PathoHR: Hierarchical Reasoning for Vision-Language Models in Pathology

Yating Huang[1]*, Ziyan Huang[2]*, Lintao Xiang[1], Qijun Yang[1], Hujun Yin[1]
[1]University of Manchester
[2]South China University of Technology
{yating.huang, hujun.yin}@manchester.ac.uk, bonnie.ziyan.huang@gmail.com

Abstract

Accurate analysis of pathological images is essential for automated tumor diagnosis but remains challenging due to high structural similarity and subtle morphological variations in tissue images. Current vision-language (VL) models often struggle to capture the complex reasoning required for interpreting structured pathological reports. To address these limitations, we propose PathoHR-Bench, a novel benchmark designed to evaluate VL models' abilities in hierarchical semantic understanding and compositional reasoning within the pathology domain. Results of this benchmark reveal that existing VL models fail to effectively model intricate cross-modal relationships, hence limiting their applicability in clinical setting. To overcome this, we further introduce a pathology-specific VL training scheme that generates enhanced and perturbed samples for multimodal contrastive learning. Experimental evaluations demonstrate that our approach achieves state-of-the-art performance on PathoHR-Bench and six additional pathology datasets, highlighting its effectiveness in fine-grained pathology representation.

1 Introduction

Recently, vision-language (VL) models have gained substantial advances (Goel et al., 2022; Li et al., 2022), fostering an integration of computer vision and natural language processing across a wide range of application domains (Wei et al., 2024; Das et al., 2024). In medical image analysis, contrastive learning-based approaches have been extensively employed to align large-scale medical images with corresponding diagnostic reports, thereby enabling zero-shot classification and grading without additional fine-tuning (Xie et al., 2024; Phan et al., 2024). This paradigm markedly reduces the reliance on high-quality annotated data and enhances the scalability of medical imaging models. Nevertheless, in contrast to other imaging modalities such as X-ray and MRI, pathological image analysis presents distinct and more formidable challenges due to its higher structural complexity and more subtle visual cues in the images.

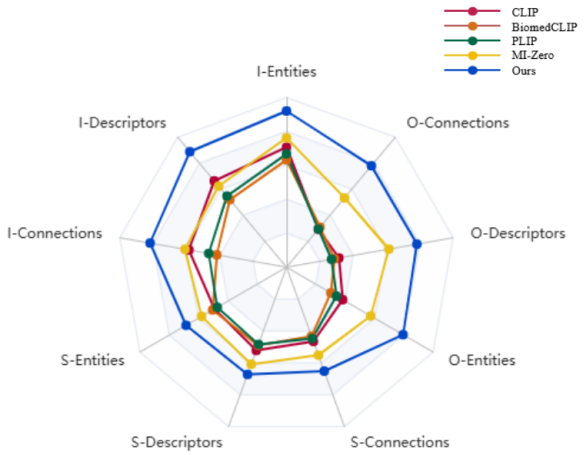


Figure 1: Radar charts for compared models on PathoHR-Bench across multiple compositional reasoning aspects. The axes correspond to three types of perturbations: I (Information Loss), S (Semantic Drift), O (Order Variation), evaluated under three semantic roles: Entities, Descriptors and Connections. Higher values indicate stronger robustness.

Pathological biopsy microscopy remain the gold standard for tumor diagnosis, as it capture intricate cellular structures and morphological details (Ruiz López et al., 2017). However, the subtle intra-class variations in images pose significant challenges for disease classification. Existing approaches predominantly rely on task-specific models tailored to individual applications, such as Gleason grading of prostate cancer and breast cancer subtyping (Pati et al., 2023; Bulten et al., 2022). In contrast, pathological texts are inherently fine-grained, characterized by specialized medical terminology, precise pathological reasoning, and structured diagnostic logic. These narratives not

*Equal contribution.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 2296–2311
November 4-9, 2025

only describe subtle visual cues but also offer critical insights for diagnostic grading and prognosis prediction.

