by reconstructing input images through an encoder–decoder framework. Variational Autoencoders (VAEs) (17) extend this idea with a probabilistic latent space, enabling controlled sampling and interpolation between latent variables.

2.1.4 From Pretext and Generative Methods to Contrastive Learning

The limitations of pretext-task and generative approaches highlighted the need for a more direct way of learning transferable features without labels. Contrastive learning emerged as a powerful solution by explicitly optimizing for invariance and discriminability: positive pairs are pulled together in the embedding space, while negatives are pushed apart. This formulation not only addressed the shortcomings of earlier methods but also enabled models to surpass supervised ImageNet pretraining on many benchmarks. The next subsection explores contrastive learning in detail, beginning with its application in unimodal representation learning and later extending to multimodal vision–language models.

2.2 Contrastive Learning in Representation Learning

Contrastive learning is a self-supervised learning paradigm that has become central to modern machine learning. The core idea is to learn representations by bringing semantically similar samples (positives) closer together in an embedding space, while pushing dissimilar samples (negatives) apart. Unlike supervised learning, contrastive learning does not require explicit labels, which makes it especially powerful for representation learning on large-scale unlabeled data.

The roots of contrastive learning can be traced back to metric learning and siamese networks (18), where the objective was to learn embeddings that preserve semantic similarity. Later, Noise Contrastive Estimation (NCE) (19) and the InfoNCE loss (20) formalized the contrastive objective in a probabilistic framework. These early formulations laid the groundwork for large-scale self-supervised learning methods that have since reshaped both computer vision and natural language processing.

2.2.1 Early Siamese Networks and Metric Learning

Before the modern wave of contrastive learning, one of the earliest and most influential architectures for similarity learning was the Siamese network (18). A Siamese network consists of two identical neural networks (sharing parameters) that process two different inputs in parallel. The outputs are then compared using a distance metric, such as Euclidean distance or cosine similarity, to determine whether the inputs are semantically similar. This design enforces that the learned representations are invariant to superficial differences, while still capturing discriminative information for distinguishing dissimilar inputs.

Siamese networks were first popularized in the context of signature verification (18), where the goal was to identify whether two handwritten signatures belonged to the same person. These early works laid the

foundation for learning representations through pairwise similarity, foreshadowing the role of contrastive objectives in large-scale self-supervised learning.

2.2.2 Contrastive Loss Functions

A key component of contrastive learning is the choice of loss function that drives the representation learning process. Several formulations have been developed, each with different properties. The original contrastive loss (21) was designed for Siamese networks. Given a pair of samples with label $y \in \{0,1\}$ indicating whether they are similar ($y = 1$) or dissimilar ($y = 0$), the loss is defined as:

$$\mathcal{L}_{\text{contrastive}} = y \cdot D^2 + (1-y) \cdot \max(0, m-D)^2$$

where D is the distance between the embeddings of the two samples, and m is a margin that enforces dissimilar pairs to be at least m apart in the embedding space. This loss encourages positive pairs to be close and negative pairs to be sufficiently far apart.

The triplet loss (22) extends the contrastive loss by considering three samples at a time: an anchor, a positive (same class), and a negative (different class). The loss enforces that the distance between the anchor and the positive is smaller than the distance between the anchor and the negative by at least a margin α:

$$\mathcal{L}_{\text{triplet}} = \max\left(0, D(a,p) - D(a,n) + \alpha\right)$$

where $D(\cdot,\cdot)$ denotes the distance metric. Triplet loss became popular in face recognition and retrieval tasks, as it provided finer control over relative similarities compared to pairwise contrastive loss. Noise Contrastive Estimation (19) introduced the idea of distinguishing data samples from artificially generated noise samples. This probabilistic formulation inspired later developments in contrastive learning by framing the task as binary classification between true and noise-distributed pairs. NCE influenced modern objectives such as InfoNCE by showing that negative sampling can approximate likelihood-based training in a computationally tractable way. The InfoNCE loss (20) is arguably the most widely used contrastive objective in modern self-supervised learning. Given one positive pair and multiple negatives, the loss is defined as:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(\mathbf{z}, \mathbf{z}^+)/\tau)}{\exp(\text{sim}(\mathbf{z}, \mathbf{z}^+)/\tau) + \sum_{j=1}^{K} \exp(\text{sim}(\mathbf{z}, \mathbf{z}_j^-)/\tau)}$$

where $\mathbf{z}$ and $\mathbf{z}^+$ are embeddings of a positive pair, $\mathbf{z}_j^-$ are embeddings of negative samples, and τ is a temperature parameter. InfoNCE elegantly generalizes contrastive learning to large batches by treating all non-matching samples in the batch as negatives. This formulation underpins methods such as SimCLR and CLIP, making it the dominant loss in contemporary contrastive learning.

2.2.3 Contrastive Learning with Unimodal Data

Initially, contrastive learning was applied within a single modality, particularly in computer vision. The central challenge was to design pretext tasks that generate positive and negative pairs without human annotation.

SimCLR (23) is one of the most influential unimodal contrastive learning methods. It constructs positive pairs by applying two different data augmentations to the same image, while treating other images in the batch as negatives. Representations are learned using an encoder (e.g., ResNet) followed by a projection head, and the similarity between representations is maximized for positive pairs using the InfoNCE loss. The simplicity of this framework, combined with heavy data augmentation and large batch sizes, demonstrated that contrastive learning can rival or surpass supervised pretraining in vision tasks. Momentum Contrast (MoCo) (24) introduced a memory bank and a momentum-updated encoder to address the limitation of requiring very large batch sizes. By maintaining a dynamic queue of negative samples, MoCo allows efficient contrastive training with manageable hardware resources. This approach made contrastive learning more practical and scalable, leading to strong results across multiple computer vision benchmarks. Bootstrap Your Own Latent (BYOL) (25) challenged the assumption that negative pairs are necessary for contrastive learning. Instead, BYOL employs two networks (an online and a target network) and trains the online network to predict the target representation of the same image under different augmentations. Surprisingly, this method avoids representational collapse and achieves state-of-the-art results without explicit negatives.

2.2.4 Limitations of Unimodal Contrastive Learning

Despite their success, unimodal contrastive learning approaches have limitations. They rely heavily on carefully designed augmentations to generate positive pairs, which may not always capture semantic invariances. Moreover, unimodal contrastive models are confined to a single modality and cannot directly capture relationships across different types of data, such as images and text. This limits their utility in tasks that require reasoning across modalities, such as image captioning, cross-modal retrieval, or visual question answering. These shortcomings motivated the extension of contrastive learning to the multimodal setting.

2.2.5 Contrastive Learning with Multimodal Data

The breakthrough for multimodal contrastive learning came with the availability of large-scale paired datasets, particularly image–text pairs from the web. Instead of relying on artificial augmentations, positive pairs can be drawn directly from naturally co-occurring modalities (e.g., an image and its caption), while negative pairs are sampled from mismatched combinations.

CLIP (Contrastive Language–Image Pretraining) (26) demonstrated that contrastive learning could scale to hundreds of millions of noisy image–text pairs collected from the web. Using a dual-encoder architecture with a Vision Transformer (ViT) for images and a Transformer-based text encoder, CLIP learns a joint embedding space by maximizing the similarity between corresponding image–text pairs and minimizing similarity for non-matching pairs within a batch. The resulting representations are highly generalizable and support zero-shot transfer: by embedding text prompts such as "a photo of a dog," CLIP can classify images without task-specific training data. ALIGN (27) scaled this idea further by training on over one billion noisy image–alt-text pairs collected without extensive filtering. Despite the lower quality of data compared to CLIP's curated set, the sheer scale of ALIGN enabled strong performance across retrieval and classification benchmarks. This work highlighted the importance of dataset scale in multimodal contrastive learning. Microsoft's Florence (28) extended the paradigm by incorporating a much larger dataset and introducing ar-

chitectural refinements. Beyond these early models, subsequent works such as ALBEF (29), BLIP (30), and FLAVA (31) combined contrastive learning with generative or fine-grained alignment objectives, enabling stronger performance on downstream multimodal reasoning tasks.

2.2.6 Adaptation to the Medical Domain

Building on the success of multimodal contrastive learning in the general domain, several recent efforts have adapted this framework to the medical setting, where paired image-text data is more specialized and less abundant. In medical vision-language modeling, the same dual-encoder architecture is used, but the data is typically sourced from scientific figures and associated captions in biomedical literature, as seen in datasets like PMC-OA (5), PMC-15M (3), and BIOMEDICA (4).

Despite using the same InfoNCE loss, the challenges in the medical domain are distinct. Figures are frequently compound (multi-panel), and captions are often sparse, ambiguous, or contextually detached. To mitigate this, models like PMC-CLIP (5) incorporate preprocessing pipelines that decompose compound figures into subfigures and enrich captions using surrounding text.

ROCO (Radiology Objects in COntext) (6) is a widely used dataset in the medical vision-language domain, consisting of 81,000 radiology images paired with corresponding captions. The dataset was curated from the PubMed Central Open Access Subset, with a focus on radiological content. Each image is associated with a caption extracted from the article's figure legend, and labeled with modality and body part metadata. ROCO includes both image-level annotations and structured text information, making it suitable for tasks such as image-text retrieval, caption generation, and multimodal representation learning. Despite its utility, ROCO is relatively small in scale compared to more recent datasets like PMC-OA and PMC-15M, and its captions are often concise and domain-specific, which may limit generalization in broader biomedical or clinical settings. Nonetheless, it remains a valuable benchmark for evaluating radiology-focused vision-language models.

PMC-15M (3) is a large-scale biomedical image-text dataset containing approximately 15 million figure-caption pairs, extracted from over 4.4 million articles in the PubMed Central Open Access Subset (PMC-OA). The dataset curation pipeline involves parsing structured XML files to extract figures and their corresponding captions, followed by quality filtering to exclude illegible or corrupted images. While PMC-15M covers a broad spectrum of biomedical image modalities, including microscopy, radiology, histopathology, and clinical charts, it does not enforce strict filtering for medically relevant content. This inclusivity introduces a degree of noise, which may hinder performance on fine-grained medical tasks. Nevertheless, the scale and diversity of PMC-15M make it a valuable resource for pretraining multimodal models. It serves as the backbone for training BiomedCLIP, a vision-language model based on the CLIP architecture, which is optimized using the InfoNCE contrastive loss. BiomedCLIP demonstrates improved performance across a variety of biomedical vision-language benchmarks, showcasing the utility of large-scale weak supervision in the medical domain.

PMC-OA (5) is a large-scale biomedical image-text dataset comprising approximately 1.65 million high-quality figure-caption and subfigure-subcaption pairs, extracted from over 2.4 million articles in the PubMed Central Open Access Subset (PMC-OA). The dataset is constructed through a rigorous, automated pipeline designed to ensure semantic alignment and domain relevance, addressing challenges such as compound fig-

ures and sparse or ambiguous captions commonly found in biomedical literature.

The curation process begins with extracting all figure-caption pairs from the PMC-OA corpus. To identify medically relevant content, a two-stage filtering strategy is applied. First, captions are screened using a predefined medical keyword list. Then, figures are classified using a ResNet-101 model trained on the DocFigure dataset to categorize them into 28 figure types. Figures predicted as any of the top four medical-related categories are retained, yielding 381,096 medically relevant figures.

Given the high prevalence of compound figures, the authors employ a DETR (DEtection TRansformer) model with a ResNet-34 backbone, trained on MedICaT data, to detect and segment subfigures. This model achieves a mean Average Precision (mAP) of 0.94 on the validation set and extracts 1,848,729 subfigures from compound images.

To strengthen image-text alignment, the corresponding captions are split into subcaptions using rule-based heuristics and formatting cues. A CLIP model is then fine-tuned on a manually annotated subset to align subfigures with subcaptions, achieving 73% alignment accuracy. This results in 1,003,911 high-quality aligned subfigure-subcaption pairs. The final dataset is constructed by combining these with remaining subfigure-caption pairs, culminating in 1,646,592 image-text pairs.

PMC-OA is used to pretrain **PMC-CLIP**, a domain-specific vision-language model built upon the CLIP architecture. PMC-CLIP adopts a dual-objective training approach, integrating image-text contrastive (ITC) loss with masked language modeling (MLM), to foster robust and semantically rich multimodal representations tailored to the biomedical domain.

BIOMEDICA (4) is a large-scale, meticulously curated biomedical image-text dataset, specifically designed to support vision-language research across diverse medical subfields. It is constructed from over 6 million full-text articles in the PubMed Central Open Access Subset, resulting in more than 24 million unique image-caption pairs. The data curation pipeline employed in BIOMEDICA is fully automated and modular, incorporating structured XML parsing, figure extraction, caption association, and metadata enrichment. Each article is first parsed to extract figures and their corresponding figure titles and captions. Unlike earlier datasets, BIOMEDICA also integrates subcaptions and in-text figure references where available, providing richer semantic grounding between the image and text modalities. They include full paragraph containing each mention of a figure in their dataset.A distinctive feature of BIOMEDICA is its use of self-supervised learning for image labeling. Specifically, the dataset employs the DINO model (32) to generate high-level visual representations of images without requiring manual annotations. This approach enables the assignment of coarse-grained labels, such as imaging modality and panel type, to each image. These labels are further refined through expert curation, ensuring high-quality annotations across the dataset.

Comparison and Motivation While these datasets and models represent significant advancements in biomedical VL research, challenges remain in terms of data quality. ROCO provides high-quality annotations but is limited in scale. PMC-OA offer larger dataset with automated extraction pipelines, yet issues with figure decomposition and caption enrichment persist. BIOMEDICA and PMC-15M demonstrate the potential of large-scale pretraining but are not publicly available, hindering reproducibility and further exploration. These limitations motivate the development of Open-PMC, our proposed dataset that emphasizes

high-quality curation, including subfigure decomposition and enriched captions, to facilitate robust and generalizable medical VL models.

2.3 Evaluation of Contrastive Vision–Language Encoders

Contrastive learning aligns images and texts in a shared embedding space, and the effectiveness of the learned representations is typically evaluated through tasks that measure cross-modal alignment and semantic generalization. Two widely used evaluation paradigms are *cross-modal retrieval* and *zero-shot classification*. Retrieval evaluates fine-grained matching across modalities, e.g., retrieving the correct caption given an image, while zero-shot classification tests a model's ability to generalize to unseen categories using natural language prompts to guide predictions.

2.3.1 Cross-Modal Retrieval

Cross-modal retrieval has emerged as a foundational benchmark for evaluating vision–language models, particularly those trained using contrastive learning objectives. It assesses a model's ability to retrieve semantically relevant items across modalities, for example, retrieving a caption given an image query (*image-to-text*) or retrieving an image given a text query (*text-to-image*), by comparing representations in a shared embedding space.

Earlier retrieval methods relied on shallow statistical techniques such as Canonical Correlation Analysis (CCA) to align image and text features. More recent approaches employ deep neural networks to learn joint embeddings, optimizing contrastive alignment so that semantically similar pairs are pulled closer and dissimilar pairs are pushed apart in the latent space (29). Leading models, such as CLIP and ALIGN, demonstrated the effectiveness of this approach by achieving strong performance on standard retrieval benchmarks like MSCOCO and Flickr30K, and receiving widespread attention for their scalability and generalization.

Despite its success in the general domain, cross-modal retrieval in biomedical contexts presents unique challenges. Biomedical images often contain subtle visual cues and complex domain-specific language. Recognizing this, recent studies benchmarked contrastive models on Chest X-rays and histopathology datasets, reporting varied retrieval performance across modalities (3; 4). To address robustness concerns, emerging benchmarks such as MediMeta-C systematically assess retrieval under realistic image corruptions, noise, blur, brightness shifts, demonstrating that models fine-tuned on well-curated datasets with subfigure-level alignment exhibit greater degradation resilience (33).

Moreover, a growing body of work has investigated out-of-distribution (OOD) generalization in retrieval: models that perform well on familiar benchmarks may fail when queried with visually or linguistically divergent samples. The WikiDO benchmark was introduced to evaluate OOD cross-modal retrieval robustness using disparate domains, highlighting the importance of evaluating retrieval beyond in-domain similarities (34). Similarly, frameworks like DeepBench explore domain shifts using LLM-generated corruptions, further informing the design of more resilient retrieval systems (35).

In summary, cross-modal retrieval remains a central evaluation for contrastive vision–language models, particularly in the medical domain. Beyond measuring alignment accuracy, recent literature emphasizes the

importance of robustness and OOD generalization,critical considerations when deploying models in real-world biomedical applications.

2.3.2 Zero-Shot Classification

Zero-shot classification has become a standard protocol for evaluating vision–language models trained with contrastive objectives. In this setting, models are assessed on their ability to assign images to semantic categories without any task-specific fine-tuning. Instead of learning a classifier from labeled examples, natural language descriptions of classes are encoded by the text encoder to serve as category prototypes. Classification is then performed by comparing the similarity between an image embedding and these text prototypes, with the label corresponding to the most similar prototype selected as the prediction (26; 27).
This evaluation paradigm directly tests the generalization capacity of pretrained encoders. Whereas retrieval benchmarks focus on instance-level matching, zero-shot classification measures whether the embedding space captures category-level abstractions that can transfer to entirely new tasks. In the biomedical domain, this is particularly important due to the scarcity of annotated data. Prior studies have shown that models trained on large-scale biomedical image–text datasets can achieve competitive zero-shot performance on specialized classification tasks, such as pneumonia detection from chest X-rays or tumor subtyping from histopathology images (3). These results highlight zero-shot evaluation as a critical tool for benchmarking how well medical vision–language models extend beyond their training distributions and support downstream clinical applications.

2.3.3 Robustness Evaluation in Medical Vision–Language Models

Robustness, how well models perform under realistic image corruptions and domain shifts, is an increasingly important evaluation dimension for medical vision–language models. Clinical imaging is often affected by acquisition artifacts, environmental variability, and preprocessing noise, yet standard benchmarks typically assess performance on clean datasets. This gap has motivated the development of corruption-based benchmarks and adaptation strategies tailored to the medical domain.
MediMeta-C (33) systematically applies seven types of perturbations, such as Gaussian noise, motion blur, brightness and contrast changes, zoom blur, pixelation, and impulse noise, at five levels of severity to five medical imaging modalities including microscopy, chest X-rays, and retinal OCT. Combined with MedMNIST-C (36), this framework enables evaluation of medical vision-language models (MVLM) degradation under OOD conditions. Their experiments revealed that many models suffer severe performance drops under corruptions, and that few-shot tuning (RobustMedCLIP) can partially improve robustness while preserving generalization (33).
In another study, robustness of retrieval-based contrastive models, CLIP, MedCLIP, CXR-RePaiR, and CXR-CLIP were benchmarked under an occlusion retrieval task (37). Images were occluded at various levels, and performance on cross-modal retrieval was measured via Recall@K. All models showed sensitivity to occlusion, with domain-specific variants (e.g., CXR-CLIP) outperforming general-purpose CLIP but still experiencing performance degradation proportional to occlusion levels (37).

Beyond the medical domain, broader frameworks like DeepBench have emerged for evaluating robustness in vision–language models across real-world domain shifts (35). DeepBench uses LLM-guided corruption generation and unsupervised metrics (e.g., label-flip probability, mean corruption error) to assess robustness. Though not medical-specific, it highlights the need for domain-aware robustness evaluation, an insight particularly relevant to biomedical applications.

