Histopathology Multi-modal Embedding for Pathology Composed Retrieval

Qifeng Zhou[1], Wenliang Zhong[1], Thao M. Dang[1], Hehuan Ma[1], Saiyang Na[1], Yuzhi Guo[1], and Junzhou Huang[1]

The University of Texas at Arlington, Arlington TX 76010, USA
{qxz8706, wxz9204, tmd4090, hehuan.ma, sxn3892, yuzhi.guo}@mavs.uta.edu
jzhuang@uta.edu

Abstract. To overcome the black-box nature of predictive AI and the hallucination risks of generative models, retrieval-based models offer an interpretable, evidence-based paradigm for pathology clinical workflow. However, real-world clinical queries are inherently interleaved (e.g., pathology images and text). Current dual-encoders suffer from an **Architectural Mismatch**, lacking the mechanism to fuse such composed queries. To address this, we formalize the task of Pathology Composed Retrieval (PCR). While Multimodal Large Language Models (MLLMs) offer deep-fusion capabilities, directly applying them exposes a **Task Mismatch** and a **Domain Mismatch**. To resolve these challenges, we propose HOMIE, a model-agnostic adaptation framework that transforms any generative MLLM into a specialized pathology retrieval expert. Evaluated on our newly introduced PCR Benchmark, a lightweight 2B-parameter HOMIE variant substantially outperforms existing paradigms, surpassing specialized 7B pathology MLLMs and dual-encoders by large margins on composed retrieval, while maintaining strong performance on traditional simple retrieval. The project page is available at https://qfchou.github.io/HOMIE_page/.

Keywords: Computational Pathology · Pathology Composed Retrieval

1 Introduction

The digitization of pathology has created a promising opportunity for Artificial Intelligence (AI) to enhance clinical diagnosis [28, 36, 42]. However, current AI models in this high-stakes domain face trust barriers. Traditional supervised models are "black boxes" that lack transparency, while emerging Large Language Models (LLMs) introduce the clinically unacceptable risk of hallucination [12,15]. Given these risks, a retrieval-based paradigm is more suitable for clinical application [4, 6]. Rather than predicting a diagnosis, a retrieval system functions as a "computational consult" [34]. It searches its database for relevant information based on a pathologist's query [1,21]. This approach empowers pathologists to examine this information and draw their own expert-guided conclusions, preserving clinical autonomy and fostering trust [5, 13]. Therefore, a powerful, specialized retrieval model is important for advancing scalable AI in pathology.

While recent multi-modal models have enabled basic pathology retrieval [28, 42], they are limited to simple retrieval (e.g., image-to-text or text-to-image) by their dual-encoder architectures. Real-world clinical queries in pathology are inherently compositional and interleaved. Architecturally, dual-encoders lack a mechanism to deeply fuse such interleaved multi-modal inputs, resulting in an **Architectural Mismatch**. Forcing these models to perform composed retrieval may discard critical cross-modal semantic interactions. To address this gap, we formally define the task of **Pathology Composed Retrieval (PCR)**: retrieving the most relevant items from a database given a query composed of an interleaved sequence of pathology images, texts, and videos. This task requires a multi-modal embedding model, which can generate a single, semantically embedding from any combination of modalities for complex retrieval tasks.

A natural intuition to overcome this architectural mismatch is Multimodal Large Language Models (MLLMs), whose deep-fusion architectures natively process interleaved inputs [26]. However, directly applying MLLMs to PCR presents some challenges. First, **Task Mismatch**: MLLM's latent spaces are optimized for generation, lacking the discriminative metric space required for retrieval [19]. Our empirical evaluations reveal that even state-of-the-art pathology MLLMs (e.g., PathoR1 [46]) perform poorly on PCR tasks. Furthermore, general MLLMs adapted for retrieval suffer from **Domain Mismatch**, as they cannot interpret pathology features like subtle cellular morphology or staining artifacts [33]. To solve these challenges, we propose HOMIE framework. Crucially, HOMIE is not a monolithic model, but a **model-agnostic, systematic adaptation framework** designed to transform any generative MLLM into a specialized pathology retrieval expert capable of generating a unified multi-modal embedding.

HOMIE systematically resolves all three aforementioned mismatches. First, by leveraging an MLLM backbone, HOMIE inherently overcomes the **Architectural Mismatch**. To resolve the remaining barriers, we employ a two-stage process. A text-only retrieval-adaptation stage forces the generative LLM to learn a discriminative metric space, directly solving the **Task Mismatch**. Subsequently, a pathology-specific tuning stage addresses the **Domain Mismatch** by implementing dynamic native-resolution processing, targeted stain augmentation, and a progressive knowledge curriculum. To evaluate this, we also introduce the first PCR Benchmark, enabling quantification of these new tasks.

Our main contributions are summarized as follows: (1) We formally define the novel PCR task and introduce the PCR Benchmark for its rigorous evaluation. (2) We propose HOMIE, a novel, model-agnostic adaptation framework that resolves the Architectural, Task, and Domain mismatches, successfully transforming generative MLLMs into pathology retrieval experts. (3) Extensive experiments demonstrate that HOMIE outperforms specialized dual-encoder and MLLM-based methods by large margins on PCR tasks and maintains strong performance on traditional simple retrieval.

2 Related Work

Vision-language Models in Pathology. Recent studies in pathology have adopted the CLIP [30] model, leveraging paired image-text data within a contrastive learning framework to align similar image-text embeddings while separating dissimilar ones. These models, including PubmedCLIP [9], BiomedCLIP [45], PLIP [16], PathCLIP [38], QuiltNet [17], PathgenCLIP [37], PathoCLIP [46], trained on amounts of pathology image-text pairs, excel at basic cross-modal retrieval tasks like image-to-text and text-to-image search. More advanced models, such as CONCH [28] and MUSK [42], achieved stronger performance by adopting more sophisticated architectures (CoCa [44] for CONCH and BEiT3 [41] for MUSK) and incorporating generative captioning objectives alongside contrastive alignment, surpassing the performance of simple CLIP-based methods. However, all these methods are constrained by an architectural paradigm reliant on separate encoders for image and text. The input is either a single text or a single image. Furthermore, all these models process images at a fixed, low resolution (e.g., 336x336), discarding the fine-grained morphological details critical for pathological analysis. Our framework is the first to move beyond this rigid design, using a unified MLLM to generate a true omni-modal embedding from interleaved, multi-modal data.