Although recent studies have explored VL pretraining and zero-shot transfer in pathology (Javed et al., 2024; Lu et al., 2023a,b), most existing models still treat pathological texts as a "bag-of-words," ignoring their hierarchical structure, syntactic dependencies, and diagnostic reasoning logic. Previous research (Yuksekgonul et al., 2022) reveals that standard VL benchmarks often inadequately evaluate a models' capacity for compositional and structural semantic understanding. This shortcoming is particularly critical in pathology, where diagnostic texts require more than surface-level pattern matching. Accurate and interpretable decision-making in pathology requires effective modeling of hierarchical clinical language in conjunction with visual and morphological cues. Current models and evaluation methods face several limitations:

i). Limited hierarchical structure awareness: Pathological tests typically follow a structured diagnostic pattern (Bera et al., 2019), (e.g., lesion region + cellular morphology + symptoms identification + diagnostic conclusion.). However, most VL models treat texts as unordered token sets, lacking the capacity to model multi-level semantic dependencies or structured reasoning patterns. Although these models may perform adequately on retrieval or classification tasks, their disregard for compositional structure undermines their interpretability and alignment with clinical diagnostic logic.

ii). Limited compositional reasoning capability: Pathology texts contain complex reasoning that integrates anatomical regions, cellular structures, and molecular markers, often within intricate spatial and functional contexts. Bag-of-words-based models would fail to capture such contextual and inferential relationships, thereby hindering accurate and nuanced diagnostic decision-making.

Based on the above analysis, the main contributions of this work are summarized as follows:

- We introduce PathoHR-Bench, a novel benchmark designed to evaluate the hierarchical structure awareness and compositional reasoning capabilities of pathological VL models. Unlike existing evaluation protocols that focus on application performance, PathoHR-Bench offers a systematic assessment of model robustness under three core perturbation types that reflect key challenges in pathology text comprehension. To better reflect the structured nature of diagnostic narratives, the benchmark further categorizes evaluations according to distinct semantic roles commonly observed in pathology reports.
- We subsequently propose a data-driven VL training scheme tailored to the pathology domain, aimed at more effectively leveraging existing pathological VL datasets. The approach generates both enhanced and perturbed samples across textual and visual modalities, and employs cross-modal contrastive learning to improve the model's ability to capture fine-grained semantic alignments and diagnostic reasoning cues.

As illustrated in Figure 1, the proposed training scheme yields consistent performance gains across all perturbation types and hierarchical components on the PathoHR-Bench, highlighting its effectiveness in capturing multi-level semantic dependencies and inferring complex pathological relationships. Moreover, in standard zero-shot diagnostic classification tasks, we observe that improved structural understanding and compositional reasoning directly enhance diagnostic accuracy, underscoring the critical importance of these capabilities for real-world clinical deployment. The full benchmark suite and source code will be publicly released upon acceptance.

2 Motivations

The lack of hierarchical reasoning capabilities significantly limits the reliability and clinical interpretability of existing VL models in pathological analysis. A further limitation lies in current evaluation practices, which predominantly rely on standard retrieval or classification metrics and offer limited insight into models' structural understanding or reasoning ability. While recent studies in the natural image domain have proposed benchmarks for evaluating compositional reasoning in VL models (Zhao et al., 2022b; Diwan et al., 2022), these approaches are not directly applicable to pathology due to fundamental domain differences in textual structure and semantics. Descriptions in natural image datasets are typically open-domain and emphasize visually salient features. In contrast, pathological texts are "closed-domain," relying on highly specialized and standardized vocabularies (e.g., MeSH, ICD-O) and structured diagnostic patterns. Unlike open-domain free text, these reports