Together, these studies underscore a growing consensus: high accuracy alone is insufficient. Robustness to realistic corruptions is essential for medical MVLMs, and recent work provides both benchmarks and strategies to measure and improve resilience. This review establishes a foundation for integrating robustness evaluations into our own methodology, ensuring our models are not only accurate but dependable under real-world variability.

2.4 Preprocessing of Image–Caption Pairs in Medical Literature

Building high-quality image–caption datasets is a crucial prerequisite for effective vision–language modeling, especially in the biomedical domain. Unlike general-domain datasets such as COCO (38) or LAION (39), biomedical corpora often contain complex multi-panel figures, lengthy and technical captions, and heterogeneous modalities ranging from radiology scans and microscopy to charts and clinical photographs. Without careful preprocessing, these datasets risk introducing noise, misalignment, or spurious correlations into the training pipeline, ultimately limiting the performance and reliability of downstream models.

Preprocessing strategies aim to improve the fidelity of image–caption pairs by addressing two major challenges. First, compound figures must be decomposed into individual subfigures so that each visual unit corresponds more directly to its descriptive text. Second, textual descriptions often exceed the input length of standard encoders, requiring summarization or restructuring to preserve semantic content while remaining compatible with model constraints. Together, these steps ensure that each image is paired with an informative and well-aligned caption, reducing noise in the training signal and enhancing cross-modal representation learning.

2.4.1 Subfigure Extraction as Object Detection

Subfigure extraction can be seen as a specialized case of document layout analysis, a long-standing problem in computer vision and digital libraries. Document parsing tasks such as table detection, diagram segmentation, and page layout understanding share many of the same challenges, including irregular spacing, overlapping elements, and ambiguous boundaries. By framing subfigure extraction as part of this broader research area, researchers have been able to adapt techniques from document analysis to scientific figure processing, thereby situating the problem within the growing field of multimodal document understanding.

In the biomedical domain, compound figures are especially prevalent in research articles, often comprising multiple related visual elements such as X-rays, histological slides, diagnostic scans, and graphical plots. These subfigures are typically arranged in grid or column layouts but can also exhibit irregular spacing, overlapping regions, or embedded annotations. Without accurate separation of these subfigures, vision-language models are likely to learn spurious associations or fail to align visual features with their corresponding textual

descriptions. As a result, subfigure extraction plays a critical role in the preprocessing pipeline of medical image–text datasets.

Classical Approaches Prior to the advent of deep learning, compound figure separation in scientific articles primarily relied on handcrafted image analysis techniques. These methods leveraged low-level visual cues such as whitespace, edges, or geometric structures to infer the boundaries between subfigures. While effective in some cases, classical approaches generally struggled with noisy, irregular, or complex layouts. Below, we summarize the main categories of these early methods.

One of the earliest strategies for compound figure separation relied on the observation that many figures are arranged in grids with uniform spacing. By scanning for horizontal and vertical whitespace bands, algorithms could detect separator regions and use them to partition the figure into subpanels (40). This approach worked well when figures exhibited consistent margins or clear background spacing between subfigures, as is common in many types of scientific plots or microscopy layouts. However, it was highly sensitive to irregular spacing, narrow gaps, or cases where whitespace was occluded by annotations, legends, or background textures. As a result, separator-based methods often failed in biomedical figures, where panels may be tightly packed or visually complex.

Another line of research approached the problem by detecting the boundaries of subfigures using edge-based operators. Techniques such as gradient-based edge detection (e.g., Sobel (41), Canny (42)) or boundary tracing could highlight the visual contours between panels. Compared to whitespace detection, edge-based methods were more effective at handling layouts with minimal gaps, since they could identify panel borders directly (43). Nonetheless, these methods were highly sensitive to noise, image artifacts, and broken edges, which frequently resulted in fragmented or incomplete boundaries. The variability of figure styles across different journals further exacerbated these weaknesses.

In summary, classical approaches laid the foundation for automated compound figure analysis by exploiting simple visual cues such as whitespace and edges. While these methods were computationally lightweight and effective under ideal conditions, they lacked the robustness to handle the irregular and annotation-rich layouts commonly encountered in biomedical publishing. These limitations motivated the shift toward learning-based methods, particularly object detection frameworks, which offer greater adaptability and accuracy.

Deep Learning-Based Detection To overcome the limitations of rule-based methods, recent research has reformulated subfigure extraction as an object detection task. In this paradigm, each subfigure is treated as an object within a larger compound figure, and the goal is to predict bounding boxes around each panel. By leveraging advances in general-purpose object detection, these approaches achieve substantially greater robustness across diverse figure layouts. Below, we review several families of deep learning models that have been applied or adapted for subfigure extraction.

The "You Only Look Once" (YOLO) family of detectors (44) introduced real-time object detection by framing detection as a single-stage regression problem. Instead of generating region proposals as in earlier approaches, YOLO directly predicts bounding box coordinates and class probabilities from a grid-based representation of the image. For subfigure extraction, YOLO's speed and efficiency make it attractive for

large-scale processing of scientific articles. (45; 46) applied YOLO variants to detect subfigures in compound figures, demonstrating that single-shot detectors can handle both regular and irregular layouts with higher robustness than band- or edge-based methods. However, YOLO can struggle with very small or tightly packed subfigures, where bounding box predictions may overlap or fail to separate adjacent panels.

More recently, transformer-based detectors such as the DEtection TRansformer (DETR) (47) have been employed for subfigure extraction. Unlike CNN-based detectors, DETR formulates detection as a set prediction problem and replaces region proposals with a transformer encoder–decoder architecture that attends to different parts of the image. Lin et al. (5) trained a DETR model on the MedICaT dataset (48), which consists of 2,069 manually annotated biomedical compound figures. The attention mechanism of DETR enables improved spatial reasoning, allowing the model to capture irregular layouts and non-uniform spacing more effectively than convolution-based detectors. Nevertheless, a limitation of MedICaT is its heavy skew toward radiology images (72%), with fewer examples of microscopy (13%) and visible light photography (3%), raising questions about cross-modality generalization.

In summary, the application of object detection architectures such as YOLO and DETR has significantly advanced the robustness of subfigure extraction. These methods have shifted the field from heuristic-based pipelines to end-to-end trainable systems capable of handling the wide variability present in scientific and biomedical figures.

Challenges and Synthetic Data One of the most significant challenges in training object detection models for subfigure extraction is the scarcity of large, high-quality annotated datasets. Constructing such datasets requires manually drawing bounding boxes around each subfigure, a process that is both time-consuming and labor-intensive. In the biomedical domain, the annotation burden is even greater: accurate labeling often requires domain expertise to correctly distinguish between panels that may contain similar visual structures but represent different modalities or experimental conditions. Furthermore, biomedical figures frequently include embedded annotations such as arrows, labels, legends, and scale bars, which complicate the task of defining precise subfigure boundaries. These factors make large-scale manual annotation impractical, limiting the availability of diverse and representative training data.

To mitigate this issue, researchers have turned to the generation of *synthetic compound figures*, in which subfigures are programmatically arranged into larger composite images using predefined or randomized layout rules. In this setting, individual subfigures are drawn from existing datasets of single-panel images and then assembled into grids, columns, or irregular arrangements that mimic the layouts commonly observed in scientific publications. Since the composition process is automated, ground-truth bounding boxes for each subfigure are available by design, eliminating the need for manual annotation. (45; 46) demonstrated that models pretrained on large collections of synthetic compound figures achieve improved robustness and generalization when fine-tuned on smaller human-annotated datasets. This strategy enables training object detectors at scale, even in domains where annotated resources are scarce.

2.4.2 Text Preprocessing and Summarization for Vision–Language Models

A practical challenge in vision–language modeling is that textual descriptions often exceed the input length limits of commonly used text encoders, such as BERT (49). This problem is particularly pronounced in scientific and biomedical domains, where figure captions, abstracts, or full-text passages can be lengthy and densely descriptive. Feeding the raw text directly into encoders is infeasible due to tokenization constraints, and truncation risks discarding essential semantic information needed for cross-modal alignment.
Several strategies have been explored in prior work to address this issue. One approach is to truncate captions or retain only selected segments (e.g., the first sentence), but this often results in significant information loss (50). Another direction leverages hierarchical encoders, where text is broken into smaller chunks, each encoded separately before being aggregated (51). More recently, large language models (LLMs) have been employed as preprocessing tools to generate abstractive or extractive summaries of long captions or articles. This summarization step condenses the input while preserving the most relevant content, ensuring compatibility with existing text encoders without substantial loss of meaning. For example, (52) highlight the role of LLM-based summarization as an effective way to reduce input length while maintaining task relevance.

2.4.3 Image Modality Classification in Biomedical Datasets

An important step in curating medical vision–language datasets is the classification of images by modality. Different imaging modalities, such as radiology, microscopy, and visible light photography, carry distinct visual properties and semantic interpretations, making modality labels crucial for both dataset organization. Several prior works have addressed modality classification during dataset construction. ROCO (6) introduced a taxonomy that distinguishes between radiology and non-radiology figures, relying primarily on metadata and manual curation. PMC-OA (5) extended this by proposing automated pipelines that parse PubMed Central articles and apply heuristic filtering, though without explicit large-scale modality annotations. More recent efforts, such as BIOMEDICA (4), combine clustering methods with pretrained encoders and expert review to assign modality labels, thereby improving coverage across radiology, microscopy, and visible light photography. Similarly, PMC-15M (3) and related large-scale collections recognize the importance of modality diversity but lack fine-grained classification.

2.5 Investigating Embedding Distributions in Multimodal Models

Moving beyond task-oriented metrics like retrieval or zero-shot performance, recent work explores how training choices influence the geometry of learned embeddings. Understanding whether models trained on different data occupy similar or distinct regions in the representation space is especially relevant for vision–language models in medical domain. Below, we survey key techniques for comparing embedding distributions and representation alignment.

2.5.1 Visual Inspection via Embedding Projections

Low-dimensional projection tools provide intuitive insights into how embeddings organize by semantic categories or modalities. t-SNE (t-Distributed Stochastic Neighbor Embedding) (53) is a widely used nonlinear dimensionality reduction technique for visualizing high-dimensional data in two or three dimensions. The algorithm operates by modeling high-dimensional pairwise similarities with a Gaussian-based probability distribution, and then seeks a low-dimensional embedding whose pairwise similarities, modeled using a Student t-distribution, match the high-dimensional ones. Optimization occurs via minimizing the Kullback–Leibler divergence between these two distributions, encouraging similar points to remain close while allowing dissimilar points to spread apart. This design excels at revealing local clustering structures, making it particularly effective for identifying subtle grouping phenomena. In biomedical contexts, researchers have used t-SNE to inspect how well models align embeddings from different medical modalities. The MEDBind (54) model in particular employed t-SNE to show that this model brings chest X-ray and ECG embeddings closer together, demonstrating stronger intermodal alignment compared to baseline models.
UMAP (Uniform Manifold Approximation and Projection) (55) is a manifold learning–based dimensionality reduction technique. It operates in two phases: first, constructing a weighted nearest-neighbor graph in the high-dimensional space to model local relationships; second, optimizing a low-dimensional embedding via a cost function that preserves both local and global structure. Although UMAP is more commonly used in genomics and single-cell omics (56), it has been adopted across biomedical research to visualize high-dimensional clinical data. For instance, it has been used to delineate patient phenotypes from electronic health records (57), revealing clinically meaningful clusters corresponding to different chief complaints. In the realm of vision–language modeling, UMAP has been applied to compare embeddings from different modalities or training regimes. For instance, in foundation models like mPLUG-2 (58), UMAP visualizations showed tighter alignment between paired image and text representations in shared latent space, suggesting stronger cross-modal alignment due to model architecture design.

2.5.2 Studying Representation Alignment via Embedding Comparison

When comparing how different models encode data, it is essential to go beyond task-level metrics and probe internal representation geometry. Several quantitative tools have been proposed in the broader machine learning literature to assess representation similarities, especially in multimodal and cross-domain settings.
CKA (Centered Kernel Alignment) (59) is a widely adopted metric for comparing internal representations of neural networks. It computes the alignment between similarity (Gram) matrices of two embedding sets and is invariant to orthogonal transformations and isotropic scaling. This makes it robust for comparing features from differently initialized models or those trained using different losses. CKA has become a standard tool for probing representation similarity. In vision–language representation learning, CKA has been leveraged to quantify whether pretraining pipelines that differ in data composition or curation yield similar embedding geometry. For instance, the study (60) used CKA to compare encoder alignment across unaligned vision and language models, revealing that some pairs exhibit semantic correspondence even without joint training. More recently, (61) used CKA to select vision–text encoder pairs with strong alignment before fine-tuning, showing that higher CKA was predictive of better downstream retrieval alignment. CKA's limitations have

been studied. (62) show it can be sensitive to outliers or transformations that preserve linear separability, potentially overstating representation similarity; remedying this may require careful feature preprocessing or complementary metrics.

MMD (The Maximum Mean Discrepancy) (63) offers a principled framework for understanding how differences in dataset curation via rigorous, quantitative analysis of embedding distributions. MMD is a kernel-based, nonparametric two-sample test that compares the mean embeddings of two distributions in a reproducing kernel Hilbert space (RKHS). For distributions P and Q, and a characteristic kernel k, MMD is defined as:

$$\mathrm{MMD}(P,Q) = \left\| \mathbb{E}_{x\sim P}[\varphi(x)] - \mathbb{E}_{y\sim Q}[\varphi(y)] \right\|_{\mathcal{H}}.$$

Its empirical estimator is:

$$\widehat{\mathrm{MMD}}^2 = \frac{1}{n^2}\sum_{i,j} k(x_i,x_j) + \frac{1}{m^2}\sum_{i,j} k(y_i,y_j) - \frac{2}{nm}\sum_{i,j} k(x_i,y_j),$$

where $\{x_i\}$ and $\{y_j\}$ are embeddings from two models. Permutation testing provides significance by estimating the null distribution. This enables statistical validation of whether the embedding distributions differ, a capability critical for assessing dataset-driven shifts in representation learning.

In addition to distributional and kernel-based methods, graph-based similarity offers an alternative way to compare embedding spaces by examining the local structural organization of data. Nearest-Neighbor Graph Similarity (NNGS) (64) measures the alignment between embedding spaces by constructing a k-nearest-neighbor graph for each and quantifying similarity between these graphs, typically via Jaccard overlap of neighbor sets. Unlike global metrics, NNGS emphasizes local structural consistency and has been demonstrated to correlate with downstream analogical reasoning and zero-shot classification performance, as shown in case studies using GloVe and CLIP embeddings. This method has also been adapted in NLP, where the Nearest-Neighbor Overlap (N2O) (65) metric evaluates the overlap in nearest neighbors across sentence embeddings to assess alignment between different encoders in a task-agnostic manner, highlighting structural similarity without relying on explicit downstream tasks.

3 OPEN-PMC: Curation and Processing

In this work, we introduce **OPEN-PMC** (66; 67), a large-scale biomedical vision-language dataset curated to enhance multimodal learning in medical domains. Unlike previous datasets, OPEN-PMC leverages figure-caption pairs, in-text references and their summary, and additional contextual signals extracted from full-text biomedical research articles. Our dataset construction builds on and extends the PMC-OA pipeline (5), integrating novel processing steps including improved subfigure extraction, in-text reference linking, image modality classification, caption segmentation, and contextual summarization using large language models such as GPT-4o (68). Figure 1 shows an overview of our pipeline.

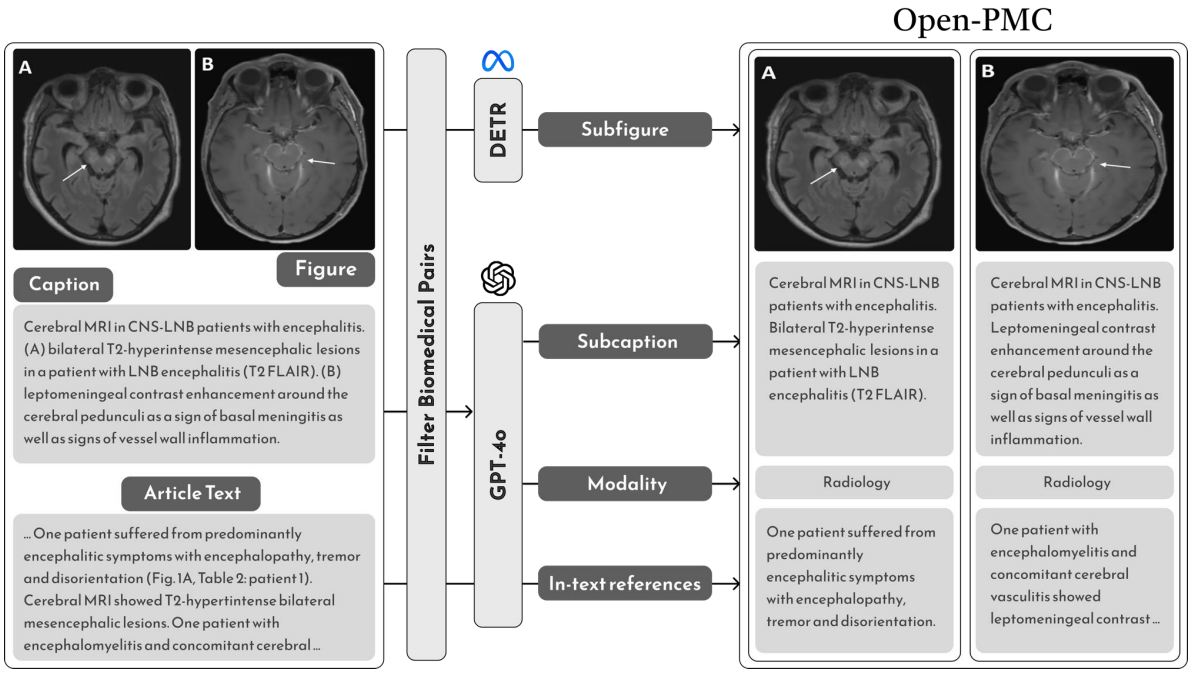

Figure 1: Overview of OPEN-PMC dataset curation pipeline.

3.1 Data Collection and Preprocessing

PubMed Central (PMC) serves as the most widely used source for biomedical vision–language dataset construction due to its scale, open accessibility, and inclusion of diverse imaging modalities. PMC is a free digital repository of full-text biomedical and life sciences journal literature hosted by the U.S. National Institutes of Health's National Library of Medicine. Previous datasets such as ROCO (6), PMC-OA (5), PMC-15M (3), and BIOMEDICA (4) have all relied on PMC articles as their primary source of figures and captions. However, raw PMC data presents several challenges: figures are often compound and unannotated, captions can be lengthy and noisy, and article XML structures are not always consistent. Without careful curation, these issues introduce substantial noise into downstream training.

Our pipeline was initially based on Build-PMC-OA (5), which automated parsing and pairing of figures with captions. We extended this pipeline with stricter semantic refinement stages. Figures, captions, and

article metadata such as PubMed ID (PMID) and PMC ID were extracted using XML parsing and regular expressions. Articles with malformed XML, missing figure–caption structures, or broken formatting were excluded, yielding a corpus of 16.7 million figure–caption pairs.