Multimodal Embedding Learning. MLLMs have extended LLMs to process multiple data modalities, achieving notable progress in understanding and reasoning across diverse input types [22,25,26]. To adapt MLLMs for retrieval task, the recent work E5-V [19] fine-tunes an LLM with summarization prompts and text-only data to extract embeddings, then integrates a vision module to obtain multimodal embeddings for zero-shot multimodal retrieval tasks. VLM2Vec [20], LamRA [26] and GME [47] leverage multimodal data and prompt-tuning to accommodate diverse queries and modalities, achieving strong performance on multimodal retrieval tasks. However, these models are designed for and pre-trained on general-domain data. This creates Domain Mismatch, as these models are not equipped to interpret pathology images, such as subtle morphological variations or staining artifacts. For pathological adaptation, directly applying these frameworks requires a domain-specific LLM or MLLM pretrained on pathology instruction data, which demands detailed and standardized annotations. In contrast, our framework, HOMIE can adapt an MLLM to generate omni-modalembeddingsg for pathology composed retrieval with only public pathology image-text pairs, bypassing these prohibitive requirements.

3 Methods

3.1 Pathology Composed Retrieval Benchmark

Formally, we define the Pathology Composed Retrieval (PCR) task as follows: Given a query q and a large candidate set $\mathcal{C} = \{c_1, c_2, \cdots, c_N\}$, the aim is to find similar candidates in $\mathcal{C}$ based on their similarity scores to q. A key feature of PCR

Table 1: The statistics of our PCR benchmark. It includes two retrieval tasks with selected samples from six data sources. Our benchmark has various combinations of Image (Img.), Text (Txt.), and Video (Vid.) modalities for query and target sides.

Task	Data Source	Query Mod.	Target Mod.	# of Samples
Composed Retrieval	Bookset [10]	Multi Img.	Txt.	600
	Bookset [10]	Img. & Txt.	Img.	488
	Quilt-VQA [32]	Img. & Txt.	Txt.	724
	Quilt-VQA-RED [32]	Img. & Txt.	Txt.	252
	Videopath [40]	Vid.	Txt.	244
Simple Retrieval	Bookset [10]	Img.	Txt.	2,688
	Educontent [36]	Img.	Txt.	947
	MMUpubmed [36]	Img.	Txt.	1,385
	Bookset [10]	Txt.	Img.	2,688
	Educontent [36]	Txt.	Img.	947
	MMUpubmed [36]	Txt.	Img.	1,385

is that both the query q and each candidate c_i are multi-modal, meaning they can be a single image, a single text, a single video, or any interleaved sequence.

To address the critical evaluation gap, we introduce the Pathology Composed Retrieval (PCR) Benchmark. This benchmark is designed to test a model's ability to perform compositional retrieval, moving beyond simple image-text retrieval. We created a suite of novel pathology composed retrieval tasks by repurposing existing public datasets:

Multi-Image to Text Retrieval $(q^i, q^i, ...) \rightarrow c^t$**:** This task measures a model's ability from multiple images to retrieve a single, corresponding caption. We extract from Bookset [10], which provides multi-image with a single caption.

Image-Text to Image Retrieval $(q^i, q^t) \rightarrow c^i$**:** This task tests fine-grained compositional reasoning. Instead of generating new content, we utilize GPT-5 strictly to parse the original expert diagnostic texts from the multi-image set in Bookset [10] and extract existing relational clauses (e.g., "a magnified view of the atypia..."). This creates a query (source image + relational text) to retrieve the correct target image while preserving absolute clinical fidelity.

Image-Text to Text Retrieval $(q^i, q^t) \rightarrow c^t$**:** This task reframes Visual Question Answering (VQA) to simulate pathologist inquiry and we extract from Quilt-VQA [32] and Quilt-VQA-RED [32]. To prevent models from "cheating" by exploiting lexical overlap between questions and answers, we apply GPT-5 purely as a linguistic formatter to restructure the phrasing of the answer. This ensures the underlying clinical semantics remain unaltered while forcing the model to genuinely ground its reasoning in the visual evidence.

Video-to-Text Retrieval $q^v \rightarrow c^t$: This task is extracted from VideoPath dataset [40]. The model should retrieve a text diagnostic description with a video query, testing the fusion of spatio-temporal features.

As shown in Table 1, these tasks can directly measure a model's ability to generate a true multi-modal embedding for pathology. Detailed data curation

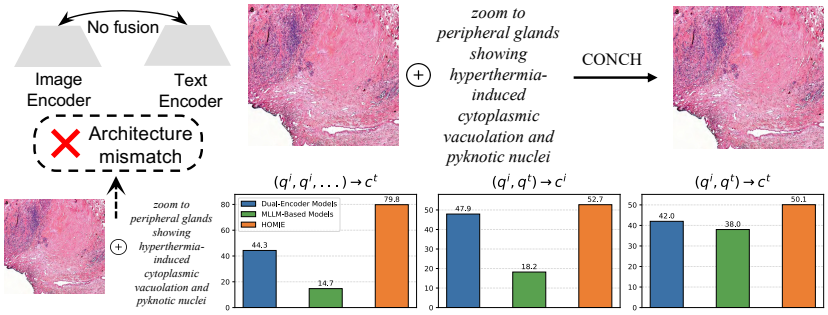

Fig. 1: Analysis of Existing Models on PCR. **Left**: Dual-encoder models face an *Architectural Mismatch*, forcing them to process interleaved queries separately. **Top Right**: Dual-encoder models fail in composed retrieval and degenerate into shallow matching, erroneously retrieving the original image. **Bottom Right**: Empirical results on the PCR Benchmark. Both paradigms exhibit substantial performance degradation: dual-encoders due to *Architectural Mismatch*, and MLLMs due to *Task and Domain Mismatches*.

protocols and the exact LLM prompts, which are explicitly engineered to strictly prohibit the generation of new content, ensuring the rewriting process serves solely as a linguistic formatter while absolutely preserving the original medical facts and clinical fidelity, are provided in the supplementary material.

3.2 Analysis of Existing Models on PCR

Before detailing the specific architecture, we first analyze why existing models fail at PCR task. As shown in the bar charts of Figure 1 (bottom right), on the $(q^i, q^i, ...) \rightarrow c^t$ task, the best dual-encoder and MLLM-based model reach only 44.3% and 14.7% Recall@1, respectively, whereas HOMIE achieves 79.8%. By investigating these empirical failures, we identify the fundamental mismatches that limit current models:

The Architectural Mismatch of Dual-Encoder Models. As illustrated in Figure 1 (left), dual-encoder models separate visual and textual inputs. Because these models lack a mechanism to fuse interleaved multi-modal queries, they rely only on independent, shallow similarity scores. As shown in Figure 1 (top right), instead of performing true multi-modal reasoning to find the correct diagnostic target, these models often reduce to visual matching, erroneously retrieving the original unannotated image itself.

The Task and Domain Mismatches of MLLMs. MLLMs inherently possess the deep-fusion architecture needed for interleaved inputs. However, directly deploying them exposes two flaws. First, a Task Mismatch: their latent spaces are optimized for token generation rather than contrastive metric learning, making their hidden states unsuited for retrieval. Second, a Domain Mismatch: general-domain MLLMs lack the morphological priors necessary to interpret complex pathology data.