During this work, the BIOMEDICA dataset (4) was released, containing 24 million parsed figure–caption pairs from PMC articles. BIOMEDICA provided a more complete and bigger base collection (BIOMEDICA is without any filtering and data processing steps and it contains only the raw image-caption pairs along with the full in-text references). Therefore, in order to save resources, we adopted BIOMEDICA as our starting point for large-scale collection. This allowed us to bypass the costly raw XML parsing stage, while still applying our own medical relevance filtering, subfigure extraction, and text augmentation. Our final dataset therefore combines the scale of BIOMEDICA with additional semantic quality controls tailored to medical imaging applications.

3.2 Initial Medical Filtering

A large proportion of figures in PMC are non-medical, including bar plots, line graphs, statistical charts, and schematic diagrams. Processing these non-medical figures through subfigure extraction and caption alignment would be both computationally expensive and semantically irrelevant for medical vision–language learning. As a preliminary step, we therefore implemented a medical filtering stage.

We constructed a curated lexicon of medical terms spanning anatomical entities (e.g., “lung”, “retina”, “dermis”), imaging modalities (e.g., “CT”, “MRI”, “ultrasound”), and pathological concepts (e.g., “tumor”, “lesion”, “metastasis”). Each caption was checked for the presence of at least one term from this lexicon. Captions containing no medical terms were flagged as non-medical, and their associated figures were excluded. This process reduced the dataset from 16.7 million to 880,294 figure–caption pairs, reflecting a 95% reduction and underscoring the prevalence of non-medical content in PMC.

To validate the reliability of this keyword-based filtering, we randomly sampled 300 figures and manually annotated them as medical or non-medical. The results showed that our method achieved 99% precision, confirming that keyword matching is a low-cost yet highly effective filtering strategy. While keyword-based approaches may miss rare cases where captions use unconventional terminology, the method offered a strong trade-off between scalability and accuracy. Alternative strategies such as classifier-based filtering or semantic similarity scoring could capture borderline cases but at much higher computational cost. Given our goal of large-scale dataset construction, keyword filtering was the most efficient choice. After this step, we were left with about 800,000 image-caption pairs.

To handle this step for the part that we used BIOMEDICA dataset, we utilized the modality labels provided by the BIOMEDICA dataset to filter out any non-medical images, retaining only those that were explicitly categorized as medical imaging. After this step, we were left with about 6,000,000 image-caption pairs, which we will call PMC-6M, that is, the baseline medical images in PMC.

3.3 Compound Image Decomposition

Scientific articles in biomedical domains frequently contain compound figures, which are composite images consisting of multiple distinct subfigures or panels (e.g., labeled (a), (b), (c), etc.). These panels often

have different image characteristics. For example, a compound figure may combine an X-ray image, a histology slide, and a bar plot summarizing experimental results. Treating such figures as atomic units for vision–language training introduces noise, since real-world images are not like this. To address this, we framed compound image decomposition as an object detection problem: the task of localizing and isolating individual subfigures within larger compound layouts, thereby producing atomic image–text pairs better suited for representation learning.

Our decomposition pipeline evolved through an iterative process. We first deployed a baseline DETR model trained on MedICaT (48) to the 800,000 image-caption pair dataset we first extracted and conducted quantitative and qualitative error analyses. While performance was strong on radiology-dominant content (consistent with MedICaT's skew), we observed frequent failures on underrepresented modalities, especially microscopy and visible-light photography, including missed panels, inaccurate bounding boxes, and under-segmentation in dense cellular imagery. These diagnostics motivated a redesign focused on modality robustness.

In response, we constructed a large, modality-balanced synthetic corpus of compound figures by composing real single-panel biomedical images into diverse layouts with randomized grids, margins, aspect ratios, and label styles. Using this corpus, we retrained the detector with DAB-DETR to improve convergence and localization in cluttered, irregular layouts. The resulting model significantly improved recall and localization across microscopy, endoscopy, dermatology, and visible-light panels, while maintaining radiology performance. Finally, we redeployed the improved detector to the full corpus and applied a post-decomposition figure-type classifier to remove non-medical content before downstream pairing and training. Figure 2 shows the overview of our pipeline for creating synthetic compound figures used to train the DAB-DETR model. A *Sampler* selects single-panel images and layout specifications from their respective pools. A *Labeler* assigns subfigure labels from a predefined label pool (e.g., 1, a, a1, a-1), placing them according to the chosen scheme.

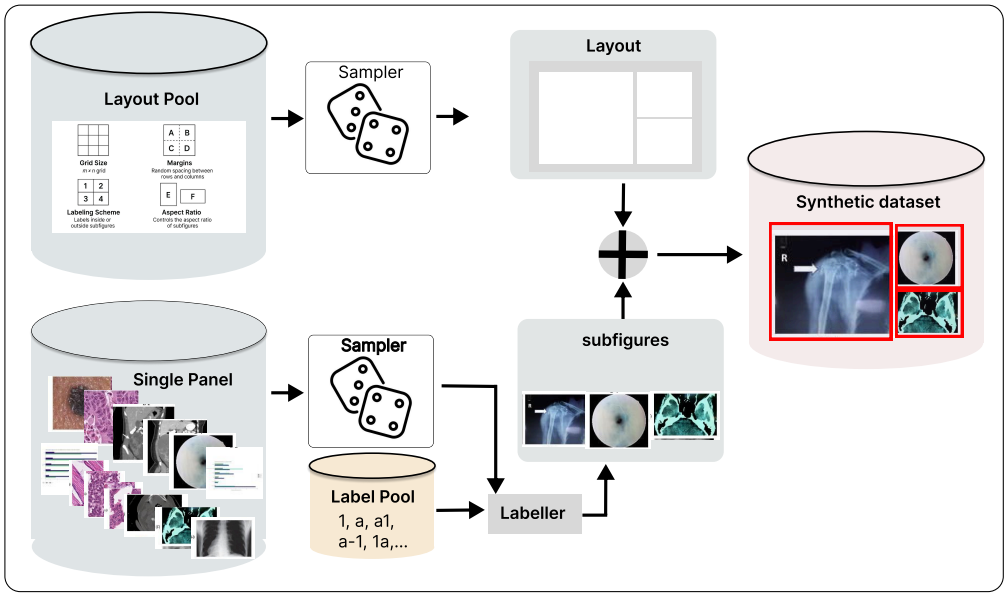

Figure 2: Overview of compound figure generation pipeline

3.3.1 Initial Subfigure Extraction Using MedICaT-DETR

As a starting point, we adopted a DETR (47) model that had been pretrained on the MedICaT dataset (48). MedICaT contains 2,069 manually annotated compound biomedical figures with bounding-box annotations for individual subfigures. Although relatively small, it remains one of the few publicly available resources with high-quality ground truth for this task. The DETR model trained on MedICaT achieved strong performance on its validation set, reporting a mean Average Precision (mAP) of 0.94. However, its training distribution was highly skewed toward radiology imagery (72% radiology, 13% microscopy, 3% visible light photography), leading to modality bias.

When applied to a broader and more diverse biomedical corpus, the MedICaT-DETR model exhibited significant limitations. While it performed reliably on radiology subfigures, performance was substantially lower for microscopy and visible light panels, with frequent missed detections and inaccurate bounding boxes. For example, dense microscopy images often lacked clear borders, leading to under-segmentation, while endoscopy and dermatology images were sometimes mislocalized or ignored altogether. These shortcomings underscored the need for a more generalized decomposition model trained on a dataset that better reflected the diversity of biomedical imaging modalities.

3.3.2 Synthetic Compound Figure Generation

The key bottleneck in advancing subfigure extraction lies in the scarcity of large-scale annotated training data. Manual annotation of subfigure boundaries is labor-intensive and requires domain expertise, particularly in biomedical figures where panels can be irregular, overlapping, or cluttered with embedded labels. To address this, we developed a scalable synthetic data generation pipeline that programmatically composes single-panel biomedical figures into compound layouts.

Our pipeline generated over 500,000 synthetic compound figures by randomly sampling real single-panel biomedical images and arranging them into multi-panel layouts. Layouts were parameterized along several axes: grid size (e.g., 2×2, 3×3, or irregular configurations), inter-panel margins, alignment, and orientation. The number of subfigures per compound ranged from 2 to 9, reflecting the variability seen in published articles. Importantly, each synthetic compound was automatically annotated with ground-truth bounding boxes, providing fully supervised labels without manual intervention.

To ensure modality diversity, we balanced the composition process across radiology, microscopy, dermatology, endoscopy, and visible light photography images. This balance helped mitigate the modality bias observed in MedICaT, enabling the detector to learn features relevant to underrepresented biomedical domains. Furthermore, we introduced variability in spacing, aspect ratios, and embedded labeling schemes (e.g., subfigure labels "(a), (b), (c)" or "1, 2, 3") to mimic the heterogeneity of real compound figures. This diversity was critical for achieving robust generalization.

In designing this synthetic dataset, particular emphasis was placed on replicating edge cases that commonly occur in real biomedical publications but are often neglected in standard grid-based generation. For instance, we generated compound figures with non-symmetric layouts, irregular panel aspect ratios, and variable inter-panel spacing, reflecting the diversity seen in actual articles. We also constructed compounds mixing different modalities within the same figure, such as radiology paired with microscopy or dermatology panels,

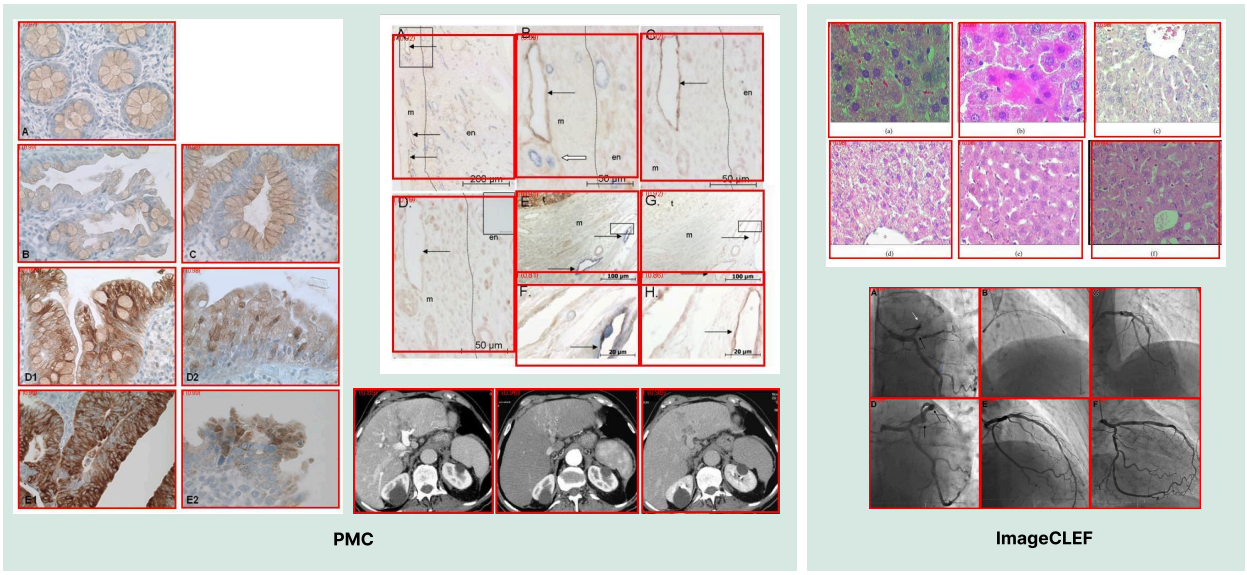

Figure 3: Qualitative results of subfigure detection using our DAB-DETR model. **Left** Real-world biomedical compound figures from PMC articles in PMC-6M (also BIOMEDICA). **Right** Examples from the ImageCLEF 2016 benchmark.

to model the heterogeneous composition typical of multi-experiment studies. Additional variations included partially overlapping panels, embedded annotations (e.g., arrows or labels), and inconsistent background colors. By deliberately incorporating such challenging configurations, the synthetic corpus provided stronger supervision for the detector, reducing brittleness and improving its robustness to the full spectrum of figure designs encountered in biomedical literature. Source subfigures are drawn from well-known benchmark datasets such as ROCO (6), SICAP (69), HAM10000 (70), PathMNIST and RetinaMNIST from MedMNIST (71; 72), PAD-UFES-20 (73), and PlotQA (74) as listed in Table 1. Figure 3 showcases examples from the ImageCLEF 2016 dataset and from a subset of PMC-6M, illustrating accurate detection of distinct subfigures across diverse panel layouts and content types.

Table 1: Datasets used for synthetic subfigure generation.

Split	Radiology	Histopathology	Dermatology	Retina	Plots
Train	ROCO	SICAP	HAM10000	RetinaMNIST	PlotQA
Sample Size	65422	18783	10015	1080	60000
Validation	ROCO (test)	PathMNIST	PAD-UFES-20	RetinaMNIST (val)	PlotQA (val)
Sample Size	8176	10004	2298	120	10000

3.3.3 Training a Modality-Robust DAB-DETR Model

Using the synthetic dataset, we trained a transformer-based object detector based on Dynamic Anchor Box DETR (DAB-DETR) (75). DAB-DETR extends the original DETR architecture by introducing dynamic anchor boxes as query embeddings, which improves convergence speed and localization accuracy, particularly

in dense or cluttered layouts. This property made DAB-DETR well-suited to the irregular panel configurations observed in biomedical compound figures. The model was trained on 500,000 synthetic compounds and validated on both a held-out synthetic set and the ImageCLEF 2016 compound figure separation benchmark (76).

Compared to the MedICaT-DETR baseline, our DAB-DETR model achieved higher mAP across all modalities, with particularly large improvements in microscopy and visible light photography. Qualitatively, the model was better at handling overlapping panels, irregular boundaries, and embedded legends, demonstrating greater robustness to the structural noise inherent in biomedical figures. The results can be found in Table 2.

Once trained, this DAB-DETR model was applied to decompose all compound figures in our dataset, producing atomic subfigure crops for downstream processing.

Table 2: Performance comparison on two datasets using mAP and F1 metrics.

Model	**Synthetic Validation**		**ImageCLEF 2016**	
	mAP (%)	**F1 (%)**	**mAP (%)**	**F1 (%)**
Previous model (MedICaT)	33.22	73.18	28.20	64.85
Our model (DAB-DETR)	**98.58**	**99.96**	**36.88**	**73.55**

3.3.4 Post-Decomposition Classification and Filtering

Decomposition alone does not guarantee medical relevance, since compound images often contain non-medical figures such as plots, tables, flowcharts, or software interface screenshots. To filter out these irrelevant figures, we applied a figure-type classifier trained on the DocFigure dataset (77), which consists of over 3,000 labeled figures spanning 28 categories. The classifier was implemented using a ResNet-101 backbone and trained for multi-class prediction. Among the categories, one was designated as "Medical," encompassing radiology, microscopy, and other diagnostic imagery.

For each subfigure, the classifier produced a ranked probability distribution over the 28 categories. We retained subfigures if the "Medical" category appeared within the top four predictions. This ranking-based cutoff was motivated by prior work (5). In practice, this allowed us to recover borderline cases (e.g., microscopy images misclassified as "illustrations") without admitting a large volume of irrelevant diagrams. Non-medical figures were excluded at this stage, ensuring that only clinically meaningful images contributed to the final dataset.

3.4 Caption Segmentation and Subfigure Alignment

Multi-panel figures in biomedical publications are almost always accompanied by captions that describe all subfigures together in a single block of text. Such captions often begin with a general description of the figure, followed by subfigure-specific descriptions introduced with markers such as "(a), (b), (c)" or numerical labels. To build a high-quality dataset for our experiments we need to extract the subcaptions of

each subfigure. Subcaptions allow us to examine how fine-grained text–image alignment influences model performance.

3.4.1 Caption Segmentation with Large Language Models

We leveraged GPT-4o to segment multi-part captions into subcaptions corresponding to each subfigure. The model was prompted with structured heuristics to identify textual markers such as "(a), (b), (c), (d)," Roman numerals ("i, ii, iii"), or numeric labels ("1, 2, 3"). In addition to explicit labels, GPT-4o was instructed to detect linguistic transitions that indicate a shift in visual focus (e.g., "In contrast," "Meanwhile," "On the right," "In the bottom panel"). This hybrid approach,combining structural cues with discourse markers,enabled the system to capture both explicit and implicit references to subfigures. For captions lacking clear structure, GPT-4o defaulted to returning the original unsegmented caption to preserve semantic integrity rather than risk over-segmentation. The prompt used for subcaption extraction is in Section A.

3.4.2 Visual Label Detection and Recognition

Textual segmentation alone is insufficient without knowing which subfigure each segment refers to. To bridge this gap, we adopted components from the Exsclaim! pipeline (78), which specializes in linking caption references to visual labels inside compound figures. First, YOLOv3 (79) was used to detect visual markers (e.g., "(a), (b), (c)") embedded in subfigure corners. These detections localized the regions of the image that contained panel labels. Next, a ResNet-152 classifier trained for character recognition was applied to the cropped marker regions, enabling robust identification of the specific letter or number used as the label. This two-step process allowed us to anchor subcaptions to explicit subfigures with high accuracy, even when label placement varied across articles.

3.4.3 Alignment of Subcaptions to Subfigures

After both caption segmentation and visual label detection, we aligned segmented text to subfigures by matching the recognized marker tokens (e.g., "(a)" in text with "a" detected in the image). In most cases, this alignment was direct. For figures where visual labels were absent or illegible, we assigned the entire original caption to each subfigure as a fallback strategy, ensuring that no subfigure was discarded. Although this fallback introduces some redundancy, it preserves contextual coverage and prevents data loss. In rare cases where captions described subfigures without consistent markers (e.g., "left" vs. "right" panels), GPT-4o's discourse heuristics were used to assign text segments based on spatial ordering of subfigures within the compound image.

3.4.4 Edge Cases and Quality Control

Biomedical figures presented several edge cases that required careful handling. Some captions contained nested references (e.g., "(a–c) show microscopy images, while (d) is an MRI scan"), which required splitting grouped descriptions into multiple aligned pairs. Other captions embedded panel references within long

sentences, which GPT-4o handled by splitting clauses into smaller subcaptions. We also encountered ambiguous situations, such as subfigures lacking any textual mention. In these cases, we defaulted to assigning the global caption. To evaluate segmentation quality, we manually inspected 500 figure–caption pairs, observing that subcaption segmentation achieved a precision of 93% in correctly isolating subfigure-specific text segments. This confirmed the effectiveness of combining LLM-driven segmentation with visual label detection.

3.5 Textual Augmentation and Contextualization

While captions provide immediate descriptions of figures, they are often brief and omit important contextual details that appear in the surrounding article text. In biomedical literature, figure mentions in the article body frequently contain richer information, such as detailed explanations of clinical findings, methodological nuances of an experiment, or interpretive insights that frame the visual evidence in a broader scientific narrative. Relying solely on captions therefore risks underrepresenting the semantic depth of figure–text pairs, whereas incorporating in-text references produces a more comprehensive alignment. This augmentation is especially critical for medical datasets, where subtle differences in interpretation (e.g., "mild infiltration" vs. "severe infiltration") can substantially influence downstream clinical tasks.