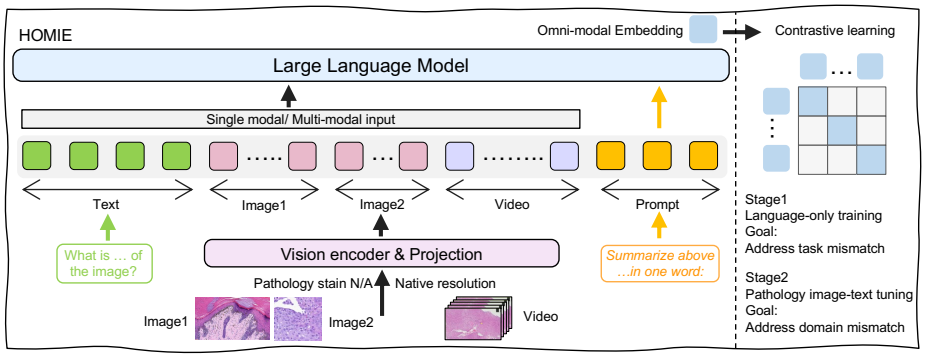

Fig. 2: Overview of the HOMIE framework. The model ingests arbitrary modalities (e.g., image, text, video) and leverages a prompt-guided LLM to generate a unified omni-modal embedding. The vision pathway is optimized for pathology data, featuring pathology-specific stain normalization/augmentation and native resolution input. We employ a two-stage contrastive learning strategy to train the model, addressing task mismatch followed by domain mismatch.

The HOMIE Solution. To systematically overcome these diagnosed barriers, we propose the HOMIE framework. We detail the architecture design and the two-stage adaptation strategy in the subsequent sections.

3.3 Architecture and Omni-modal Embedding

Our model consists of three critical components: a vision encoder f_v, an MLP-based projector f_p, and a LLM f_φ. The method overview is shown in Figure 2.

Native Resolution and Video Processing. To process visual inputs without discarding crucial diagnostic details, we utilize the SigLIP-2 ViT architecture [39]. For static images, instead of forcibly resizing inputs to a fixed, low resolution (e.g., 336×336), we dynamically adjust the image dimensions to multiples of a base patch size (14×14). Then we apply 2D Rotary Positional Embeddings (2D-RoPE) [35] to keep the position information. This design guarantees that the model processes pathology images at their native resolutions efficiently. For video inputs, we utilize frame rate sampling and interleaved multi-RoPE [14] for robust spatio-temporal understanding.

Multi-modal Embedding Extraction. Given a visual input V (image, video, or interleaved sequence), it is encoded and projected into a sequence of visual tokens $h_v = f_p(f_v(V))$. The LLM f_φ then processes h_v alongside any text tokens h_t to produce a multi-modal embedding $E = f_\varphi(h_v, h_t)$. To extract a single, dense representation from a LLM, we employ the Explicit One-word Limitation (EOL) prompting strategy [18]. The rationale behind EOL lies in the architectural difference between masked and causal language models. Unlike masked language models (e.g., BERT) that use bidirectional attention to inherently condense semantic information into a special `[CLS]` token, LLMs rely on

causal attention. Conventional pooling strategies, such as mean pooling across all tokens, are sub-optimal for causal LLMs because early tokens lack access to future context, inherently diluting the global representation. Without explicit constraints, the hidden states of LLMs are naturally optimized for open-ended text generation rather than discriminative metric learning [18]. Recent literature [18] has empirically demonstrated that the EOL strategy overcomes this limitation and significantly outperforms traditional pooling methods, establishing it as the SOTA approach for extracting dense representations from LLMs. By explicitly appending the phrase "in one word:" to the prompt, we force the LLM to encapsulate and compress the entirety of the preceding multi-modal context into the target token's hidden state. Specifically, we use the following prompts: (i) for image-only input, we use: `<image> Summarize above image in one word: <emb>`; (ii) for text-only input, we use: `<text> Summarize above sentence in one word: <emb>`; (iii) for mixed image-text input, we utilize: $\texttt{<image}_1\texttt{><text}_1\texttt{>...<image}_i\texttt{><text}_j\texttt{> Summarize above image and sentence in one word: <emb>.}$ where `<image>` and `<text>` denote placeholders for the input image and text, respectively. We use the last hidden state immediately preceding the `<emb>` token as the representation of the input.

3.4 The HOMIE Framework

We introduce HOMIE, a systematic framework to adapt an MLLM into an omni-modal, pathology-specific retrieval model. Our framework is a two-stage process. The first stage solves task mismatch by adapting LLM for retrieval using language-only pre-training. The second stage solves the domain mismatch by pathology-specific tuning. This framework is trained using only public data, which we systematically collated and curated to support this process.

Training Dataset Curation. Our training data includes text-only data and image-text pairs. Text-only data is used for Stage one. We use text pairs from Natural Language Inference (NLI) dataset [11], MedNLI [31] and MedMCQA [29]. To ensure high domain specificity, we curated a pathology-specific subset from MedMCQA [29]. Image-text pairs are used for Stage two and are aggregated from PathGen-1.6M [37], PathCap [38], and Quilt-1M [17]. We observed that naively mixing these heterogeneous sources leads to suboptimal performance. We identified the primary cause: Noisy image–text pairs collected from the web present challenges for training and may degrade the model performance. Therefore, a core part is bootstrap [23] approach: we first trained a based model on the unfiltered Quilt-1M data, then used this model to score the alignment of all pairs in Quilt-1M. All pairs with an image-text similarity score below a predefined threshold $\lambda = 0.1$ were discarded. This curation and filtering process is the first step in our solution, ensuring that our specialization stages are built upon high-quality, domain-relevant data.

Stage One: Adapting for Retrieval. The first stage addresses task mismatch. MLLMs are optimized for generative tasks, not retrieval, meaning their latent space isn't structured for metric learning. To solve this issue, we train LoRA modules on our curated, pathology-specific text data to adapt the LLM for

retrieval tasks. This stage forces the LLM to learn a semantically structured embedding space, making it suitable for retrieval before it ever sees images.

Stage Two: Pathology-Specific Curriculum Tuning. The second stage addresses the domain mismatch by fine-tuning the entire MLLM on our curated pathology image-text pairs. Different from natural images, pathology images present unique challenges, including chiefly staining variability and complex morphological features, which simple fine-tuning cannot solve. We therefore employ a pathology-specific tuning that incorporates three key adaptations. First, at the input level, we apply pathology stain augmentation and normalization [33], forcing the model to learn stain-invariant morphology. Second, we leverage our architecture's ability to process images at their native, original resolution. This is crucial as it ensures the model can analyze the multi-scale, fine-grained details that fixed-resolution models discard. Finally, we employ a Progressive Knowledge Curriculum instead of simply mixing datasets. This staged pathology curriculum tuning: the model is first exposed to Pathgen-1.6M [37], which emphasizes tissue-cell morphology and spatial organization to build foundational morphological priors. Then the model is trained on PathCap [38] and our filtered Quilt-1M to learn how to associate these morphologies with high-level, multimodal diagnostic knowledge that includes diagnostic information.