3.5.1 Extraction of In-Text References

We extracted in-text references by leveraging XML cross-reference tags and figure identifiers available in PMC articles. Each figure and its subfigures were linked to all textual mentions scattered across the article body, capturing references that appeared in the methods, results, or discussion sections. This process produced a pool of candidate contextual sentences for each visual element. To isolate relevant fragments, paragraphs containing figure mentions were segmented into sentences using rule-based regular expressions. The dispersed nature of scientific writing introduced several challenges. Figure mentions were often scattered across multiple paragraphs, sometimes separated by hundreds of words. Furthermore, not all references were equally informative: some merely repeated the figure number without adding content, while others provided indirect contextual cues without explicit figure mentions. This heterogeneity made it difficult to simply concatenate all figure-related text, as doing so would introduce noise and redundancy. Another challenge was balancing detail with scalability, since large amounts of raw text could exceed the input length of downstream encoders and lead to inefficiencies during training.

3.5.2 Summarization of Contextual References

To address these issues, we aggregated all candidate text fragments associated with each figure or subfigure and processed them using the GPT-4o-mini API to produce focused summaries. The summarizer was guided to extract clinically or experimentally meaningful information, including diagnostic findings, methodological details, and interpretative statements, while filtering out tangential or repetitive content. Each summary was constrained to a maximum of 250 tokens, providing a consistent input length that ensured compatibility with

transformer-based text encoders. By condensing dispersed figure references into a unified representation, the summarization step improved semantic clarity and reduced noise. The prompt used for this task is in A.

For compound figures, textual augmentation was applied at the subfigure level. Summarized contextual references were linked to the specific subcaption generated during caption segmentation, ensuring that each subfigure was paired not only with a localized description but also with its broader clinical or experimental context.

The final output of this stage was a set of enriched image–text pairs that combined three layers of supervision: (1) the raw caption, (2) the segmented subcaption, and (3) the contextual summary. This hierarchical text representation enabled flexible experimentation in downstream training, allowing us to test the impact of different text granularities on encoder performance. Moreover, by incorporating context from the article body, our dataset more closely mirrored how figures are used in real-world biomedical communication, where visual evidence is rarely interpreted in isolation but instead situated within a broader narrative.

3.6 Image Modality Classification

In addition to segmenting and contextualizing figure–caption pairs, it is important to annotate each image with its underlying medical modality. Modality information provides valuable metadata that can support modality-specific analyses, stratified evaluations, and controlled experiments examining whether certain encoders specialize in different imaging domains. For example, radiology images such as CT or MRI scans convey very different visual characteristics and semantic cues compared to microscopy slides or dermatology photographs.

3.6.1 Automated Modality Classification with GPT-4o-mini

To perform large-scale annotation, we employed GPT-4o-mini as an automated image modality classifier. Each image was first categorized into one of two high-level groups: *diagnostic* or *non-diagnostic*. The diagnostic branch was then refined into three subcategories: (1) *Radiology*, which encompassed common imaging modalities such as X-ray, ultrasound, computed tomography (CT), and magnetic resonance imaging (MRI); (2) *Visible Light Photography (VLP)*, which included dermatology, endoscopy, and other clinical photographs; and (3) *Microscopy*, covering both light and fluorescence microscopy. This hierarchical classification scheme reflects the dominant modalities present in biomedical publishing and aligns with prior work on modality-aware representation learning. This classification was performed at the subfigure level following decomposition, ensuring that each atomic visual unit received its own label. Some figures were visually ambiguous. For example, a poorly scanned grayscale image might be confused between an X-ray and a microscopy slide. To mitigate these issues, GPT-4o-mini was guided with explicit modality definitions and reference examples during inference, helping to reduce uncertainty in borderline cases. The prompt used for this task is in A.

Table 3: Comparison of medical paired image-text datasets.

Dataset	Size (M)	In-text Ref	Summ Refs	Subfigures	Medical Only	Modality	Open Access
ROCO	0.08	✗	✗	✓	✓	✓	✓
PMC-OA	1.6	✗	✗	✓	✓	✗	✓
PMC-15M	15	✗	✗	✗	✗	✗	✗
BIOMEDICA	24	✓	✗	✗	✗	✓	✓
OPEN-PMC-2M	2.2	✓	✓	✓	✓	✓	✓

3.6.2 Validation with Expert Review

To evaluate the reliability of the automated classifications, we conducted a validation study with three expert reviewers. The reviewers independently annotated a random sample of 1,000 images drawn from across the corpus. Agreement between automated predictions and majority-vote human labels was 87%, indicating strong reliability given the diversity of imaging modalities. Disagreements were most common in edge cases, such as ambiguous grayscale figures or images spanning multiple modalities. In cases of disagreement among reviewers themselves, qualitative inspection revealed that even human annotators sometimes disagreed on modality boundaries, suggesting that the 87% agreement rate is a conservative estimate of classifier reliability.

3.7 Final Curated Datasets

In this work, we use two curated dataset variants: **OPEN-PMC-2M** and **OPEN-PMC-18M**. Both are produced using the same overall curation framework, but differ in their source data, subfigure extraction models, and the level of downstream processing applied.

OPEN-PMC-2M is based on our independently collected PMC-OA corpus and uses the MedICaT-trained DETR model for subfigure extraction. This dataset includes full sub-caption extraction, subfigure–caption alignment, and in-text reference summarization, enabling us to conduct detailed ablation studies and assess the impact of each component on pre-training performance. This dataset was curated initially to enable the empirical assessment of how each processing step contributes to overall model performance.

OPEN-PMC-18M, in contrast, starts from the BIOMEDICA (4) release, which provides a comprehensive parsed version of PMC data. For this version, we first filter out the non-medical images and get **PMC-6M**, then employ our modality-robust DAB-DETR model trained on a large-scale synthetic compound figure dataset. Due to resource limitations and experiments in Section 4, we did not perform sub-caption extraction, alignment, or in-text reference summarization for this dataset. However, OPEN-PMC-18M offers both greater scale and improved modality coverage.

Throughout the manuscript, we refer to these datasets by name when discussing results: OPEN-PMC-2M represents the smaller, fully processed version used for controlled analysis, while OPEN-PMC-18M represents the larger, modality-generalized version used to combine high quality with broad coverage. Table 3 shows a comparison between OPEN-PMC-2M and the previous datasets.

4 Experiments

We pre-trained the vision and language encoders using the curated OPEN-PMC datasets to initialize them with medically relevant representations. During pre-training, the objective was to align each image with its paired caption in a shared embedding space, so that semantically related visual and textual inputs are drawn closer together while unrelated pairs are pushed apart. This alignment provides the basis for evaluating how different dataset designs influence representation learning. By holding the encoder architecture fixed and varying the training data, we are able to directly measure the effect of dataset curation on the quality of learned embeddings and their performance on downstream medical tasks.

In this section, we first describe our training and evaluation setup. We then present a series of experiments designed to assess how each curation step, introduced to improve dataset quality, affects the performance of the trained model. Then, we report the results obtained using the best configuration identified in these experiments, comparing models trained on OPEN-PMC-2M and OPEN-PMC-18M against previous state-of-the-art baselines. Finally, we conduct a representation analysis to examine the distinctiveness and stability of embeddings learned from OPEN-PMC, providing deeper insights beyond standard benchmark performance.

4.1 Model Architecture Details

Our training setup followed a single unified configuration, in which one image encoder was used to process all biomedical image modalities and one text encoder was used to process the paired textual data. This design choice was motivated by prior work such as BiomedCLIP (3) and Biomedica (4), which adopt the same strategy to enable fair comparisons across modalities and datasets. By embedding radiology, microscopy, and visible-light images into the same representational space alongside their captions, the model is able to learn cross-modal associations that are both modality-agnostic and semantically consistent. To align the outputs of the two encoders, we employed a standard contrastive learning objective following the vanilla contrastive loss formulation (26), which encourages matched image–text pairs to be pulled closer together while pushing apart non-matching pairs within the same training batch. This framework not only facilitates direct comparison with existing medical vision–language models but also provides a robust foundation for evaluating the effect of dataset curation on representation quality.

Pre-training of the vision and text encoders was therefore designed to establish a shared embedding space for figure–caption data. Formally, we denote the image encoder as f_I and the text encoder as f_T. Given an input figure and its associated caption, f_I maps the visual data to an embedding vector, while f_T maps the textual data to another vector in the same space. The contrastive objective then enforces that semantically corresponding pairs $(f_I(\text{image}), f_T(\text{text}))$ are positioned close together, while non-corresponding pairs are positioned further apart. Through this process, the model learns representations that capture meaningful associations between biomedical figures and their descriptions.

4.1.1 Vision Encoder

For the visual backbone, we employed a Vision Transformer (80) (ViT-B/16) architecture, which divides each input image into non-overlapping patches of size 16×16 pixels and projects them into a sequence of

embedding vectors. These patch embeddings are augmented with positional encodings and passed through a stack of 12 transformer layers, each consisting of multi-head self-attention and feed-forward blocks. The model was initialized with weights pre-trained on ImageNet-21k, providing a strong prior for generic visual features. This initialization gives us a more robust baseline to train on using our curated dataset (3).
Each image undergoes a preprocessing pipeline in which it is randomly cropped and resized to a fixed resolution of 224×224 pixels. Pixel values are then normalized using channel-wise mean and standard deviation values for the red, green, and blue channels. During evaluation, a deterministic transformation is applied: the image is resized so that its shortest side is 224 pixels, followed by a center crop to produce a 224×224 image. The resulting image is passed through the image encoder, which produces a 768-dimensional representation, from which we retain only the `[CLS]` token as the representation of the entire image. This vector is then passed through a linear projection layer, mapping it into a 512-dimensional shared embedding space aligned with the text encoder. Before computing the contrastive loss, image embeddings were L2-normalized.

4.1.2 Text Encoder

For the textual backbone, we employed a BERT-base architecture consisting of 12 transformer layers, 12 self-attention heads, and a hidden dimension of 768. As our domain-specific encoder, we used PubMedBERT (81), which is pretrained on a large corpus of PubMed abstracts. Importantly, our dataset (Open-PMC-2M), constructed from figure captions and associated texts in PubMed Central articles, does not overlap with the pretraining corpus of PubMedBERT, ensuring no data leakage.
Text preprocessing followed the WordPiece tokenization algorithm (82), using the vocabulary provided by BiomedCLIP (3). If a caption exceeded the encoder's maximum sequence length (512 tokens), it was truncated; otherwise, it was padded to the required length. The resulting token sequence was then passed through the text encoder, yielding contextualized token embeddings of size 768. We retained only the `[CLS]` token as the aggregate representation of the caption. This vector was subsequently mapped into a 512-dimensional shared embedding space via a linear projection layer, ensuring alignment with the image encoder. Before computing the contrastive loss, text embeddings were L2-normalized.

4.2 Pre-training of Vision and Language Encoders

The training of our vision and language encoders was guided by a contrastive learning objective. The central idea of this objective is to associate each figure with its correct caption while distinguishing it from all unrelated captions within the same batch. This design ensures that the learned embedding space captures meaningful cross-modal correspondences between images and text.
Formally, let $B = \{(x_i, t_i)\}_{i=1}^{N}$ denote a batch of N paired training samples, where x_i represents a figure and t_i its corresponding caption. The encoders map each input into a shared representation space: $f_I(x_i)$ for the image and $f_T(t_i)$ for the text. The training objective minimizes the distance between matched pairs while maximizing the distance between mismatched pairs, encouraging correct alignment and discouraging spurious associations.
The contrastive loss is defined as follows:

$$\ell_{\text{con}}(x_i, t_i; B) = -\left(\log \frac{\exp(\langle f_I(x_i), f_T(t_i) \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle f_I(x_i), f_T(t_k) \rangle / \tau)} + \log \frac{\exp(\langle f_I(x_i), f_T(t_i) \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle f_I(x_k), f_T(t_i) \rangle / \tau)} \right), \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes a similarity function such as cosine similarity, and $\tau > 0$ is a temperature parameter that adjusts the sharpness of the similarity distribution. A lower value of τ makes the model more sensitive to subtle differences in similarity scores, while higher values produce smoother distributions.
Multimodal contrastive learning trains encoders f_I and f_T by minimizing Eq. 1 over the pairs in B:

$$\mathcal{L}_{\text{multimodal}} = \min_{f_I, f_T} \mathbb{E}_B \left[\frac{1}{N} \sum_{i=1}^{N} \ell_{\text{con}}(f_I(x_i), f_T(t_i)) \right], \tag{2}$$

All pre-training experiments used a batch size of 2048, distributed across eight A100 GPUs (256 per GPU). The temperature parameter, initialized at 0.07, is learnable but constrained to a minimum of 0.01. Optimization employed AdamW with $\beta_0 = 0.9$, $\beta_1 = 0.98$, $\varepsilon = 10^{-6}$ and weight decay is 0.2. A cosine decay learning rate scheduler with linear warmup (10% of total steps) was used, with a peak learning rate of 10^{-6} and a minimum value near zero. Mixed precision (16-bit) was enabled, and the random seed was set to zero. Pretraining on OPEN-PMC was conducted over 64 epochs.

4.3 Evaluation Tasks

To evaluate the impact of data quality on model performance, the curated dataset was benchmarked against existing larger, less curated datasets. We assess our vision–language encoders on three major dimensions of evaluation: (1) cross-modal retrieval, (2) zero-shot classification, and (3) robustness to input perturbations. Retrieval and zero-shot classification measure how effectively the model can align visual and textual information and generalize to novel categories without task-specific fine-tuning. Robustness, in turn, evaluates the stability of these representations under common image transformations such as rotations, flips, brightness changes, and scaling, factors that frequently occur in biomedical imaging due to variability in acquisition and publication pipelines. Together, these evaluations are carried out across several medical and biological imaging benchmarks, spanning diverse modalities and task settings, providing a comprehensive picture of model performance under both standard and perturbed conditions.

4.3.1 Cross-Modal Retrieval

Cross-modal retrieval involves matching inputs across modalities, specifically, retrieving text given an image query (image-to-text retrieval, or I2T) and retrieving an image given a text query (text-to-image retrieval, or T2I). In both cases, the model encodes the input from one modality and ranks all candidates from the other modality based on the cosine similarity between their embeddings. A successful retrieval indicates that the model has learned to align visual and textual semantics in a shared embedding space.
Let $\mathbf{v}_i = f_v(I_i)$ denote the embedding of image I_i and $\mathbf{t}_j = f_t(T_j)$ denote the embedding of caption T_j, both normalized to unit length. A similarity score between an image and a text is computed via cosine similarity:

$$\text{sim}(\mathbf{v}_i, \mathbf{t}_j) = \frac{\mathbf{v}_i^\top \mathbf{t}_j}{\|\mathbf{v}_i\| \|\mathbf{t}_j\|}.$$

Retrieval procedure. Given a query (image or text), similarities are computed against all items in the opposite modality, and candidates are ranked in descending order. The evaluation is typically performed using metric Recall@K which is the percentage of queries for which the correct match appears in the top K retrieved results. Retrieval directly probes the alignment quality of the joint embedding space. High retrieval performance indicates that semantically corresponding pairs are well-clustered.
We evaluate retrieval performance using Recall@200, which indicate the fraction of correct targets found among the top-200 retrieved results. High scores on these metrics demonstrate the model's ability to capture fine-grained associations between image and text.

Datasets for Retrieval We evaluate retrieval performance on three domain-specific datasets. We make sure to have at least one dataset from each of the modality groups that we categorized our dataset into.

Quilt-1M (83) is a large-scale vision–language dataset tailored to histopathology. The initial version, Quilt, was automatically curated from educational histopathology videos on YouTube, spanning over 1,087 hours and yielding approximately 802,000 image–text pairs. Through the use of a combination of models, large language models, handcrafted algorithms, human knowledge databases, and automatic speech recognition, the creators extracted high-quality image-text alignments from these videos.
To further expand its scale, Quilt was merged with additional sources such as Twitter, scientific publications, and other internet content, resulting in Quilt-1M, a dataset containing one million paired microscopy images and natural language descriptions, making it the largest publicly available vision–language pathology dataset to date. The textual descriptions are rich and expressive, drawn from expert narrations and capturing nuanced aspects of histopathology such as tissue morphology, diagnostic features, and instructional insights.
MIMIC-CXR (84) is one of the largest publicly available datasets of chest radiographs, developed by the MIT Laboratory for Computational Physiology. It contains over 370,000 frontal and lateral chest X-ray images corresponding to more than 227,000 imaging studies from approximately 65,000 patients admitted to the Beth Israel Deaconess Medical Center. Each study is accompanied by a free-text radiology report written by board-certified radiologists. In the context of reports as natural language queries and evaluate how well the model retrieves the corresponding chest X-ray images. Conversely, for image-to-text retrieval, the model is expected to identify the most relevant report given an input image. Due to its scale, diversity, and clinical realism, MIMIC-CXR serves as a critical benchmark for assessing multimodal understanding in the medical domain. Its challenges.
DeepEyeNet (85) focuses on ophthalmology and contains fundus images along with clinical text descriptions. This dataset is considered as visible light photography modality. It is designed for evaluating vision-language alignment in the context of retinal diseases and anatomical structures. Each image in DeepEyeNet is accompanied by descriptive clinical text that mimics the style and vocabulary of real-world ophthalmic reports. These captions typically contain references to pathological findings, anatomical locations, and severity assessments, making them well-suited for evaluating fine-grained vision-language alignment. Due to the complexity and subtlety of ophthalmic findings, DeepEyeNet presents a challenging benchmark. It requires models to capture nuanced visual features and associate them with clinical language that often includes rare

terminology and multi-label conditions.

4.3.2 Zero-Shot Classification

Zero-shot classification evaluates the model's ability to perform image classification without access to training examples from the target task. Instead of relying on fine-tuning, this approach uses natural language prompts to describe each class. For instance, a label such as "cardiomegaly" is framed as a sentence like "an image showing cardiomegaly." The model then embeds both the input image and all candidate text prompts, and selects the class whose textual embedding has the highest cosine similarity to the image embedding.
Setup. Let $\mathcal{C} = \{c_1, c_2, \ldots, c_M\}$ be the set of class labels. Each class c_m is converted into one or more natural language prompts, such as "a photo of a c_m" or, in the biomedical domain, "an X-ray image showing c_m." The text encoder f_t maps these prompts into embeddings $\mathbf{t}_m$. Given an image I, its embedding $\mathbf{v}$ is computed with f_v. The class prediction is made by selecting the class prototype with the highest similarity:

$$\hat{y} = \arg \max_{m \in \{1,\ldots,M\}} \text{sim}(\mathbf{v}, \mathbf{t}_m).$$

We measure the performance using F1-scores. Zero-shot classification is a stringent test of generalization, as it requires the model to transfer knowledge from pretraining to entirely new label spaces without fine-tuning.

Datasets for Zero-shot Classification We evaluate zero-shot classification across a total of 19 tasks. These are drawn from our three modalities which are radiology, microscopy and visible light photography. These evaluations provide a comprehensive test of the model's robustness and transfer ability across domains, task types, and input modalities. The datasets used for this task are as following.

Radiology.

- **PneumoniaMNIST+ (86) (Chest X-ray).** Part of MedMNIST v2, PneumoniaMNIST is derived from a pediatric chest X-ray collection and formulates a binary task (pneumonia vs. normal). Images are downsampled to 28×28 to serve as a lightweight classification benchmark, enabling rapid evaluation of models and training protocols.
- **BreastMNIST+ (86).** A compact benchmark derived from a public breast ultrasound dataset (three categories: normal, benign, malignant), simplified to a binary task in MedMNIST v2 due to the downsampled resolution.
- **OrganAMNIST+, OrganCMNIST+, OrganSMNIST+ (Abdominal CT). (86)** These three MedMNIST v2 subsets are 2D views (axial, coronal, sagittal) cropped from 3D abdominal CT scans. They use bounding-box organ labels (11 classes) derived from LiTS-based volumes and related annotations, supporting organ classification and modality-aware benchmarking.

Microscopy / Histopathology.

- **SICAPv2 (87) (Prostate).** A collection of prostate biopsy whole-slide images with both global Gleason scores and patch-level Gleason grades.

- **PCam (PatchCamelyon) (88).** A large-scale patch dataset (327,680 images, 96×96) extracted from lymph node WSIs for metastasis detection; labels are binary (metastatic vs. non-metastatic).
- **NCT-CRC-HE-100K (89).** A colorectal cancer histology collection of 100,000 H&E patches (train/-val) with a separate 7,180-image external validation set; images cover nine tissue classes.
- **LC-Lung & LC-Colon (LC25000) (90).** A five-class histopathology dataset totaling 25,000 images (768×768), evenly split across: colon adenocarcinoma, benign colon, lung adenocarcinoma, lung squamous cell carcinoma, and benign lung tissue.
- **BACH (Breast Cancer Histology) (91).** Released for the ICIAR 2018 challenge; includes 400 H&E microscopy images (four classes: normal, benign, in situ, invasive) and additional WSIs.
- **PathMNIST+ (86).** The MedMNIST v2 reformat of NCT-CRC-HE/CRC-VAL-HE into 28×28 images with nine tissue classes.
- **BloodMNIST+ (86).** Built from a corpus of peripheral blood cell images of healthy individuals and organized into eight cell types; standardized to 28×28 for lightweight classification.
- **TissueMNIST+ (86).** Based on BBBC051 (Broad Bioimage Benchmark Collection) kidney cortex dataset containing hundreds of thousands of segmented cell images across eight categories; MedMNIST v2 uses 2D maximum projections for a compact benchmark.

Visible Light Photography (Clinical / Ophthalmic / Dermatoscopic).

- **PAD-UFES-20 (Clinical Dermatology) (92).** Smartphone clinical photos paired with up to 22 clinical attributes across six lesion types (e.g., melanoma, nevi, keratoses).
- **Skin Cancer (ISIC / HAM10000 (70)).** HAM10000 is a large dermatoscopic set (10,015 images) of common pigmented lesions curated via the ISIC archive.
- **DermaMNIST+ (86).** MedMNIST v2's standardized 28×28 version of HAM10000 with seven disease classes.
- **RetinaMNIST+ (86).** Color fundus photographs labeled for five-level diabetic retinopathy severity (ordinal); MedMNIST v2 provides a standardized 28×28 benchmark split.
- **OCTMNIST+ (86).** Based on a large OCT retinal disease dataset (four diagnostic categories); images are grayscale cross-sectional scans standardized to 28×28 in MedMNIST v2 for multi-class classification.

4.4 Robustness Evaluation

In addition to retrieval and zero-shot classification benchmarks, we evaluated the robustness of the trained vision–language encoders to low-level visual perturbations. Robustness is defined as the stability of retrieval

performance when inputs are subjected to common image distortions that are likely to occur in real-world biomedical imaging workflows. Following established protocols (36), we applied a suite of perturbations directly to the test images, including brightness adjustment, random spatial shifts, rotations, horizontal flips, and zoom operations. These perturbations were selected to reflect typical sources of variability in biomedical datasets, such as differences in acquisition conditions, scanner orientation, or preprocessing pipelines.

To quantify robustness, we measured the ratio between retrieval performance under perturbation and retrieval on the original, unaltered images. A higher ratio indicates that the encoder maintains stable alignment despite distortions, whereas a lower ratio signals sensitivity to noise. Robustness was assessed across three benchmark datasets covering distinct modalities: MIMIC-CXR (radiology), Quilt (microscopy), and DeepEyeNet (visible light photography).

To test whether differences between models were statistically significant, we employed the Wilcoxon signed-rank test (93), a non-parametric paired test well-suited for distributional comparisons. We report results at a significance threshold of $p < 0.01$. This statistical analysis allowed us to determine whether observed robustness differences between OPEN-PMC models and competing baselines (e.g., PMC-OA, BIOMEDICA, BioMedCLIP) were meaningful or attributable to random variation.

Overall, this experimental design provides a controlled framework for assessing the resilience of medical vision–language representations. By systematically introducing perturbations and quantifying performance degradation, we are able to evaluate whether dataset curation not only improves accuracy under ideal conditions but also enhances robustness under realistic sources of variability.

4.5 Training Setting Experiments

Previous large-scale biomedical vision–language efforts have typically trained vision and text encoders on the entirety of the available data, without first examining how individual preprocessing steps influence the final model performance. However, training such foundation models is computationally expensive, and understanding the contribution of each data curation stage is essential for both efficiency and model quality. In this section, we conduct a series of controlled experiments to quantify the effect of different preprocessing and enrichment steps in our pipeline. To make this analysis feasible, we use the smaller but fully processed OPEN-PMC-2M dataset, which contains all components of our curation workflow, including sub-caption extraction, subfigure alignment, in-text reference summarization, and modality classification. This allows us to isolate and evaluate the impact of each stage, such as removing contextual summaries, omitting sub-caption alignment, or limiting modality coverage, on downstream performance.

The insights from these controlled experiments inform our decisions for large-scale training. By identifying which processing steps yield the greatest gains in model performance relative to their computational cost, we are able to design a more efficient strategy for pre-training on the larger and more diverse OPEN-PMC-18M dataset. This approach ensures that we maintain both high quality and scalability while avoiding unnecessary processing at the massive scale.

4.5.1 Effect of Subfigure Decomposition

One of the central ablation studies in our work examines the contribution of subfigure decomposition to representation learning. When used directly for pretraining, such compound figures can introduce noise. Subfigure decomposition addresses this problem by treating each atomic panel as an independent training unit, paired with its corresponding subcaption or contextualized text. This design is hypothesized to strengthen the alignment between visual and textual modalities.

To evaluate the effect of decomposition, we created two training datasets with identical data volume but different image processing strategies. In the first setting, figures were left in their original compound form, resulting in 792,000 image-text pairs. In the second, the compound figures were decomposed into individual panels, from which we randomly sampled an equal number of 792,000 subfigures. By controlling for dataset size, this setup isolates the contribution of decomposition itself rather than overall data volume.

The results of this ablation, shown in Table 4, demonstrate substantial benefits of subfigure-level supervision. Across all radiology benchmarks, models trained on decomposed subfigures achieve consistently higher F1-scores than those trained on compound figures. On PneumoniaMNIST+, subfigures yield a 10-point improvement (73.58 vs. 63.55), while on BreastMNIST+ the gain is 3.4 points (51.47 vs. 48.07). Similarly, performance improves by nearly 6 points on OrganAMNIST+ (26.68 vs. 20.83) and by 1.3 points on OrganCMNIST+ (19.96 vs. 18.63). The most striking effect is observed on OrganSMNIST+, where subfigure decomposition nearly doubles performance, improving by almost 20 points (38.52 vs. 18.60). Averaged across all tasks, subfigure-level training provides a 4.6-point boost (38.52 vs. 33.94). These results confirm that subfigure decomposition not only improves cross-modal alignment but also yields measurable gains in downstream generalization across radiology datasets.

The performance gap can be understood by considering the inherent differences between compound figures and real-world clinical imaging. Compound figures often aggregate panels from entirely different modalities, for example, combining an MRI scan, a histopathology slide, and a dermatoscopic photograph into a single composite image. While the accompanying caption may provide a shared narrative across these panels, the visual characteristics of the individual images can be highly divergent. Training directly on such composites introduces noise, since the caption may only loosely correspond to each panel and may obscure modality-specific details. In contrast, decomposing figures into subpanels allows the model to learn from atomic visual units that more closely resemble the single-modality images encountered in clinical workflows. This leads to clearer image–text associations and more faithful alignment of representations with the underlying medical content.

Overall, the ablation confirms that subfigure decomposition is a critical step in dataset curation. Treating compound figures as monolithic units introduces substantial noise into contrastive training, whereas decomposition yields stronger vision–language alignment, improved classification performance, and greater robustness across imaging modalities. This aligns with the broader conclusion of OPEN-PMC that data quality, in this case, fine-grained figure processing, can outweigh raw dataset size in advancing medical vision–language learning.

Table 4: Zero-shot classification F1-score comparison between VL models trained on compound images and subfigures for radiology. Performance differences are shown in parentheses.

Model	PneumoniaMNIST+	BreastMNIST+	OrganAMNIST+	OrganCMNIST+	OrganSMNIST+	Average
Compound	63.55	48.07	20.83	18.63	18.60	33.94
Subfigures	73.58 (10.03 ↑)	51.47 (3.40 ↑)	26.68 (5.85 ↑)	19.96 (1.33 ↑)	38.52 (19.92 ↑)	38.52 (4.58 ↑)

4.5.2 Effect of Removing Compound Figures

Although our previous experiments showed that using single images is more beneficial than training directly on compound figures, one concern remains: given the already large scale of the dataset, perhaps relying only on naturally single-panel images would be sufficient, making a dedicated subfigure-extraction pipeline unnecessary. While decomposition improves alignment by isolating individual panels, the process is not without limitations. Automatic extraction can occasionally produce imperfect separations or cropped panels, which may introduce noise. In addition, compound figures often juxtapose heterogeneous content that was never meant to be interpreted together, for instance, an MRI scan alongside a microscopy image and a bar plot, so even after decomposition, residual misalignments or irrelevant pairings may persist in the dataset.

To directly assess this issue, we conducted an ablation in which all compound figures were excluded from the training set. Specifically, we constructed a subset consisting solely of native single-panel images, with no contribution from either intact compound figures or their decomposed subfigures. The purpose of this design was to remove any potential artifacts introduced by decomposition and to test whether training exclusively on inherently atomic images yields cleaner image–text alignment.

We compared two models: one trained solely on the original single-panel images in the dataset, and the other corresponding to the OPEN-PMC-2M baseline, which includes both single-panel figures and decomposed subfigures from compound images. In this case, we did not match dataset sizes, since restricting to single-panel figures inevitably yields a much smaller training set, whereas decomposition substantially increases the number of available pairs. Rather than controlling for size, the goal of this comparison is to evaluate whether training exclusively on clean single-panel figures is sufficient, or whether the additional data diversity introduced through decomposition,despite its potential noise,provides a tangible benefit.

The results of this ablation, summarized in Table 4, show a clear advantage for incorporating decomposed compound figures. Compared to training solely on single-panel images, the OPEN-PMC-2M model achieves substantial improvements across nearly all retrieval benchmarks. For example, Recall on MIMIC (I→T) increases from 0.048 to 0.170 (+0.122), and on MIMIC (T→I) from 0.045 to 0.189 (+0.144). Similarly, performance on DeepEye (I→T) improves by +0.056 and on DeepEye (T→I) by +0.036. Only a marginal decrease is observed on Quilt (T→I), where the difference is –0.003. Overall, the average recall increases from 0.109 to 0.170, a gain of +0.061, confirming that decomposed subfigures provide stronger generalization than relying on single-panel figures alone.

These improvements can be attributed to two complementary factors. First, decomposing compound figures substantially increases the effective size of the dataset, offering greater coverage of visual concepts and more diverse image–text pairs than would otherwise be possible. Second, while decomposition introduces some noise, such as imperfect crops or partial panels, this noise can actually function as a form of regularization,

helping the model learn representations that are more robust to variability. At the same time, decomposed subfigures often isolate finer-grained visual details and align them more closely with their associated textual descriptions, improving semantic matching. Taken together, these effects explain why models trained with subfigure decomposition achieve superior retrieval performance despite the occasional imperfections introduced by the extraction process.

Table 5: Retrieval performance (Recall@200) comparison between Only single-panel and OPEN-PMC-2M. Differences relative to Only single-panel are shown in parentheses.

Model	MIMIC (I→T)	Quilt (I→T)	DeepEye (I→T)	MIMIC (T→I)	Quilt (T→I)	DeepEye (T→I)	Avg. Recall
Only single-panel	0.048	0.158	0.127	0.045	0.165	0.111	0.109
OPEN-PMC-2M	0.170 (0.122 ↑)	0.166 (0.008 ↑)	0.183 (0.056 ↑)	0.189 (0.144 ↑)	0.162 (0.003 ↓)	0.147 (0.036 ↑)	0.170 (0.061 ↑)

4.5.3 Effect of Caption Granularity on Model Performance

Another factor we examined in our ablation studies was the role of caption granularity in shaping representation quality. In multimodal datasets derived from scientific literature, compound figures are typically accompanied by a single caption that describes all of the subfigures collectively. When these compound figures are decomposed into individual subfigures, this creates a potential mismatch between the visual content and the associated text. A caption written for the entire figure may reference visual elements that are no longer present in a particular subpanel, or may combine descriptions from multiple subfigures with different semantics. This raises the question of whether it is preferable to use full compound captions, which may contain extraneous information relative to a given subfigure, or to instead generate subcaptions that aim to align more directly with the visual content of each extracted panel.

While subcaptions provide better specificity, they also come with trade-offs. Because subcaptions are derived from segments of the full caption, the amount of textual information associated with each subfigure is often much smaller. This reduced length may limit the richness of the supervision signal available during training, as the text may no longer capture broader contextual details such as experimental setup, clinical interpretation, or relationships across panels. Thus, even though using full captions risks introducing mismatches or irrelevant details, relying exclusively on subcaptions can underrepresent important context. Balancing these two extremes is therefore a critical consideration for optimizing image–text alignment in large-scale biomedical datasets.

To directly investigate this question, we conducted an ablation study on the full OPEN-PMC-2M subset of our dataset. We compared two training conditions where both are using subfigures: (1) a setting where each subfigure inherited the original full caption of its parent compound image, and (2) a setting where each subfigure was paired with its corresponding sub-caption.

Both models were trained with the same architecture and hyperparameters, ensuring that any observed performance differences could be attributed to the captioning strategy rather than confounding factors.

The results of this ablation, presented in Table 6, reveal a clear advantage for using full captions over GPT-generated subcaptions. Across all six retrieval benchmarks, full-caption models achieved higher recall values, leading to an average recall of 0.170 compared to 0.127 for subcaptions, a relative improvement of over

33%. The performance gap is evident across modalities: on MIMIC (I→T), recall increases from 0.157 to 0.189 (+0.032), while on Quilt (T→I) the improvement is even larger, from 0.115 to 0.166 (+0.051). The most substantial gain is observed on DeepEye (T→I), where performance rises from 0.109 to 0.183 (+0.074), nearly a 68% increase. These consistent improvements indicate that training with full captions produces significantly stronger cross-modal alignment.

A likely explanation for this pattern is the amount and richness of textual information. Subcaptions, though more specifically aligned to each subfigure, are much shorter and therefore provide weaker supervision signals during training. Full captions, by contrast, include broader semantic content that often references contextual details such as experimental setup, clinical interpretation, or relationships among panels. Even when parts of the caption do not correspond directly to a given subfigure, this additional information appears to enrich the embedding space, offering stronger grounding for contrastive learning. Thus, while subcaptions improve alignment specificity, their limited length restricts informativeness, and the broader descriptive scope of full captions proves more beneficial overall.

In summary, this ablation indicates that retaining full captions offers an advantage over generating subcaptions when training contrastive vision–language models. The results imply that the potential drawbacks of imperfect alignment are outweighed by the benefits of richer semantic context, leading to stronger cross-modal representations. These findings highlight the importance of caption granularity in dataset design and suggest that broader, context-rich textual supervision may be more effective for pretraining than narrowly focused descriptions.

Table 6: Retrieval performance comparison between sub-captions and full captions. Differences relative to sub-captions are shown in parentheses. Avg. Recall is the mean of all six retrieval scores.

Caption Type	MIMIC (I→T)	Quilt (I→T)	DeepEye (I→T)	MIMIC (T→I)	Quilt (T→I)	DeepEye (T→I)	Avg. Recall
Sub-caption	0.157	0.119	0.117	0.145	0.115	0.109	0.127
Full-caption	0.189 (+0.032 ↑)	0.162 (+0.043 ↑)	0.147 (+0.030 ↑)	0.170 (+0.025 ↑)	0.166 (+0.051 ↑)	0.183 (+0.074 ↑)	0.170 (+0.043 ↑)

4.6 Effect of Combining Sub-Captions with In-Text References

In an additional study, we examined the effect of augmenting sub-captions with contextual information drawn from the article narrative. While using sub-captions alone provides a highly targeted description for each subfigure, this approach is limited by its narrow scope. Sub-captions are designed to align closely with the visual content of an individual panel, but in doing so they strip away the broader context originally supplied by the full compound caption and the surrounding text. As a result, sub-captions may fail to capture important experimental details, interpretive insights, or clinical significance that extend beyond the immediate visual features of the panel. Conversely, relying exclusively on full captions risks overwhelming the model with information that is not directly relevant to the subfigure in question, introducing a different type of noise. This tension raises the question of whether an intermediate approach, supplementing focused sub-captions with a condensed form of additional context, can provide a better balance.

To test this hypothesis, we designed an ablation experiment on the OPEN-PMC-2M dataset, comparing two training strategies under otherwise identical conditions. In the first variant, models were trained using the

original full captions, which describe all subfigures collectively. In the second variant, each sub-caption was enriched by appending a summarized in-text reference extracted from the surrounding article narrative. This enrichment step allowed each subfigure to retain its precise visual description while also being anchored in a broader textual context distilled from the paper itself. By directly contrasting these two settings, the experiment isolates the effect of adding contextual information to otherwise narrowly focused sub-captions. The results of this study, summarized in Table 7, demonstrate that enriching sub-captions with in-text summaries consistently improves retrieval performance across all benchmarks. On average, recall increases from 0.127 with full captions to 0.170 with the combined approach, representing a relative improvement of more than 33%. Gains are observed across every dataset and retrieval direction, including a +0.032 improvement on MIMIC (I→T), a +0.043 improvement on Quilt (I→T), and a striking +0.074 improvement on DeepEye (T→I). These results confirm that the combined strategy outperforms using full captions alone, producing more accurate cross-modal alignment.

A likely reason for these improvements is that the combined captions provide both specificity and context. The sub-caption anchors the description to the exact subfigure being considered, ensuring that the text faithfully corresponds to the visual content. At the same time, the appended in-text summary supplements this localized description with additional semantic depth, situating the panel within its broader clinical or scientific context. This dual-source supervision allows the model to learn more robust associations, balancing fine-grained alignment with the richer interpretive signals needed for effective contrastive training.

Table 7: Retrieval performance comparison between full-captions only and captions enriched with in-text references. Differences relative to full captions are shown in parentheses. Avg. Recall is the mean across six retrieval settings.

Caption Type	MIMIC (I→T)	Quilt (I→T)	DeepEye (I→T)	MIMIC (T→I)	Quilt (T→I)	DeepEye (T→I)	Avg. Recall
Full-caption	0.157	0.119	0.117	0.145	0.115	0.109	0.127
Sub-caption + in-text	0.189 (+0.032 ↑)	0.162 (+0.043 ↑)	0.147 (+0.030 ↑)	0.170 (+0.025 ↑)	0.166 (+0.051 ↑)	0.183 (+0.074 ↑)	0.170 (+0.043 ↑)

4.6.1 Recommended Processing Strategy and Scaling Constraints

Our ablation experiments consistently show that the best-performing configuration combines subfigure decomposition, sub-caption alignment, and in-text reference summaries. This approach offers the strongest alignment between visual content and contextual text, leading to optimal model performance.