3.5 Training Objective

We employ contrastive learning with the InfoNCE loss [30] for both language-only pre-training and pathology tuning stages. Specifically, given a batch size of B, the embeddings of i-th query q_i should be positioned close to the embeddings of its positive target c_i and far away from other negative instances, formulated as:

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B}\log\left[\frac{\exp\left[cos\left(q_i, c_i\right)/\tau\right]}{\sum_{j=1}^{B}\exp\left[cos\left(q_i, c_j\right)/\tau\right]}\right],$$

where τ is a temperature parameter and $cos(q_i, c_i)$ represent the cosine similarity of q_i and c_i in contrastive learning.

4 Experiments

Training Datasets and Metrics. Our framework is trained entirely on publicly available data, as detailed in Sec 3.1. For stage one, we use the full NLI dataset [11] and a curated text corpus consisting of 14k pairs from MedNLI [31] and 15k pathology-specific QA pairs filtered from MedMCQA [29]. For stage two, we follow our pathology progressive tuning. We first use 1.6M pairs from PathGen-1.6M [37] to instill foundational knowledge in tissue-cell morphology and spatial organization. We then use 0.2M pairs from PathCap [38] and 0.5M pairs from Quilt-1M [17] (after our rigorous filtering) to develop broader multimodal knowledge that includes diagnostic information. For the retrieval tasks, we primarily utilize Recall@K as the evaluation metric.

Baseline Methods. To demonstrate the effectiveness and necessity of HOMIE, we evaluate against two categories of baselines. We emphasize that existing paradigms fundamentally struggle with the PCR task due to architectural or objective mismatches. These comparisons aim to highlight the bottlenecks of current approaches rather than demonstrating the advantages of model scale. (1) **Pathology Dual-Encoder Models.** This group represents the current SOTA in pathology simple retrieval. We evaluate standard CLIP-based architectures, including PubmedCLIP [9], BiomedCLIP [45], PLIP [16], PathCLIP [38], QuiltNet [17], PathgenCLIP [37], PathoCLIP [46], alongside advanced models like CONCH [28] and MUSK [42] (based on CoCa [44] and BEiT3 [41]). Due to their rigid dual-encoder architectures, these models inherently lack the physical mechanism to deeply fuse interleaved multi-modal queries. To enable their evaluation on the PCR task, we employ parameter-free vector addition and reciprocal rank fusion (RRF), establishing the upper bound of traditional non-MLLM paradigms and exposing their **Architectural Mismatch**. (2) **MLLM-based Models.** We evaluate MLLMs by categorizing them into three sub-types. First, *General MLLMs* (e.g., Qwen2.5-VL-7B [3], Qwen3-VL-2B [2] and Qwen3-VL-8B [2]) serve as our foundation control, lacking both retrieval optimization and pathology priors, thus suffering from both task and domain mismatches. Second, *Retrieval-adapted General MLLMs* (e.g., E5-V [19], LamRA [27], and GME [47]) have been specifically instruction-tuned for cross-modal retrieval in natural images. Their catastrophic failure on the PCR benchmark strictly isolates and exposes the severe **Domain Mismatch**. Finally, *Pathology MLLMs* (e.g., HuatuoGPT-Vision-7B [7], Lingshu-7B [43], and PathoR1-7B [46]) possess comparable parameter scales ($\sim$7B) and rich pathology priors. Their inclusion demonstrates that generative pathology knowledge does not naturally translate to a discriminative metric space, explicitly isolating the **Task Mismatch**. Together, these sub-types validate the absolute necessity of HOMIE's two-stage adaptation framework.

Implementation Details. We implement HOMIE using PyTorch and DeepSpeed ZeRO-2. All training is performed on 8 NVIDIA H100 GPUs with BF16 precision and FlashAttention-2 [8] enabled. For stage 1, we train for 2 epochs with a global batch size of 576. The learning rate is set to 2×10^{-4} with a cosine decay and 0.03 warmup ratio. In this stage, the vision encoder and projector are frozen. We apply LoRA to the LLM with $r = 64$ and $\alpha = 128$. For stage 2, we initialize from the Stage 1 checkpoint and train for 2 epochs with a global batch size of 384. The learning rate is adjusted to 1×10^{-4}. Crucially, we unfreeze the vision encoder and the projector to capture domain-specific features. The LLM is trained via LoRA with an increased rank of $r = 128$ and $\alpha = 256$.

4.1 Zero-shot Composed Retrieval

Table 2 reports zero-shot composed retrieval results, closely mirroring real-world clinical workflows (e.g., interleaved images and text). These results validate our framework while exposing the limitations of existing paradigms.

Table 2: Zero-shot composed retrieval performance, reporting Recall (R@k) metrics (in percentage %) at various thresholds. Bold and underlined values indicate the best and second-best performance, respectively. Dashed lines (-) indicate non-applicability, as these methods cannot process video inputs. Add and RRF denote parameter-free vector addition and reciprocal rank fusion, respectively.