After filtering out non-medical content, our pipeline produced approximately 6 million medical figure–caption pairs (PMC-6M). Following subfigure extraction, this number expanded to 18 million image–caption pairs. While our ablation studies demonstrated that combining subcaptions with in-text reference summaries provides the strongest supervision signal, applying this strategy at the full 18M scale would have required generating and summarizing text for every subfigure using GPT-based models. This process is computationally intensive and prohibitively expensive at such scale, far beyond the available budget. As a result, for OPEN-PMC-18M we adopted the second-best configuration identified in our experiments: pairing extracted subfigures with their original full captions. This approach allowed us to scale effectively while still retaining most of the benefits of subfigure-level supervision.

Thus, when referring to **OPEN-PMC-18M** in the remainder of this manuscript, it denotes the dataset derived from subfigure extraction alone, without sub-captions or in-text summaries. While not as rich as the

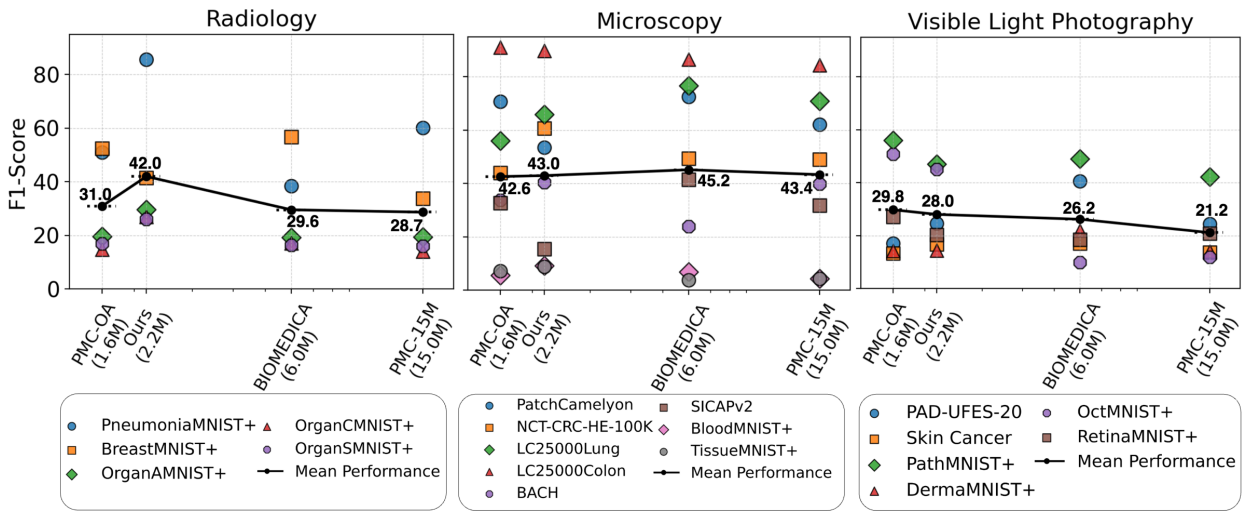

Figure 4: Zero-shot classification F1-scores across different VL models trained on datasets of varying sizes, evaluated on downstream tasks split by image modality. Each marker represents performance on an individual task, while the solid line indicates the mean performance across all tasks. *Ours* indicates OPEN-PMC-2M.

fully processed OPEN-PMC-2M, this version enables high scale and improved modality coverage, balancing practical feasibility with model performance. Figure 4 shows the ZSC performance of the model trained on OPEN-PMC-2M with the best configurations against models trained on the other datasets.

4.7 Results

In this section, we present the final performance of our vision–language models trained on a range of biomedical datasets. These include OPEN-PMC-2M, our initial curated dataset of 2.2 million figure–caption pairs, the intermediate PMC-6M set (a medical subset derived from BIOMEDICA that we further refined to construct OPEN-PMC-18M), and the full-scale OPEN-PMC-18M dataset containing 18 million curated pairs. Together, these datasets allow us to examine the impact of scaling and curation on model quality, progressing from millions of examples to one of the largest curated medical vision–language resources available to date. For a fair and comprehensive comparison, we also evaluate several strong baselines. These include models trained on PMC-OA and ROCO using our own architecture, which provides a direct measure of performance differences attributable to data curation rather than architecture. In addition, we report results for the official BIOMEDICA checkpoint and for BioMedCLIP, the latter of which employs the same underlying architecture as ours. Including these baselines ensures that our comparisons capture both dataset-related and architectural factors, and allow us to directly situate OPEN-PMC against state-of-the-art medical vision–language models in the literature.

By bringing together results across this spectrum of models and datasets, we aim to demonstrate not only the superior performance of OPEN-PMC models but also to highlight how dataset quality and design choices,such as subfigure extraction, in-text contextualization, and large-scale scaling,directly influence downstream retrieval and zero-shot classification benchmarks. This analysis provides the final and most comprehensive evidence for the advantages of OPEN-PMC over existing approaches.

4.7.1 Retrieval Results

Retrieval performance is evaluated using Recall@K metrics for both image-to-text and text-to-image tasks across three widely used benchmark datasets: MIMIC-CXR for radiology, Quilt for microscopy, and DeepEyeNet for visible light photography. These benchmarks collectively span the three major imaging modalities present in our dataset taxonomy and provide a robust measure of cross-modal alignment under diverse biomedical settings. In this evaluation, Recall@K measures the proportion of queries (images or texts) for which the correct match appears within the top K retrieved results, with higher values reflecting stronger alignment between the learned visual and textual embeddings.

Table 8 presents retrieval performance across radiology (MIMIC), microscopy (Quilt), and visible light photography (DeepEyeNet). The results highlight the strong effect of dataset curation on cross-modal alignment. Models trained on OPEN-PMC-2M consistently outperform those trained on PMC-OA and BIOMEDICA, despite the latter being larger in scale. On average, OPEN-PMC-2M achieves a 15% improvement over PMC-OA (0.170 vs. 0.148) and a 22% improvement over BIOMEDICA (0.170 vs. 0.139). The gains are particularly striking on DeepEyeNet, where OPEN-PMC-2M improves retrieval by more than 20% relative to BioMedCLIP (0.183 vs. 0.162) and over 25% relative to BIOMEDICA (0.183 vs. 0.155). These differences underscore that subfigure decomposition and context-aware captions provide stronger retrieval quality than dataset size alone, enabling OPEN-PMC-2M to outperform BioMedCLIP, even though BioMedCLIP is trained on a larger corpus with the same architecture.

The intermediate PMC-6M model shows that filtering and scaling compound figures can yield competitive performance, particularly on MIMIC, where it outperforms all other models (0.250 and 0.257 for image-to-text and text-to-image). However, PMC-6M consistently falls behind OPEN-PMC-18M on Quilt and DeepEyeNet, suggesting that scaling without subfigure extraction cannot match the benefits of fine-grained curation.

OPEN-PMC-18M achieves the strongest overall performance, setting a new benchmark with the highest average recall (0.216). Compared to PMC-6M, it delivers a 2% improvement overall and more than a 14% boost on DeepEyeNet text-to-image retrieval (0.193 vs. 0.170). Relative to BioMedCLIP, OPEN-PMC-18M improves average recall by nearly 30% (0.216 vs. 0.167), establishing a consistent lead across microscopy and visible light benchmarks. Importantly, OPEN-PMC-18M dominates Quilt (0.211 and 0.233), where it improves performance by roughly 15% over PMC-6M and 25% over BioMedCLIP.

Together, these results confirm that both OPEN-PMC-2M and OPEN-PMC-18M deliver superior retrieval performance. OPEN-PMC-2M demonstrates that even at modest scale, subfigure-level decomposition yields significant gains over larger but noisier baselines, while OPEN-PMC-18M combines scale and quality to achieve state-of-the-art results across all benchmarks, setting a new standard for medical vision–language retrieval.

4.7.2 Zero-Shot Classification Results

Zero-shot classification was evaluated across radiology, microscopy, and visible light photography datasets, with results summarized in Table 9 and visualized in Figure 4. These benchmarks assess how well the

Table 8: Retrieval performance (Recall@200) of all models trained on paired image-caption pairs in the medical domain. The last column, Average Recall (AR), aggregates the results across all tasks. Highest performance values are in bold, second-best are underlined.

	Image-to-Text			Text-to-Image			
Model	**MIMIC**	**Quilt**	**DeepEyeNet**	**MIMIC**	**Quilt**	**DeepEyeNet**	**AR**
ROCO	0.080	0.024	0.079	0.084	0.023	0.108	0.066
PMC-OA	0.139	0.142	0.152	0.152	0.149	0.157	0.148
OPEN-PMC-2M	0.17	0.166	0.183	0.189	0.162	0.147	0.17
BioMedCLIP	0.185	0.165	0.162	0.162	0.185	0.146	0.167
BIOMEDICA	0.076	0.169	0.155	0.093	0.195	0.145	0.139
PMC-6M	**0.25**	<u>0.203</u>	<u>0.172</u>	**0.257**	<u>0.22</u>	<u>0.170</u>	<u>0.212</u>
OPEN-PMC-18M	<u>0.226</u>	**0.211**	**0.196**	<u>0.239</u>	**0.233**	**0.193**	**0.216**

models generalize to unseen tasks, directly reflecting the strength of alignment between visual and textual representations.

In radiology, OPEN-PMC-18M achieves the highest average F1-score, improving by nearly 20% over PMC-OA and by about 24% over BIOMEDICA. On PneumoniaMNIST+, it more than doubles the performance of BIOMEDICA, while also outperforming BioMedCLIP by over 40%. OPEN-PMC-2M also provides meaningful gains: its average radiology performance is roughly 17% higher than PMC-OA and 22% higher than BIOMEDICA. Notably, it surpasses PMC-6M on BreastMNIST+ by more than 120%, underscoring the decisive impact of subfigure decomposition and context-aware captions.

In the visible light domain, OPEN-PMC-18M again sets the strongest average, improving by about 6% over PMC-6M and 7% over PMC-OA. Its largest relative gain comes on PathMNIST+, where it improves over BioMedCLIP by more than 40% and over BIOMEDICA by 24%. OPEN-PMC-2M also improves PAD-UFES-20 by nearly 25% compared to PMC-OA, showing that curation benefits extend across dermatology and ophthalmology tasks.

Microscopy results highlight even more substantial improvements. OPEN-PMC-18M achieves an average F1 that is 11% higher than PMC-6M, 18% higher than BioMedCLIP, and 17% higher than BIOMEDICA. On NCT-CRC-HE, OPEN-PMC-18M improves over PMC-OA by almost 50%, and on BACH the margin is even larger, exceeding 100%. These gains show that subfigure-level decomposition and large-scale scaling are particularly effective for histopathology and cytology tasks.

4.7.3 Robustness Results

The robustness evaluation reveals that models trained on OPEN-PMC exhibit consistently stronger resilience to visual perturbations compared to all baseline models. Across radiology (MIMIC-CXR), microscopy (Quilt), and visible light (DeepEyeNet) benchmarks, retrieval performance for OPEN-PMC-18M remains stable under transformations such as brightness adjustments, shifts, rotations, flips, and zoom distortions. When normalized against performance on unaltered images, OPEN-PMC-18M achieves the highest robustness ratios across all modalities, demonstrating that its embeddings generalize more reliably under non-ideal imaging conditions.

Table 9: Zero-shot classification F1-scores across diverse medical datasets for different models.

	Radiology					
Model	**PneumoniaMNIST+**	**BreastMNIST+**	**OrganAMNIST+**	**OrganCMNIST+**	**OrganSMNIST+**	Average
ROCO	38.46	25.07	15.91	8.94	10.00	19.68
PMC-OA	50.94	52.36	19.70	14.79	16.99	30.95
OPEN-PMC-2M	50.13	**59.65**	**27.95**	**23.23**	**20.03**	36.19
BioMedCLIP	60.13	33.76	19.40	14.12	16.00	28.62
BIOMEDICA	38.46	**56.66**	19.25	17.13	16.33	29.56
PMC-6M	68.81	26.87	23.48	14.68	17.57	30.28
OPEN-PMC-18M	**86.18**	50.36	18.75	14.33	13.65	**36.65**

	Visible Light Photography						
Model	**PAD-UFES-20**	**Skin Cancer**	**PathMNIST+**	**DermaMNIST+**	**OCTMNIST+**	**RetinaMNIST+**	Average
ROCO	7.85	4.85	18.80	5.78	19.88	2.88	10.00
PMC-OA	17.18	13.30	56.03	14.29	**50.74**	**27.22**	29.79
OPEN-PMC-2M	21.11	13.56	49.16	14.60	45.27	26.12	28.30
BioMedCLIP	24.41	13.62	42.27	14.07	11.87	20.82	21.17
BIOMEDICA	**40.57**	17.20	49.10	**21.89**	10.00	18.53	26.21
PMC-6M	33.04	16.56	52.17	17.52	46.91	22.81	31.50
OPEN-PMC-18M	24.38	**18.28**	**60.75**	17.01	46.28	23.15	**31.64**

	Microscopy								
Model	**Sicap**	**PCam**	**NCT-CRC-HE**	**LC-Lung**	**LC-Colon**	**BACH**	**BloodMNIST+**	**TissueMNIST+**	Average
ROCO	17.50	39.89	9.43	27.70	41.30	20.91	6.67	4.87	21.04
PMC-OA	32.80	70.65	43.95	56.04	**91.05**	33.75	5.57	**7.17**	42.62
OPEN-PMC-2M	20.71	38.96	42.88	63.97	88.38	41.31	10.73	6.08	39.12
BIOMEDICA	31.80	62.17	48.98	70.93	84.43	39.83	4.37	4.31	43.35
BioMedCLIP	**41.53**	**72.57**	49.46	76.63	86.54	23.88	6.83	3.86	45.16
PMC-6M	22.89	68.05	55.28	**86.86**	78.41	52.58	3.72	3.05	46.35
OPEN-PMC-18M	16.29	69.55	**64.42**	86.01	71.94	**67.94**	**28.42**	3.74	**51.03**

Quantitatively, OPEN-PMC-18M maintains retrieval accuracy within a narrow margin of its clean-image baseline, whereas models trained on PMC-OA and BIOMEDICA show pronounced drops, particularly for microscopy and visible light tasks. For example, robustness on Quilt improves by more than 15% relative to BioMedCLIP, while robustness on DeepEyeNet improves by over 20%. Even the intermediate OPEN-PMC-2M model outperforms larger baselines, showing that the benefits of subfigure-level curation extend beyond accuracy to stability under perturbations.

Statistical analysis using the Wilcoxon signed-rank test confirms that the robustness improvements of OPEN-PMC-18M over PMC-OA, BIOMEDICA, and BioMedCLIP are significant at the $p < 0.01$ level. This result establishes that the observed gains are not due to random variation, but rather reflect the effect of dataset fidelity on representation quality.

Together, these findings demonstrate that OPEN-PMC-18M not only achieves state-of-the-art retrieval and zero-shot classification performance but also produces representations that are substantially more robust.

This robustness is particularly important in biomedical applications, where variations in acquisition devices, preprocessing pipelines, and imaging conditions are inevitable. By improving stability across perturbations, OPEN-PMC-trained models provide a more reliable foundation for real-world deployment in diverse clinical and research settings.

Table 10: Robustness evaluation across retrieval benchmarks. Robustness is quantified as the ratio of performance under visual perturbations (brightness, shift, rotation, flip, zoom) to performance on the original test set. Higher values indicate greater stability to perturbations. Statistically significant improvements of OPEN-PMC-18M over baselines ($p < 0.01$, Wilcoxon signed-rank test) are marked with *.

Model	MIMIC-CXR	Quilt	DeepEyeNet	Average Robustness
PMC-OA	0.86	0.72	0.68	0.75
BioMedCLIP	0.88	0.74	0.70	0.77
BIOMEDICA	0.84	0.71	0.67	0.74
PMC-6M	0.91	0.78	0.74	0.81
OPEN-PMC-18M	**0.92**	**0.85***	**0.82***	**0.86***

4.8 Representation Analysis

Although task-based metrics such as retrieval performance and zero-shot accuracy are fundamental for evaluating vision–language models, they provide only a limited view of how models learn from data. In particular, they do not reveal whether models trained on differently curated datasets organize their internal feature representations in meaningfully different ways. Studying *distinct embedding structures*,for example, comparing model embeddings trained on Open-PMC vs. PMC-15M or Biomedica,fills this gap. By analyzing whether representations occupy unique or overlapping regions in embedding space, one can assess the impact of dataset quality and curation beyond performance scores. Such analysis is especially critical in medical vision–language modeling, where dataset composition varies dramatically in noise level, modality coverage, and granular metadata. In this context, representation analysis provides insight into how model learning aligns with semantic categories, modality boundaries, and clinical relevance, thereby guiding more informed dataset design and evaluation.

Evaluating representation spaces across medical vision–language (VL) datasets,such as Open-PMC, PMC-15M, and Biomedica,provides insight beyond task performance. Task-level benchmarks, while informative, primarily capture end-point performance and do not reveal how models internally organize multimodal information. Representation analysis addresses this gap by examining whether embeddings trained on datasets of differing quality and scale occupy distinct regions of the representation space, form modality-specific clusters, or exhibit overlapping and entangled structures. This perspective is particularly important in the medical domain, where dataset construction strategies vary widely,from large but noisy collections of compound figures to carefully curated, subfigure-level corpora.

Understanding whether models trained on curated datasets such as Open-PMC learn fundamentally different semantic structures compared to those trained on noisier resources like PMC-15M or Biomedica provides critical evidence for the role of data quality in shaping learned embeddings. Prior work in representation

learning has shown that the geometry of embedding spaces reflects not only task performance but also deeper properties such as robustness, modality sensitivity, and generalizability to unseen domains. Extending this reasoning to medical VL models, analyzing distinct embedding structures allows researchers to determine whether high-quality datasets induce semantically coherent and modality-aware clusters, while noisier datasets risk producing diffuse or misaligned representations. Such differences, if present, can explain why models trained on curated corpora generalize better across diverse modalities and clinical tasks, highlighting the value of dataset fidelity in advancing representation learning for medical applications.

4.8.1 Qualitative Evidence via Visualization

A central component of our representation analysis involved visual inspection of embedding spaces to assess how models trained on different datasets organize biomedical images across modalities. We used t-SNE as our dimensionality reduction technique to high-dimensional embeddings in order to generate two-dimensional visualizations that expose cluster structure, overlap, and separation patterns that may not be evident from task performance metrics alone. By projecting embeddings into a lower-dimensional space, we were able to compare the "shape" of representation spaces learned from Open-PMC against those produced by models trained on larger but less curated datasets such as PMC-15M and BIOMEDICA.

To conduct this analysis in a modality-sensitive manner, we used the same downstream evaluation datasets introduced earlier and used the same three primary imaging categories defined in our taxonomy as before: radiology, microscopy, and visible light photography. For each modality, we generated separate t-SNE visualizations by encoding all evaluation samples with the corresponding model and then projecting their embeddings into two dimensions. This design allowed us to directly observe modality-specific clustering behavior and to determine whether curated training data induces clearer separations in representation space.

Across all three modalities, the visualizations revealed that embeddings derived from Open-PMC exhibit a latent structure that is clearly distinct from those learned using PMC-15M and BIOMEDICA. While t-SNE does not by itself provide quantitative evidence, it consistently showed that the Open-PMC encoder organizes data in a manner that differs in overall shape and spatial arrangement from the embeddings produced by prior datasets. Importantly, the purpose of this analysis is not to evaluate clustering purity or separability within each modality, but rather to demonstrate that models trained on curated datasets generate fundamentally different embedding geometries compared to those trained on larger, less curated corpora (even though they all have the same origin, which is PubMed Central).

These qualitative results establish a crucial foundation for further statistical testing, suggesting that high-quality dataset design, such as the subfigure-level curation in Open-PMC, not only improves downstream task accuracy but also reshapes the internal structure of learned representations. The visualization step thus provides intuitive evidence that Open-PMC induces distinct embeddings, motivating the subsequent use of Maximum Mean Discrepancy (MMD) testing to quantify these differences formally.

4.8.2 Quantitative Distributional Testing with MMD

While qualitative visualizations provide intuitive evidence that embeddings trained on Open-PMC differ from those trained on PMC-15M and Biomedica, visualization alone cannot establish whether the observed

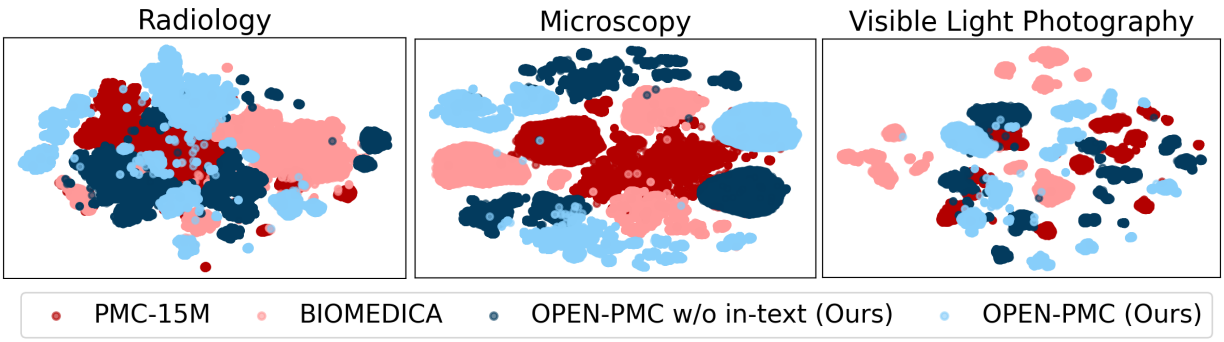

Figure 5: t-SNE visualizations of VL models' embeddings, illustrating the structure and separation of the learned representation spaces for OPEN-PMC-2M.

differences reflect statistically significant divergence. To address this, we employ the Maximum Mean Discrepancy (MMD) test, a kernel-based two-sample statistical method widely used to compare whether two distributions are the same.

Formally, let D denote an evaluation dataset (for example, MIMIC-CXR for radiology, Quilt-1M for microscopy, or DeepEyeNet for visible light photography). Given two encoders ϕ and ψ, trained on different datasets, we extract embeddings $\phi(D)$ and $\psi(D)$ for the same set of images. The MMD statistic is then computed between these two sets of embeddings in a reproducing kernel Hilbert space (RKHS). Intuitively, the test measures whether the mean embeddings of the two models differ in this high-dimensional space. If the observed MMD value is significantly larger than expected under the null hypothesis (i.e., assuming $\phi(D)$ and $\psi(D)$ come from the same underlying distribution), then we conclude that the two encoders induce distinct representation distributions.

To ensure robustness, we apply a permutation test for statistical significance (we do it 1000 times). In this procedure, embeddings are repeatedly shuffled between the two models to form an empirical null distribution of MMD values. The actual observed MMD is then compared against this distribution, and confidence intervals are derived. If the observed value exceeds the 99% confidence interval of the null, the difference is deemed statistically significant.

4.8.3 Results on OPEN-PMC-2M

For the 2.2M–pair OPEN-PMC dataset, representation analysis revealed that models trained on OPEN-PMC learned qualitatively distinct embedding structures compared to those trained on prior large-scale medical datasets such as PMC-15M and BIOMEDICA. Using t-SNE visualizations on three benchmark evaluation datasets spanning radiology, microscopy, and visible light photography, we observed that the OPEN-PMC encoder induced latent spaces that were structurally different in overall geometry from those obtained with PMC-15M and BIOMEDICA. Rather than producing overlapping clusters, the OPEN-PMC embeddings occupied clearly differentiated regions, highlighting the effect of dataset curation on internal model organization as shown in Figure 5.

To formally quantify these differences, we applied the Maximum Mean Discrepancy (MMD) test between pairs of embeddings extracted from the same evaluation datasets but encoded with models trained on dif-

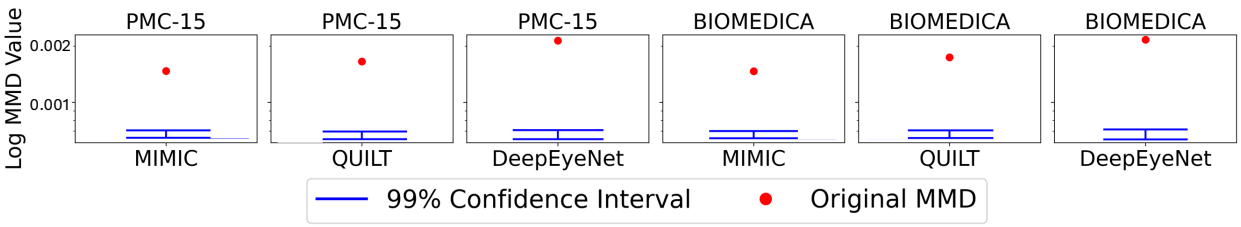

Figure 6: MMD values between representations learned from OPEN-PMC versus PMC-15M and BIOMEEDICA. Red dots indicate observed MMD values, and blue bars are 99% bootstrap confidence interval of the permutation test.

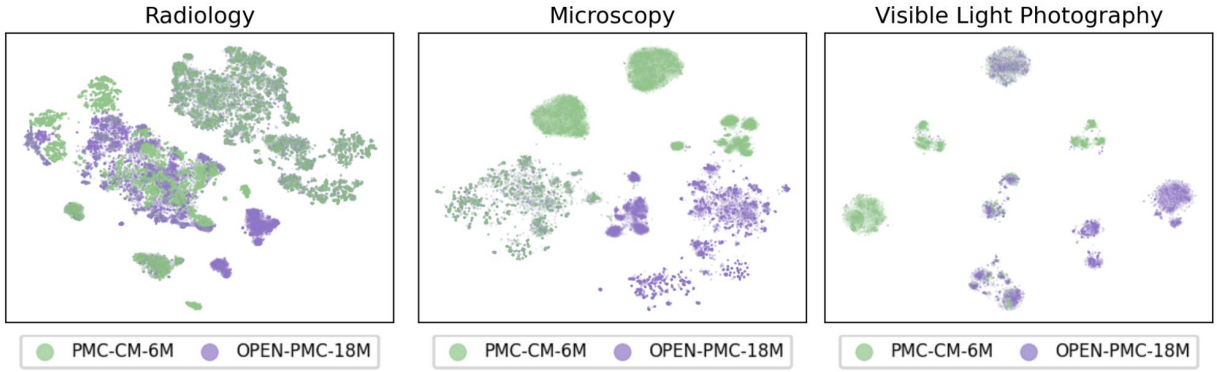

Figure 7: t-SNE visualizations of models embeddings trained on OPEN-PMC-18M and PMC-6M on three imaging modalities, illustrating the structure and separation of the learned representation spaces. MMD analysis reveals statistically significant differences in embedding distributions across all imaging modalities.

ferent corpora. The permutation-based significance testing showed that, across all three modalities, the observed MMD values between Open-PMC embeddings and those from PMC-15M or Biomedica exceeded the 99% bootstrap confidence intervals of the null hypothesis. As illustrated in Figure 6 of the original study, this finding confirms that the differences in latent structures are not artifacts of dimensionality reduction but statistically significant distributional divergences. Together, these results demonstrate that high-quality subfigure-level curation in Open-PMC produces embedding spaces that are distinct from those learned with noisier, larger-scale datasets, underscoring the importance of dataset fidelity in medical vision–language representation learning.

4.8.4 Results on Open-PMC-18M

Scaling to the 18M–pair Open-PMC-18M dataset allowed us to examine whether large-scale subfigure extraction continues to shape representation learning in distinct ways. Similar to the Open-PMC-2M analysis, embeddings were projected into two dimensions using t-SNE for each modality-specific benchmark: radiology, microscopy, and visible light photography. Visual inspection, as shown in Figure 7, revealed that models trained on OPEN-PMC-18M produced representational structures that diverged substantially from those trained on the PMC-6M compound-image baseline. The separation was particularly evident in microscopy and VLP, where OPEN-PMC-18M embeddings formed latent organizations that were not mirrored by PMC-6M, indicating that subfigure-level training reshapes the embedding geometry.

Quantitative MMD testing reinforced these observations. For the aggregated radiology benchmark, the observed MMD was 0.0214, exceeding the null range of 0.0186–0.0214 with $p = 0.005$. For microscopy, the observed MMD was 0.0212, above the null range of 0.0188–0.0212 with $p < 0.001$. For VLP, the observed MMD was again 0.0214, greater than the null interval of 0.0186–0.0214 with $p = 0.007$. In all cases, the differences were statistically significant, demonstrating that the embedding spaces produced by Open-PMC-18M differ in a measurable and consistent manner from those of the compound-image baseline. These findings confirm that large-scale subfigure extraction not only improves downstream retrieval and classification performance but also yields representational structures that are qualitatively and quantitatively distinct across modalities.

4.8.5 Interpretive Significance

Taken together, the results from both qualitative visualization and quantitative MMD testing demonstrate that OPEN-PMC produces representation spaces that are not only different from those of prior medical datasets but also superior in structure and downstream utility. The embeddings learned from OPEN-PMC-2M and Open-PMC-18M exhibit distinct latent geometries compared to PMC-15M, BIOMEDICA, and compound-image baselines (PMC-6M), with statistically significant divergence confirmed across radiology, microscopy, and visible light photography. These differences indicate that subfigure-level curation fundamentally reshapes how models internalize multimodal medical data.

Crucially, these representational distinctions align with improvements observed in retrieval, zero-shot classification benchmarks, and robustness evaluation, where OPEN-PMC consistently outperforms models trained on larger but noisier corpora. The combination of superior task performance and distinct representation structure provides convergent evidence that high-quality dataset design, rather than scale alone, drives both the effectiveness and reliability of medical vision–language models.

Thus, the evidence supports a clear conclusion: Open-PMC not only delivers better accuracy across downstream evaluations but also produces fundamentally different and higher-quality internal representations. This dual outcome highlights the central role of careful dataset curation in advancing medical vision–language learning, showing that fidelity of supervision translates into both stronger performance and qualitatively superior embedding spaces.

5 Limitations, Future Work and Conclusion

5.1 Limitations

An important limitation of our work lies in the potential biases inherent in the PubMed Central Open Access (PMC-OA) corpus, which serves as the sole source of data for our dataset. Because PMC reflects the distribution of published biomedical research, the dataset inherits disparities in what types of studies, diseases, modalities, and populations are more frequently represented. For example, well-funded institutions and high-income countries are far more likely to publish in indexed medical journals, whereas research from low- and middle-income countries often remains underrepresented. This imbalance implies that models trained on PMC-derived data may capture patterns that generalize well to dominant research communities but fail to adequately represent underreported regions, diseases, or imaging practices. Consequently, while the dataset offers wide coverage across modalities, it may not fully translate to unbiased performance in real-world clinical settings.

These concerns are supported by bibliometric evidence. A large-scale analysis of more than 10,000 articles published between 2010 and 2019 in five leading medical journals found that just four countries, USA (48.2%), UK (15.9%), Canada (5.3%), and Australia (3.2%), accounted for the majority of authorship, with 32 countries collectively representing nearly all publications during the study period (94). Such findings highlight the strong geographic skew in biomedical publishing, which inevitably shapes the data available through PMC. Although our dataset provides broader diversity than those collected from a single hospital or regional source, it remains limited by the systemic imbalances of the underlying literature.

A second limitation relates to resource constraints during the construction of Open-PMC-18M. Due to the cost and computational effort required, we were unable to extract the full set of in-text figure references at scale. Our ablation studies demonstrated that in-text references provide valuable contextual grounding and improve the quality of learned representations. even though Open-PMC-18M outperformed all the previous baselines, the absence of these references in it likely limited the extent to which contextual information could be leveraged, representing a missed opportunity to further strengthen alignment between visual and textual modalities.

Although Open-PMC models achieve strong performance on retrieval and zero-shot classification benchmarks, and is more robust compared to previous SOTA, these evaluations are limited to research datasets and do not constitute clinical validation. Before models of this type can be deployed in healthcare environments, they must undergo rigorous assessment for safety, reliability, and fairness, including testing against gold-standard diagnostic workflows and diverse patient populations. Without such validation, there remains a risk that performance observed in controlled research benchmarks may not translate to real-world clinical decision-making, where variability in image quality, patient demographics, and diagnostic context can be substantial.

Another limitation arises from the automated pipeline used to extract figures, captions, and in-text references from PubMed Central XML files. This process depends on optical character recognition (OCR) and rule-based parsing, both of which are vulnerable to errors. For example, malformed XML, mislabeled figure identifiers, or OCR mistakes in scanned documents can result in incorrect figure–caption pairings or partial

caption loss. While filtering procedures were implemented to reduce noise, some residual parsing errors are likely to remain, introducing misaligned or low-quality training pairs that may weaken the quality of learned representations.

5.2 Future Work

Building on the current work, several directions can further strengthen the utility and impact of OPEN-PMC. First, future iterations of OPEN-PMC-18M should incorporate the complete set of in-text figure references and their corresponding summaries. Our ablation studies demonstrated that in-text references provide valuable contextual grounding for figures, and scaling this process to the entire dataset would likely yield richer and more semantically aligned image-text pairs. Overcoming the computational challenges associated with large-scale reference extraction is therefore a priority for improving representation quality.

Second, while our evaluations provide strong evidence of improved representation learning, rigorous validation in clinical environments remains an essential next step. At present, all analyses have been conducted on benchmark datasets derived from research literature, which, while valuable for controlled comparisons, do not capture the full complexity of clinical practice. Before such models can be considered for deployment in real-world healthcare settings, they must be tested against gold-standard diagnostic tasks and carefully curated clinical ground truth, ensuring that their outputs are both accurate and clinically interpretable. This requires evaluation on diverse patient populations that reflect demographic, geographic, and disease variability, as well as imaging conditions that are less standardized and often noisier than those found in published research figures. Furthermore, validation studies should measure not only accuracy but also fairness, and reliability, with explicit attention to how errors may impact clinical decision-making. Establishing such validation pipelines would help determine whether the improvements observed in retrieval and zero-shot classification benchmarks translate into meaningful clinical utility, providing the necessary assurance of safety and trustworthiness required for clinical adoption.

Finally, OPEN-PMC provides a foundation for extending beyond retrieval-style supervision toward instruction-based training of generative vision–language models. Current contrastive learning approaches, while effective for aligning images and text in a shared embedding space, primarily support tasks such as retrieval and zero-shot classification. These are valuable benchmarks but do not fully capture the kinds of interactive, task-oriented capabilities that are increasingly expected of modern multimodal systems. A promising future direction is to adapt OPEN-PMC into an instruction-tuned dataset, where each medical figure (or subfigure) is paired with prompts and responses that simulate real-world tasks. For example, prompts could ask the model to describe a figure in plain language, identify abnormalities, compare two imaging modalities, or explain the clinical context of a visualization. Responses would then provide the appropriate explanation, grounded in the figure and its caption, potentially augmented with in-text references.

This direction draws inspiration from recent instruction-tuned vision–language models such as LLaVA-Med, which extend general-purpose architectures with domain-specific instruction datasets. By creating a similar instruction-based framework in the biomedical domain, OPEN-PMC could support training models that not only align images and text but also follow structured instructions and generate coherent, medically relevant outputs. Furthermore, incorporating reasoning supervision into such datasets would enable models to artic-

ulate intermediate steps or justify their predictions, rather than simply producing final answers. This shift from recognition and alignment toward reasoning and explanation is critical for building models that clinicians and researchers can trust, as it allows outputs to be more transparent, interpretable, and aligned with clinical decision-making practices.

Beyond improving interpretability, instruction-based datasets could also support a broader range of downstream applications. For instance, models trained in this way could be used to generate question–answer pairs for medical education, to assist with literature summarization by linking visual evidence to textual claims, or to power interactive diagnostic assistants capable of engaging in dialogue with clinicians. Taken together, these directions highlight that the value of OPEN-PMC is not limited to contrastive learning benchmarks but extends to serving as a foundation for the next generation of explainable and interactive medical vision–language models.

5.3 Conclusion

This thesis explored the role of dataset quality in advancing medical vision–language (VL) representation learning. While prior works such as PMC-15M, BIOMEDICA, and PMC-OA have demonstrated the value of large-scale data collection, they primarily emphasized scale over semantic fidelity. In contrast, our work systematically investigated whether careful dataset curation,through subfigure decomposition, caption segmentation, contextual summarization, and modality-aware filtering,could yield stronger and more reliable models than relying on size alone.

We introduced OPEN-PMC, a new family of curated biomedical VL datasets derived from PubMed Central articles. OPEN-PMC explicitly emphasizes quality by incorporating multiple novel refinement steps: (i) robust subfigure decomposition using a modality-balanced DAB-DETR model trained on a large synthetic compound figure corpus; (ii) segmentation of captions into subcaptions with the aid of large language models; (iii) precise alignment of subfigures with their corresponding textual references; (iv) augmentation of captions with in-text references, summarized by GPT-based models to preserve essential context; and (v) automated modality classification with expert validation. Through these processes, we produced two dataset variants: **OPEN-PMC-2M**, a smaller but fully processed collection that enabled controlled ablation studies, and **OPEN-PMC-18M**, a large-scale dataset of 18 million image–text pairs that combines subfigure-level fidelity with unprecedented coverage.

Our controlled experiments on OPEN-PMC-2M revealed several important findings. First, subfigure decomposition consistently improved cross-modal alignment, yielding large performance gains in radiology and histopathology benchmarks. Second, while subcaptions offered fine-grained supervision, full captions provided richer context and stronger overall performance. Third, the best-performing configuration combined subcaptions with in-text reference summaries, showing that contextual grounding enhances semantic alignment without overwhelming the model with irrelevant details. These insights demonstrate that dataset quality, not just scale, has a decisive impact on downstream performance.