Model	Bookset			Bookset			Quilt-VQA			Quilt-VQA-Red			Videopath		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Pathology Dual-Encoder Models (Add)															
PubmedCLIP [9]	0.8	1.8	3.0	7.7	17.3	23.9	21.1	31.4	35.5	30.2	45.2	50.8	-	-	-
BiomedCLIP [45]	19.0	40.0	57.2	17.8	39.9	52.6	10.9	24.0	31.1	19.0	33.7	46.0	-	-	-
PLIP [16]	6.8	20.2	31.0	26.8	49.6	59.8	22.1	33.7	39.0	26.6	42.5	46.0	-	-	-
PathCLIP [38]	11.2	36.8	48.5	22.4	44.6	57.9	21.3	36.3	40.6	33.7	46.4	50.8	-	-	-
QuiltNet [17]	12.3	28.2	37.7	26.6	45.5	52.6	18.9	29.4	34.7	28.2	36.9	41.3	-	-	-
CONCH [28]	41.8	71.7	79.7	43.7	70.8	81.1	9.7	25.6	36.0	18.3	42.5	58.3	-	-	-
Pathgen-CLIP-L [37]	22.5	47.7	57.5	39.8	70.8	80.7	26.4	39.6	44.8	36.9	49.2	53.2	-	-	-
Pathgen-CLIP [37]	23.0	52.8	64.3	39.2	66.5	78.9	28.2	41.6	46.0	35.3	50.0	56.3	-	-	-
MUSK [42]	43.0	72.7	83.0	47.9	74.2	84.3	34.1	53.0	58.4	50.0	69.8	77.8	-	-	-
Patho-CLIP-L [46]	44.3	77.5	88.0	40.7	69.7	78.8	22.1	40.5	48.6	35.3	58.7	66.3	-	-	-
Patho-CLIP [46]	43.3	75.8	85.7	35.6	61.8	73.6	14.1	28.7	35.5	21.8	40.1	47.2	-	-	-
Pathology Dual-Encoder Models (RRF)															
PubmedCLIP [9]	0.4	1.7	3.0	2.5	9.8	16.4	1.7	7.5	14.4	3.6	15.5	27.8	-	-	-
BiomedCLIP [45]	15.8	38.3	51.5	15.8	37.5	52.5	3.9	10.1	16.6	7.9	21.8	36.1	-	-	-
PLIP [16]	5.0	17.8	23.5	12.1	35.2	53.9	2.9	10.5	15.1	8.3	24.2	35.3	-	-	-
PathCLIP [38]	9.8	27.0	42.2	19.3	39.8	56.4	3.3	9.9	16.3	12.3	28.6	39.3	-	-	-
QuiltNet [17]	10.3	29.5	41.0	15.4	34.6	46.3	4.8	11.9	17.5	14.7	27.8	36.9	-	-	-
CONCH [28]	36.5	68.5	78.3	34.4	63.9	75.6	4.3	13.5	23.6	13.5	33.3	49.2	-	-	-
Pathgen-CLIP-L [37]	22.2	56.7	71.8	27.7	58.2	74.6	8.0	19.5	26.2	26.6	45.6	53.6	-	-	-
Pathgen-CLIP [37]	19.7	51.3	66.5	28.3	59.8	71.5	9.9	19.9	27.5	23.0	44.0	54.0	-	-	-
MUSK [42]	44.3	80.3	88.3	34.0	68.9	82.2	17.5	30.5	39.5	37.3	62.3	71.4	-	-	-
Patho-CLIP-L [46]	43.2	76.8	86.5	33.6	62.3	77.0	9.8	22.4	30.1	25.0	48.8	59.1	-	-	-
Patho-CLIP [46]	42.8	74.3	85.8	30.1	58.2	71.1	6.9	16.2	22.8	14.7	33.7	44.8	-	-	-
MLLM-Based Models															
Qwen2.5-VL-7B [3]	3.5	8.7	16.0	3.9	12.7	23.2	14.4	27.6	35.1	24.2	42.1	50.0	1.6	4.5	8.6
Qwen3-VL-2B [2]	0.5	2.5	4.0	1.8	5.5	8.4	3.6	8.7	11.9	2.0	9.6	13.9	0.8	2.9	4.9
Qwen3-VL-8B [2]	1.7	6.0	10.3	3.3	10.9	15.6	7.2	14.1	18.2	16.7	29.4	34.1	1.2	3.3	5.3
E5-V-7B [19]	2.3	8.8	12.5	2.0	6.4	10.9	3.6	7.6	10.4	9.1	15.5	20.6	-	-	-
LamRA-7B [27]	5.8	15.7	24.0	7.4	20.5	29.3	23.6	37.4	44.3	52.4	66.7	73.4	2.5	10.2	16.8
GME-7B [47]	4.3	14.3	22.2	3.7	14.5	22.1	23.5	33.3	39.9	48.8	61.1	67.1	5.3	10.7	17.2
HuatuoGPT-V-7B [7]	3.5	11.3	18.8	6.1	18.9	25.8	11.6	22.5	29.0	19.0	34.1	40.9	4.9	14.8	22.5
Lingshu-7B [43]	6.8	22.8	33.2	10.2	24.8	35.0	17.5	30.7	37.0	23.8	39.7	48.4	4.5	12.3	16.4
PathoR1-7B [46]	14.7	33.2	45.0	18.2	43.2	52.3	14.2	28.7	35.1	20.2	41.7	52.4	19.3	50.4	61.5
HOMIE (Qwen3-VL-2B)	75.2	94.2	96.7	49.8	76.0	85.9	34.4	61.9	69.8	**64.3**	**89.7**	**92.9**	22.1	59.0	76.6
HOMIE (Qwen3-VL-8B)	**79.8**	**96.7**	**98.3**	**52.7**	**79.5**	**88.5**	**35.8**	**64.1**	**72.8**	57.5	89.3	**92.9**	**30.7**	**65.6**	**79.9**

Architectural Mismatch of Dual-Encoders. Lacking internal mechanisms for multi-modal fusion, SOTA dual-encoders (e.g., MUSK [42], CONCH [28]) are restricted to parameter-free late-fusion heuristics, such as vector addition and reciprocal rank fusion (RRF). Consequently, their performance saturates at ∼44% Recall@1 on the Multi-Image to Text task.

Domain and Task Mismatches in MLLMs. Directly deploying generative MLLMs is insufficient. First, general-domain MLLM-based methods perform poorly, reflecting a Domain Mismatch with complex pathology morphology. Second, despite possessing rich medical priors, ∼7B pathology MLLMs reach only 14.7% R@1. This indicates a Task Mismatch.

Table 3: Performance comparison with baseline methods on simple retrieval datasets, reporting Recall (R@k) metrics (in percentage %) at different thresholds. Slash-separated values indicate image-to-text (i2t) / text-to-image (t2i) retrieval performance, respectively. Bold and underlined values indicate the best and second-best performance, respectively.

Model	Bookset			Educontent			MMUpubmed		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Pathology Dual-Encoder Models									
PubmedCLIP [9]	0.2/0.1	1.0/0.8	1.6/1.2	0.2/0.4	1.3/1.6	2.6/2.6	0.4/0.1	1.5/0.8	3.0/1.2
BiomedCLIP [45]	9.9/9.4	23.5/23.4	32.0/32.0	3.4/4.0	8.4/9.1	14.9/15.5	6.2/6.6	16.5/19.2	23.4/26.1
PLIP [16]	3.3/2.3	9.2/8.1	14.2/13.2	1.2/2.0	6.2/6.3	9.0/9.7	0.7/1.4	4.4/5.2	7.7/8.8
PathCLIP [38]	9.3/8.8	22.7/21.3	30.7/29.5	3.8/4.3	11.8/11.1	16.9/16.5	6.4/8.3	16.0/18.8	23.2/26.1
QuiltNet [17]	3.6/2.9	12.6/9.5	18.2/15.6	3.1/3.7	8.1/8.9	13.4/13.0	2.1/2.2	6.4/6.5	10.8/10.0
CONCH [28]	25.0/23.6	50.0/48.7	62.2/59.9	5.4/6.2	15.8/15.9	22.4/23.5	6.1/7.1	17.4/19.1	25.0/26.5
Pathgen-CLIP-L [37]	12.8/14.7	28.7/33.9	37.2/44.7	5.8/8.6	16.9/22.1	24.1/31.4	6.4/8.4	18.4/20.3	27.1/29.0
Pathgen-CLIP [37]	15.0/14.1	32.3/34.2	41.4/45.2	7.4/9.0	18.9/21.6	26.5/32.1	4.5/6.3	14.1/15.5	21.7/23.7
MUSK [42]	23.9/25.3	49.0/50.1	62.3/62.0	11.1/9.0	26.6/28.2	35.2/38.2	10.6/11.0	25.6/26.7	35.8/33.5
Patho-CLIP-L [46]	22.5/21.9	50.6/49.3	62.0/62.6	4.2/3.6	13.4/13.4	21.8/21.0	4.0/3.8	11.9/10.3	17.8/16.2
Patho-CLIP [46]	26.7/27.6	55.1/55.8	67.2/67.6	3.7/4.0	10.9/12.4	16.5/18.0	2.7/3.2	9.2/9.2	14.9/13.4
MLLM-Based Models									
Qwen2.5-VL-7B [3]	0.9/0.4	2.8/1.6	4.7/2.6	1.9/1.0	5.8/3.0	8.8/4.5	1.1/0.4	4.5/1.3	7.4/3.0
Qwen3-VL-2B [2]	0.5/0.1	1.7/0.8	2.8/1.3	0.7/0.2	2.3/1.8	3.7/2.7	0.5/0.4	1.7/1.4	3.0/2.3
Qwen3-VL-8B [2]	0.9/0.4	3.1/1.5	4.3/3.0	1.4/0.8	5.3/2.9	8.0/5.5	1.1/0.3	3.5/2.0	5.8/3.5
E5-V-7B [19]	1.7/0.5	5.5/1.8	8.6/2.4	2.6/0.7	7.7/1.6	9.7/3.2	2.3/0.8	7.8/2.6	10.7/4.5
LamRA-7B [27]	0.6/0.3	2.1/1.2	3.4/1.7	2.4/0.7	6.2/2.2	9.0/3.6	2.3/0.8	6.4/2.5	9.2/4.0
GME-7B [47]	1.9/0.5	5.1/2.3	7.8/4.1	4.1/1.5	8.8/3.7	11.7/5.5	3.3/1.4	7.4/3.8	10.5/6.0
HuatuoGPT-V-7B [7]	0.9/0.4	4.6/2.0	7.0/3.2	1.9/1.2	5.6/3.4	8.7/4.5	1.2/0.3	5.3/2.7	8.5/4.0
Lingshu-7B [43]	2.9/1.3	8.6/4.8	13.2/6.8	2.4/1.3	8.3/5.2	11.6/8.1	2.1/1.6	8.0/5.5	13.9/9.6
PathoR1-7B [46]	9.5/7.0	23.8/18.4	32.3/25.2	3.0/6.1	9.3/17.2	13.1/21.6	1.7/3.7	3.7/11.0	5.4/15.2
HOMIE (Qwen3-VL-2B)	32.0/29.7	58.5/54.7	69.4/65.1	11.1/12.7	28.5/29.7	36.5/38.3	**13.3**/11.0	**31.3**/27.1	40.1/35.5
HOMIE (Qwen3-VL-8B)	**34.4**/**31.5**	**61.4**/**60.0**	**73.0**/**72.7**	**12.1**/**14.6**	**29.8**/**31.5**	**38.0**/**40.5**	13.1/**11.1**	30.9/**28.7**	**41.4**/**37.5**

The Efficacy and Efficiency of HOMIE. By contrast, HOMIE addresses these bottlenecks. Our lightweight HOMIE (Qwen3-VL-2B) variant achieves 75.2% R@1, outperforming the best dual-encoders and exceeding PathoR1-7B [46] by over 60 percentage points. This suggests that HOMIE's gains stem from our two-stage adaptation rather than parameter scaling. Further scaling to an 8B backbone shows that HOMIE scales as a model-agnostic paradigm for composed pathology retrieval.

4.2 Zero-shot Simple Retrieval

A key question is whether gaining complex compositional reasoning capabilities compromises a model's performance on traditional simple (image-to-text and text-to-image) retrieval tasks. As shown in Table 3, HOMIE not only retains this fundamental capability but establishes a new SOTA. Both our 2B and 8B variants consistently outperform all specialized pathology dual-encoders (e.g., MUSK [42], Patho-CLIP [46]) across the Bookset [10], Educontent [36], and MMUpubmed [36] benchmarks. Furthermore, while other MLLM-based baselines continue to struggle due to inherent task and domain mismatches, HOMIE consistently outperforms them. This indicates that our two-stage adaptation framework instills fine-grained pathology knowledge into the MLLM. HOMIE achieves this performance using only public training data, without proprietary datasets.

Table 4: Performance of the HOMIE framework across diverse MLLM backbones. We report Recall@1 (%) metrics. Quilt-V. is abbreviated for Quilt-VQA.

Model	$(q^i, q^i, ...) \to c^t$	$(q^i, q^t) \to c^i$	$(q^i, q^t) \to c^t$		$q^v \to c^t$	$q^i \to c^t$		
	Bookset	Bookset	Quilt-V.	Quilt-V.-Red	Videopath	Bookset	Educontent	MMUpubmed
HOMIE (Qwen3-VL-2B)	75.2	49.8	33.4	64.3	22.1	32.0	11.1	13.3
HOMIE (Qwen3-VL-8B)	79.8	52.7	35.8	57.5	30.7	34.4	12.1	13.1
HOMIE (Qwen2.5-VL-7B)	77.2	49.2	32.5	62.3	28.3	30.0	12.2	11.7
HOMIE (HuatuoGPT-V-7B)	76.3	47.5	32.0	57.9	28.7	31.3	11.6	12.1
HOMIE (Lingshu-7B)	78.0	50.6	33.8	59.9	26.6	31.4	12.6	13.4
HOMIE (PathoR1-7B)	78.5	54.1	38.4	64.3	30.7	33.6	14.7	11.9

4.3 Generality of the HOMIE Framework

A defining characteristic of HOMIE is that it is not a monolithic model, but rather a model-agnostic, systematic adaptation framework. To assess its generality, we swapped the core MLLM backbone with various foundation models, including general-domain models (Qwen series) and SOTA pathology-specific MLLMs. As shown in Table 4, HOMIE transforms every evaluated generative MLLM into a capable pathology retrieval model. Two observations emerge from these results. First, **synergy with domain priors**: while pathology-specific MLLMs (like PathoR1-7B) perform poorly at zero-shot retrieval on their own (as shown in Table 2), applying the HOMIE framework leverages their pathology knowledge. Consequently, HOMIE (PathoR1-7B) consistently outperforms HOMIE built on the vanilla Qwen2.5-VL-7B across almost all composed retrieval metrics. Second, **framework scalability**: by comparing Qwen3-VL-2B with Qwen3-VL-8B, we observe a consistent scaling trend. As the parameter size and reasoning capabilities of the base model increase, the retrieval performance of the HOMIE-adapted model increases accordingly. These results establish HOMIE as an extensible framework. This suggests that as more capable generative MLLMs become available, the framework offers a direct pathway to adapt them for pathology retrieval.