Scaling to OPEN-PMC-18M, we confirmed that combining quality and scale delivers state-of-the-art results. Models trained on OPEN-PMC-18M outperformed strong baselines such as BioMedCLIP, PMC-OA, and BIOMEDICA across retrieval, zero-shot classification, and robustness benchmarks. In particular, OPEN-

PMC-18M achieved superior resilience to visual perturbations and statistically significant improvements in representation stability, establishing a new benchmark for medical VL pretraining. Representation analysis further revealed that embeddings learned from OPEN-PMC datasets occupy distinct and semantically coherent regions of the latent space compared to those trained on noisier corpora. Maximum Mean Discrepancy (MMD) tests confirmed that these differences are statistically significant, underscoring the representational distinctiveness induced by high-quality curation.

Together, these findings establish three main contributions of this thesis. First, we demonstrated that dataset quality,via subfigure decomposition, contextual augmentation, and modality-aware processing,substantially improves model accuracy, generalization, and robustness. Second, we introduced OPEN-PMC-2M as a high-quality benchmark for controlled ablation studies, providing concrete guidelines for optimal dataset design. Third, we delivered OPEN-PMC-18M, one of the largest curated biomedical VL datasets to date, which sets new standards for medical multimodal pretraining.

In conclusion, this work highlights that fidelity of supervision is not merely a preprocessing detail but a fundamental determinant of model quality. By curating datasets that more faithfully capture the semantics of biomedical figures, we enable vision–language models that are not only more accurate but also more robust, interpretable, and reliable. OPEN-PMC thus serves as both a practical resource for the community and a methodological case study, showing how careful dataset design can advance the next generation of medical multimodal foundation models.

Bibliography

[1] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34892–34916, 2023.

[2] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 49250–49267, 2023.

[3] S. Zhang, H. Xu, N. Usuyama, J. Bagga, and H. Poon, "Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," *arXiv preprint arXiv:2303.00915*, 2023.

[4] A. Lozano, M. W. Sun, J. Burgess, L. Chen, J. J. Nirschl, J. Gu, I. Lopez, J. Aklilu, A. W. Katzer, C. Chiu, *et al.*, "Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature," *arXiv preprint arXiv:2501.07171*, 2025.

[5] W. Lin, Z. Zhao, X. Zhang, C. Wu, Y. Zhang, Y. Wang, and W. Xie, "Pmc-clip: Contrastive language-image pre-training using biomedical documents," *arXiv preprint arXiv:2303.07240*, 2023.

[6] O. Pelka, S. Koitka, A. Meier, F. Nensa, F. Denzinger, D. Rebholz-Schuhmann, and S. Ahmed, "Roco: An image dataset for medical image understanding," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 180–188, Springer, 2018.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[8] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[10] 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*, pp. 770–778, 2016.

[11] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

[12] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015.

[13] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European conference on computer vision*, pp. 69–84, Springer, 2016.

[14] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification," *ACM Transactions on Graphics (ToG)*, vol. 35, no. 4, pp. 1–11, 2016.

[15] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[16] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[17] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations (ICLR)*, 2014.

[18] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," *Advances in neural information processing systems*, vol. 6, 1993.

[19] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304, JMLR Workshop and Conference Proceedings, 2010.

[20] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[21] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, vol. 2, pp. 1735–1742, IEEE, 2006.

[22] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," *arXiv preprint arXiv:1412.6622*, 2015. Retrieved from arXiv.

[23] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020. arXiv preprint arXiv:2002.05709.

[24] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *arXiv preprint arXiv:1911.05722*, 2019.

[25] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[26] A. Radford, J. W. Kim, J. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, M. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, pp. 8748–8763, PMLR, 2021.

[27] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning*, pp. 4904–4916, PMLR, 2021.

[28] L. Yuan, Y. Chen, X. Chen, Z. Yu, Y. Wang, L. Wang, C. Liu, D. Yu, *et al.*, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021.

[29] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv preprint arXiv:2107.07651.

[30] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "Blip: Bootstrapping language–image pre-training for unified vision–language understanding and generation," *arXiv preprint arXiv:2201.12086*, 2022.

[31] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela, "Flava: A foundational language and vision alignment model," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. arXiv preprint arXiv:2112.04482.

[32] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

[33] R. Imam, R. Marew, and M. Yaqub, "On the robustness of medical vision-language models: Are they truly generalizable?," *arXiv preprint arXiv:2505.15425*, 2025.

[34] P. K. Tankala, P. S. Pasi, S. N. Dharod, A. A. Motiwala, P. Jyothi, A. Chaudhary, and K. Srinivasan, "Wikido: A new benchmark evaluating cross-modal retrieval for vision-language models," in *NeurIPS 2024 Datasets and Benchmarks Track*, 2024.

[35] A. Daghighfarsoodeh, C.-Y. Wang, H. Taherkhani, M. Sepidband, M. Abdollahi, H. Hemmati, and H. V. Pham, "Deep-bench: Deep learning benchmark dataset for code generation," *arXiv preprint arXiv:2502.18726*, 2025.

[36] F. Di Salvo, S. Doerrich, and C. Ledig, "Medmnist-c: Comprehensive benchmark and improved classifier robustness by simulating realistic image corruptions," *arXiv preprint arXiv:2406.17536*, 2024.

[37] F. de Anda *et al.*, "Benchmarking medical vision-language models," in *[Conference]*, 2025.

[38] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015.

[39] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Schubert, *et al.*, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021.

[40] P. Li, X. Jiang, C. Kambhamettu, and H. Shatkay, "Compound image segmentation of published biomedical figures," *Bioinformatics*, vol. 34, no. 7, pp. 1192–1199, 2018.

[41] N. Kanopoulos, N. Vasanthavada, and R. L. Baker, "Design of an image edge detection filter using the sobel operator," *IEEE Journal of solid-state circuits*, vol. 23, no. 2, pp. 358–367, 1988.

[42] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.

[43] M. Taschwer and O. Marques, "Compound figure separation combining edge and band separator detection," in *International conference on multimedia modeling*, pp. 162–173, Springer, 2016.

[44] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263–7271, 2017.

[45] S. Tsutsui and D. J. Crandall, "A data driven approach for compound figure separation using convolutional neural networks," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1, pp. 533–540, IEEE, 2017.

[46] T. Yao, C. Qu, Q. Liu, R. Deng, Y. Tian, J. Xu, A. Jha, S. Bao, M. Zhao, A. B. Fogo, *et al.*, "Compound figure separation of biomedical images with side loss," in *MICCAI Workshop on Deep Generative Models*, pp. 173–183, Springer, 2021.

[47] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, pp. 213–229, Springer, 2020.

[48] S. Subramanian, L. L. Wang, S. Mehta, B. Bogin, M. van Zuylen, S. Parasa, S. Singh, M. Gardner, and H. Hajishirzi, "Medicat: A dataset of medical images, captions, and textual references," *arXiv preprint arXiv:2010.06000*, 2020.

[49] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT (1)*, pp. 4171–4186, 2019.

[50] e. a. Carlsson, "Fewer truncations improve language modeling," *arXiv preprint*, 2024.

[51] S. Khandve, V. Wagh, A. Wani, I. Joshi, and R. Joshi, "Hierarchical neural network approaches for long document classification," in *arXiv preprint*, 2022.

[52] A. et al., "Dscisum: Detailed summarization of long scientific documents," *Expert Systems with Applications*, 2025.

[53] L. van der Maaten and G. E. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[54] Y. Gao, S. Kim, D. E. Austin, and C. McIntosh, "Medbind: Unifying language and multimodal medical data embeddings," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 218–228, Springer, 2024.

[55] L. McInnes, J. Healy, N. Saul, and L. Großberger, "Umap: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[56] V. H. Do and S. Canzar, "A generalization of t-sne and umap to single-cell multimodal omics," *Genome Biology*, vol. 22, no. 1, p. 130, 2021.

[57] N. C. Hurley, A. D. Haimovich, R. A. Taylor, and B. J. Mortazavi, "Visualization of emergency department clinical data for interpretable patient phenotyping," *arXiv preprint*, 2019.

[58] H. Xu, Q. Ye, M. Yan, Y. Shi, J. Ye, Y. Xu, C. Li, B. Bi, Q. Qian, W. Wang, G. Xu, J. Zhang, S. Huang, F. Huang, and J. Zhou, "mplug-2: A modularized multi-modal foundation model across text, image and video," in *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*, pp. 38728–38748, 2023.

[59] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of neural network representations revisited," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, vol. 97 of *Proceedings of Machine Learning Research*, pp. 5728–5737, 2019.

[60] M. Maniparambil, R. Akshulakov, Y. A. Dahou Djilali, S. Narayan, M. El Amine Seddik, K. Mangalam, and N. E. O'Connor, "Do vision and language encoders represent the world similarly?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14334–14343, 2024.

[61] M. Maniparambil, R. Akshulakov, Y. A. Dahou Djilali, S. Narayan, A. Singh, and N. E. O'Connor, "Harnessing frozen unimodal encoders for flexible multimodal alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[62] M. Davari, S. Horoi, A. Natik, G. Lajoie, G. Wolf, and E. Belilovsky, "Reliability of cka as a similarity measure in deep learning," *arXiv preprint arXiv:2210.16156*, 2022.

[63] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, pp. 723–773, 2012.

[64] T. F. Tavares, F. J. Ayres, and P. Smaragdis, "Measuring similarity between embedding spaces using induced neighborhood graphs," in *International Conference on Learning Representations (ICLR) 2025*, 2025. Submitted manuscript introducing Nearest-Neighbor Graph Similarity (NNGS).

[65] L. H. Lin and N. A. Smith, "Situating sentence embedders with nearest neighbor overlap," *arXiv preprint arXiv:1909.10724*, 2019.

[66] N. Baghbanzadeh, S. Ashkezari, E. Dolatabadi, and A. Afkanpour, "Open-pmc-18m: A high-fidelity large scale medical dataset for multimodal representation learning," *arXiv preprint arXiv:2506.02738*, 2025.

[67] N. Baghbanzadeh, A. Fallahpour, Y. Parhizkar, F. Ogidi, S. Roy, S. Ashkezari, V. R. Khazaie, M. Colacci, A. Etemad, A. Afkanpour, *et al.*, "Advancing medical representation learning through high-quality data," *arXiv preprint arXiv:2503.14377*, 2025.

[68] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[69] J. Silva-Rodr´iguez, A. Colomer, M. A. Sales, R. Molina, and V. Naranjo, "Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection," *Computer methods and programs in biomedicine*, vol. 195, p. 105637, 2020.

[70] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[71] J. Yang, R. Shi, and B. Ni, "Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 191–195, IEEE, 2021.

[72] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification," *Scientific Data*, vol. 10, no. 1, p. 41, 2023.

[73] A. Pacheco, G. Lima, A. Salomao, B. Krohling, I. Biral, G. de Angelo, F. Alves Jr, J. Esgario, A. Simora, P. Castro, *et al.*, "Pad-ufes-20: a skin lesion dataset composed of patient data and clinical images collected from smartphones. data brief 32: 106221," 2020.

[74] N. Methani, P. Ganguly, M. M. Khapra, and P. Kumar, "Plotqa: Reasoning over scientific plots," in *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pp. 1527–1536, 2020.

[75] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in *International Conference on Learning Representations*, 2022.

[76] A. Garc´ia Seco de Herrera, R. Schaer, S. Bromuri, and H. Müller, "Overview of the ImageCLEF 2016 medical task," in *Working Notes of CLEF 2016 (Cross Language Evaluation Forum)*, September 2016.

[77] K. Jobin, A. Mondal, and C. Jawahar, "Docfigure: A dataset for scientific document figure classification," in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 1, pp. 74–79, IEEE, 2019.

[78] E. Schwenker, W. Jiang, T. Spreadbury, N. Ferrier, O. Cossairt, and M. K. Chan, "Exsclaim!–an automated pipeline for the construction of labeled materials imaging datasets from literature," *arXiv preprint arXiv:2103.10631*, 2021.

[79] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[80] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[81] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

[82] T. Kudo, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *arXiv preprint arXiv:1808.06226*, 2018.

[83] J. Ikezogwo and Others, "Quilt: A microscopy dataset for cross-modal retrieval," *Journal Name*, 2024. Placeholder entry. Update with correct details.

[84] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr: A large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[85] J.-H. Huang, C.-H. H. Yang, F. Liu, M. Tian, Y.-C. Liu, T.-W. Wu, I. Lin, K. Wang, H. Morikawa, H. Chang, *et al.*, "Deepopht: medical report generation for retinal images via deep models and visual explanation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2442–2452, 2021.

[86] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "Medmnist v2—a large-scale lightweight benchmark for 2d and 3d biomedical image classification," *Scientific Data*, vol. 10, no. 1, p. 41, 2023.

[87] J. Silva-Rodr´iguez, A. Colomer, M. A. Sales, R. Molina, and V. Naranjo, "Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection," *Computer Methods and Programs in Biomedicine*, vol. 195, 2020.

[88] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Patchcamelyon (pcam): A patch-based histopathology classification benchmark," 2018. arXiv preprint arXiv:1806.03962.

[89] J. N. Kather, N. Halama, and A. Marx, "100 000 histological images of human colorectal cancer and healthy tissue," *Scientific Data*, vol. 5, p. 180122, 2018.

[90] 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)," *arXiv preprint arXiv:1912.12142*, 2019.

[91] T. A. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, and et al., "Bach: Grand challenge on breast cancer histology images," in *Proceedings of the International Conference on Image Analysis and Recognition (ICIAR)*, 2018.

[92] A. G. C. Pacheco, G. R. Lima, A. S. Salomão, B. A. Krohling, I. P. Biral, G. G. de Angelo, F. C. R. Alves Jr, J. G. M. Esgario, A. C. Simora, P. B. C. Castro, and et al., "Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data Brief*, vol. 32, p. 106221, 2020.

[93] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[94] O. Brück, "A bibliometric analysis of geographic disparities in the authorship of leading medical journals," *Communications Medicine*, vol. 3, no. 1, p. 178, 2023.

Appendices

A GPT prompts

We utilized the **GPT-4o and GPT-4o-mini** models for several tasks. First, GPT-4o used it to **extract subcaptions** associated with each compound image. Additionally, GPT-4o mini was employed to **summarize in-text references** related to figures and to **identify the modality** of the images. In this section, we describe the prompts used for each task.

A.1 Subcaption Segmentation and Alignment

We utilized GPT-4o to segment full captions into panel-specific subcaptions. For captions that could not be reliably separated, the entire caption was retained for the corresponding single-panel figure. To further refine panel-text alignment, we used object localization (YOLOv3) and recognition (ResNet-152) models from Exsclaim to detect labels within subfigures. These detected labels enable automatic matching between subfigures and their corresponding subcaptions. For subfigures without identifiable labels, we assign the entire original caption to preserve textual context.

GPT-4o Prompt for Subcaption Extraction

System Prompt
Subfigure labels are letters referring to individual subfigures within a larger figure. Check if the caption contains explicit subfigure label. If not, output "NO" and end the generation. If yes, output "YES", then generate the subcaption of the subfigures according to the caption.

The output should use the template:
YES
Subfigure-A: ...
Subfigure-B: ...
...

The label should be removed from subcaption.

User Prompt
Caption: CAPTION

A.2 In-text References

In-text references provide crucial contextual information about biomedical images, containing clinical interpretations, methodology descriptions, and result analyses that complement the caption. Using XML cross-reference tags, we extract all paragraphs referencing each figure in the article text and break them down into individual sentences using regular expressions. Here, we face three challenges: in-text references span across multiple paragraphs which often exceeds our context length, relevant information may exist outside

sentences directly referencing the target subfigure, and for compound figures, it is difficult to determine which references describe which specific subfigures. To address these issues, we employ GPT-4o-mini using the following prompt to generate focused summaries:

GPT-4o-mini Prompt to Summarize In-text References

You are provided with a **medical text** (**Context**) and a **target subfigure** (**Target Subfigure**). Your task is to extract a comprehensive figure caption from the Context that is specific to the Target Subfigure.
Required Elements to Include (if present in Context):

- The **main finding or observation** shown in the Target Subfigure.
- Key **experimental conditions or methods** relevant to the Target Subfigure.
- Any **relevant information** from the Context necessary to understand the Target Subfigure.

Critical Requirements:

- Do **not include information** about other subfigures.
- Do **not use external knowledge**; rely solely on the provided Context.
- Do **not add interpretation or analysis** beyond what is explicitly stated.
- Ensure all information is **explicitly present** in the Context.
- Keep the caption concise, **under 250 words**, while preserving essential details.
- Focus on content that is **concise, relevant, objective, and factual**.

Output Format:

- **Target Subfigure**: [subfigure number_label]
- **Caption**: [detailed caption following the elements above]

Context: {CONTEXT}
Target Subfigure: {FIGURE}

A.3 Modality Classification

After separating the compound images and extracting the subcaptions, we utilized **GPT-4o-mini** to determine the modality of each individual image. For this task, we provided the subcaption of each image to GPT and instructed it to classify the image into one of the following categories: **Medical**, which includes **Radiology**, **Microscopy**, and **Visible Light Photography**, or **Ambiguous**. The **Ambiguous** category accounts for cases where captions are either too brief or lack sufficient information to determine the image modality. The following prompt was used to classify the image modality.

GPT-4o-mini Prompt for Image Modality Classification

You are an AI assistant specialized in biomedical topics. You are provided with a text description of a figure (**Figure Caption**) from a biomedical research paper.
Your task is to classify the image into **one of four categories** by following a two-step process:
Step 1: Determine if the caption is Diagnostic or Non-Diagnostic

- **Diagnostic Captions**: These are captions of medical images used to directly **visualize biological structures, detect abnormalities, or assist in clinical decision-making**. They fall into one of the following categories:
  - **Radiology (R)**: X-rays, CT scans, MRI, PET scans, Ultrasound, Angiography, or similar medical imaging methods.
  - **Visible Light Photography (V)**: Clinical photographs of external or internal body structures, such as dermatology images (skin lesions), endoscopic images (internal organs), or ophthalmology images (retinal scans).
  - **Microscopy (M)**: Images obtained using a microscope to analyze tissue samples, cells, bacteria, or subcellular structures, including light, electron, fluorescence, and transmission microscopy.
- **Non-Diagnostic Captions**: Captions for images that do not directly aid in medical diagnosis but are used for illustration or explanation. They include:
  - **Statistical or Analytical Figures**: Graphs, bar charts, line plots, and pie charts.
  - **Diagrams and Schematics**: Flowcharts, system overviews, block diagrams, and hand-drawn illustrations.
  - **Data Representations**: Tables, program listings, gene sequences, and molecular structures.
  - **Screenshots or UI Elements**: Software interfaces, algorithm visualizations, or non-medical computational images.
  - **Non-Clinical Photographs**: Images of lab equipment, experimental setups, or environmental objects.

Step 2: If the caption is Diagnostic, classify it into one of the following categories:

1. **R (Radiology)**: Medical imaging techniques used to visualize internal body structures, such as X-rays, CT scans, MRI, PET, Ultrasound, and Angiography.
2. **V (Visible Light Photography)**: Photographic images of external or internal body structures, including dermatology, endoscopy, and ophthalmology images.
3. **M (Microscopy)**: Any image captured using a microscope, such as light microscopy, electron microscopy, fluorescence microscopy, or transmission microscopy.

If the caption of the image is **not diagnostic**, return **N (Non-Diagnostic)**.
Response Format: Your response must be a **single letter only**, with no explanations:

- **R** for Radiology
- **V** for Visible Light Photography
- **M** for Microscopy
- **N** for Non-Diagnostic

Caption of the image is: {CAPTION}