4.4 Qualitative Results and Embedding Space Visualization

Figure 3 (left) illustrates the practical consequences of the Architectural Mismatch. In the top Image-Text to Image example, HOMIE performs the required spatial-semantic reasoning. Conversely, CONCH [28] ignores the textual instruction and erroneously retrieves the original source image, consistent with dual-encoders reducing to visual matching given interleaved inputs. Similarly, HOMIE grounds visual questions to diagnostic texts in the bottom example, whereas baselines return irrelevant descriptions. This behavior stems from our structurally aligned latent space. Figure 3 (right) visualizes the modality gap via UMAP. While pathology CLIP baselines (e.g., Pathgen-CLIP [37], MUSK [42]) show a large modality gap, segregating features into separate clusters (high $||\Delta||_{gap} \geq 0.571$), HOMIE reduces this gap. By leveraging our two-stage adaptation, HOMIE achieves the lowest modality gap ($||\Delta||_{gap} = 0.325$), mapping

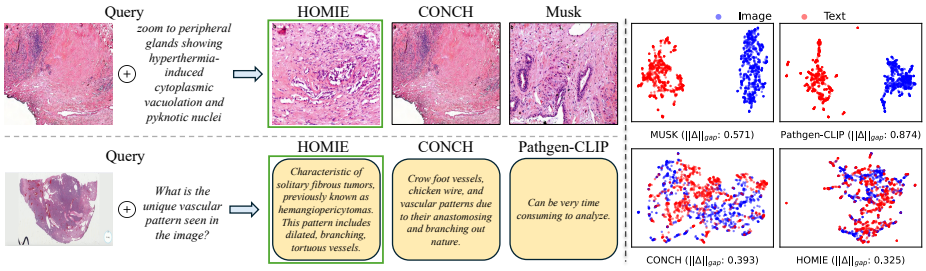

Fig. 3: Qualitative results and visualization. **Left**: Examples of composed retrieval. HOMIE performs multi-modal reasoning, whereas dual-encoder models do not. **Right**: UMAP visualization of image (red) and text (blue) embeddings from the EduContent dataset. $||\Delta||_{gap}$ denotes the modality gap metric [24].

visual and textual concepts into a shared metric space rather than adjacent regions.

4.5 Ablation Studies

We conduct an ablation study to assess the contribution of each component, rather than attributing performance to scaling or curated data alone. As shown in Table 5, we disable each pathology-specific adaptation component on the HOMIE (Qwen3-VL-8B) backbone and observe consistent performance degradation across task categories.

(1) Progressive Knowledge Curriculum: Disabling the staged training causes the largest performance drop on complex reasoning tasks. For instance, Video-to-Text retrieval drops from 30.7% to 18.4%, and Image-Text to Image retrieval drops from 52.7% to 46.7%. This supports our hypothesis: an MLLM must first build foundational morphological priors before learning high-level, multi-modal diagnostic reasoning.

(2) Native Resolution: Forcing the model to use standard, low-resolution inputs leads to notable drops on tasks requiring fine-grained visual details. For example, Image-to-Text retrieval on Bookset [10] drops by 8.9%. This indicates that fine-grained morphological features discarded by fixed-resolution models are important for pathology retrieval.

(3) Data Filtering: Training on unfiltered, noisy web data degrades performance (e.g., from 30.7% to 23.4% on Video-to-Text). This indicates that our bootstrap filtering helps construct a well-aligned embedding space.

(4) Stain Augmentation: Removing pathology-specific stain normalization and augmentation leads to consistent performance drops across multiple tasks (e.g., a 2.1% drop in Multi-Image to Text retrieval). This suggests it helps the model learn stain-invariant morphological representations. Together, these ablations indicate that each component contributes to the overall performance.

(5) Prompting Robustness: To understand the embedding extraction mechanism, we ablate our Explicit One-word Limitation (EOL) prompt at inference

Table 5: Ablation study of HOMIE's key components and prompts. We report the Recall@1 (%) metrics. All evaluated models are based on the Qwen3-VL-8B backbone. Quilt-V. is abbreviated for Quilt-VQA.

Ablations	$(q^i, q^i, ...) \to c^t$	$(q^i, q^t) \to c^i$	$(q^i, q^t) \to c^t$		$q^v \to c^t$	$q^i \to c^t$		
	Bookset	Bookset	Quilt-V.	Quilt-V.-Red	Videopath	Bookset	Educontent	MMUpubmed
Training Strategy Ablations								
w/o data filter	79.5	51.1	34.6	52.2	23.4	31.3	9.5	12.3
w/o stain N/A	77.7	52.1	35.1	57.3	28.4	32.0	11.1	12.3
w/o curriculum	76.3	46.7	30.9	52.6	18.4	29.3	10.5	**13.1**
w/o native resolution	79.5	46.3	29.0	46.2	23.8	25.5	7.3	10.6
Inference Prompt Robustness								
No Prompt	78.0	51.8	34.3	56.0	29.1	32.0	11.1	12.9
No One Word	78.5	51.5	34.8	56.3	30.0	33.8	11.5	12.6
HOMIE	**79.8**	**52.7**	**35.8**	**57.5**	**30.7**	**34.4**	**12.1**	**13.1**

time. Extracting the last hidden state without explicit instructional constraints (*No Prompt*) or with a weaker prompt (*No One Word*) still yields competitive results. This robustness suggests that training adapts the model to condense multi-modal semantics into the final token, reducing reliance on the inference-time prompt. Applying the EOL constraint (*HOMIE*) further aligns the inference behavior with the trained representation. The phrase "in one word" acts as a semantic bottleneck that reduces residual generation noise and yields the best performance (79.8%).

5 Conclusion

In this paper, we formally define the task of Pathology Composed Retrieval (PCR) and introduce the first comprehensive benchmark to evaluate interleaved multi-modal queries. To overcome the Architectural mismatch of dual-encoders and the Task/Domain mismatches of generative models, we propose HOMIE—a model-agnostic, systematic adaptation framework. HOMIE transforms general Multimodal Large Language Models (MLLMs) into specialized pathology retrieval models. Ablations indicate that the two-stage design accounts for the observed performance. Stage 1 aligns the generative latent space for retrieval, while Stage 2 instills morphological priors through dynamic native-resolution processing, stain augmentation, and a progressive knowledge curriculum. Our lightweight 2B-parameter HOMIE variant substantially outperforms existing paradigms on composed retrieval, while maintaining strong performance on traditional tasks. By generating a single, unified multi-modal embedding for complex clinical queries, HOMIE provides a reliable retrieval system that empowers pathologists to find similar information and make their own independent, evidence-based diagnoses. Future work will extend this framework to integrate broader modalities, such as genomics and spatial transcriptomics, moving toward a clinical AI assistant for pathology.

Acknowledgements

This work was partially supported by US National Science Foundation IIS-2412195, CCF-2400785, the Cancer Prevention and Research Institute of Texas (CPRIT) award (RP230363), the National Institutes of Health (NIH) R01 award (1R01AI190103-01) and Microsoft Accelerate Foundation Models Research (2024).

References

1. Alfasly, S., Alabtah, G., Hemati, S., Kalari, K.R., Garcia, J.J., Tizhoosh, H.: Validation of histopathology foundation models through whole slide image retrieval. Scientific Reports **15**(1), 3990 (2025)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Bera, K., Schalper, K.A., Rimm, D.L., Velcheti, V., Madabhushi, A.: Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. Nature reviews Clinical oncology **16**(11), 703–715 (2019)
5. Cai, C.J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., Wattenberg, M., Viegas, F., Corrado, G.S., Stumpe, M.C., et al.: Human-centered tools for coping with imperfect algorithms during medical decision-making. In: Proceedings of the 2019 chi conference on human factors in computing systems. pp. 1–14 (2019)
6. Chen, C., Lu, M.Y., Williamson, D.F., Chen, T.Y., Schaumberg, A.J., Mahmood, F.: Fast and scalable search of whole-slide images via self-supervised deep learning. Nature Biomedical Engineering **6**(12), 1420–1434 (2022)
7. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Cai, Z., Ji, K., Wan, X., et al.: Towards injecting medical visual knowledge into multimodal llms at scale. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 7346–7370 (2024)
8. Dao, T.: Flashattention-2: Faster attention with better parallelism and work partitioning. arXiv preprint arXiv:2307.08691 (2023)
9. Eslami, S., Meinel, C., De Melo, G.: Pubmedclip: How much does clip benefit visual question answering in the medical domain? In: Findings of the Association for Computational Linguistics: EACL 2023. pp. 1181–1193 (2023)
10. Gamper, J., Rajpoot, N.: Multiple instance captioning: Learning representations from histopathology textbooks and articles. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16549–16559 (2021)
11. Gao, T., Yao, X., Chen, D.: Simcse: Simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 6894–6910 (2021)
12. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)
13. Hu, D., Jiang, Z., Shi, J., Xie, F., Wu, K., Tang, K., Cao, M., Huai, J., Zheng, Y.: Histopathology language-image representation learning for fine-grained digital pathology cross-modal retrieval. Medical Image Analysis **95**, 103163 (2024)

14. Huang, J., Liu, X., Song, S., Hou, R., Chang, H., Lin, J., Bai, S.: Revisiting multimodal positional encoding in vision-language models. arXiv preprint arXiv:2510.23095 (2025)
15. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems **43**(2), 1–55 (2025)
16. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)
17. Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1m: One million image-text pairs for histopathology. Advances in neural information processing systems **36** (2024)
18. Jiang, T., Huang, S., Luan, Z., Wang, D., Zhuang, F.: Scaling sentence embeddings with large language models. In: Findings of the association for computational linguistics: EMNLP 2024. pp. 3182–3196 (2024)
19. Jiang, T., Song, M., Zhang, Z., Huang, H., Deng, W., Sun, F., Zhang, Q., Wang, D., Zhuang, F.: E5-v: Universal embeddings with multimodal large language models. arXiv preprint arXiv:2407.12580 (2024)
20. Jiang, Z., Meng, R., Yang, X., Yavuz, S., Zhou, Y., Chen, W.: Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In: The Thirteenth International Conference on Learning Representations
21. Kalra, S., Tizhoosh, H.R., Choi, C., Shah, S., Diamandis, P., Campbell, C.J., Pantanowitz, L.: Yottixel–an image search engine for large archives of histopathology whole slide images. Medical Image Analysis **65**, 101757 (2020)
22. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895 (2024)
23. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
24. Liang, V.W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.Y.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. Advances in Neural Information Processing Systems **35**, 17612–17625 (2022)
25. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26296–26306 (2024)
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
27. Liu, Y., Zhang, Y., Cai, J., Jiang, X., Hu, Y., Yao, J., Wang, Y., Xie, W.: Lamra: Large multimodal model as your advanced retrieval assistant. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4015–4025 (2025)
28. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**(3), 863–874 (2024)
29. Pal, A., Umapathi, L.K., Sankarasubbu, M.: Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In: Conference on health, inference, and learning. pp. 248–260. PMLR (2022)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from

natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
31. Romanov, A., Shivade, C.: Lessons from natural language inference in the clinical domain. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 1586–1596 (2018)
32. Seyfioglu, M.S., Ikezogwo, W.O., Ghezloo, F., Krishna, R., Shapiro, L.: Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13183–13192 (2024)
33. Shen, Y., Luo, Y., Shen, D., Ke, J.: Randstainna: Learning stain-agnostic features from histology slides by bridging stain augmentation and normalization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 212–221. Springer (2022)
34. Shmatko, A., Ghaffari Laleh, N., Gerstung, M., Kather, J.N.: Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. Nature cancer **3**(9), 1026–1038 (2022)
35. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
36. Sun, Y., Wu, H., Zhu, C., Zheng, S., Chen, Q., Zhang, K., Zhang, Y., Wan, D., Lan, X., Zheng, M., et al.: Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In: European Conference on Computer Vision. pp. 56–73. Springer (2024)
37. Sun, Y., Zhang, Y., Si, Y., Zhu, C., Zhang, K., Shui, Z., Li, J., Gong, X., LYU, X., Lin, T., et al.: Pathgen-1.6 m: 1.6 million pathology image-text pairs generation through multi-agent collaboration. In: The Thirteenth International Conference on Learning Representations (2025)
38. Sun, Y., Zhu, C., Zheng, S., Zhang, K., Sun, L., Shui, Z., Zhang, Y., Li, H., Yang, L.: Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5034–5042 (2024)
39. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
40. Vuong, T.T., Kwak, J.T.: Videopath-llava: Pathology diagnostic reasoning through video instruction tuning. arXiv preprint arXiv:2505.04192 (2025)
41. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pretraining for vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19175–19186 (2023)
42. Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., et al.: A vision–language foundation model for precision oncology. Nature **638**(8051), 769–778 (2025)
43. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
44. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)

45. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
46. Zhang, W., Zhang, P., Guo, J., Cheng, T., Chen, J., Zhang, S., Zhang, Z., Yi, Y., Bu, H.: Patho-r1: A multimodal reinforcement learning-based pathology expert reasoner. arXiv preprint arXiv:2505.11404 (2025)
47. Zhang, X., Zhang, Y., Xie, W., Li, M., Dai, Z., Long, D., Xie, P., Zhang, M., Li, W., Zhang, M.: Gme: Improving universal multimodal retrieval by multimodal llms. arXiv preprint arXiv:2412.16855 (2024)