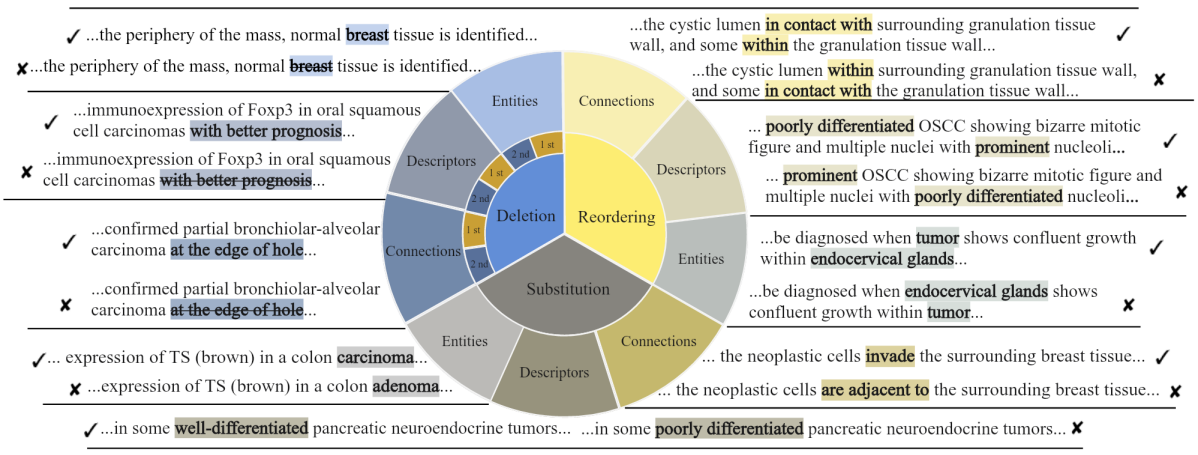


Figure 2: Cross-dimensional taxonomy in PathoHR-Bench: Text perturbation levels and semantic role levels.

follow controlled terminology and precise reasoning logic, resulting in much tighter semantic relationships and long-range inferential chains.

For instance, in describing a gastric cancer pathological image: "Disordered glandular arrangement with gland fusion and lumen disappearance, consistent with poorly differentiated adenocarcinoma." Here, the reasoning between "gland fusion" and "poorly differentiated" determines the malignancy grade of the lesion (Grillo et al., 2020), making this logical relationship crucial for both model learning and final diagnostic decision-making. In addition, pathological texts often contain a significant amount of "implicit reasoning". This type of inference typically requires integration with external medical knowledge bases (Jha et al., 2017) rather than solely relying on visual features for text generation. These characteristics underscore the need for a domain-specific benchmark tailored to pathology. In this context, PathoHR-Bench is designed to fill this gap by providing a structured and diagnostic-relevant framework for evaluating the compositional and inferential reasoning capabilities of VL models in the pathological domain.

3 Pathology Hierarchical Reasoning Benchmark (PathoHR-Bench)

Inspired by VL-CheckList (Zhao et al., 2022a) and ARO (Yuksekgonul et al., 2023) that utilize generated adversarial samples and image-text matching as primary evaluation objectives, we propose a novel pathology VL benchmark, termed as PathoHR-Bench, which is designed to systematically assess pathology-related VL models on hierarchical structure awareness and compositional reasoning capabilities. As shown in Figure 2, it adopts a cross-dimensional structure, enabling a comprehensive assessment of VL models across two independent yet interrelated dimensions: text perturbation and semantic role. This design allows for an in-depth analysis of model robustness, fine-grained comprehension, and compositional reasoning when processing pathological VL data. By identifying model limitations more effectively, PathoHR-Bench can serve as a valuable resource for advancing VL model development and optimization for pathology applications. Details of the PathoHR-Bench are illustrated in Figure 3.

At the text perturbation level, we have designed three core tasks to simulate the challenges encountered in real-world pathology texts, including information loss, semantic drift, and order variation:

i). Assessing sensitivity to information loss: We employ saliency-driven phrase deletion by leveraging the Pathology Language Image Pretraining (PLIP) (Huang et al., 2023) model to compute the similarity between each phrase and the corresponding image. The two most salient pathological terms are removed to create adversarial text variants, simulating the model's reliance on critical textual cues and assessing its behavior under different deletion orders. It reflects how the model allocates information between visual and textual inputs while evaluating its robustness to local information losses.

ii). Assessing sensitivity to semantic drift: We employ LLM-guided token substitution, where a randomly selected word in each text is masked and then predicted using a pretrained medical BERT model (BioBERT) (Kenton and Toutanova, 2019; Lee et al., 2020). The generated adversarial sam-

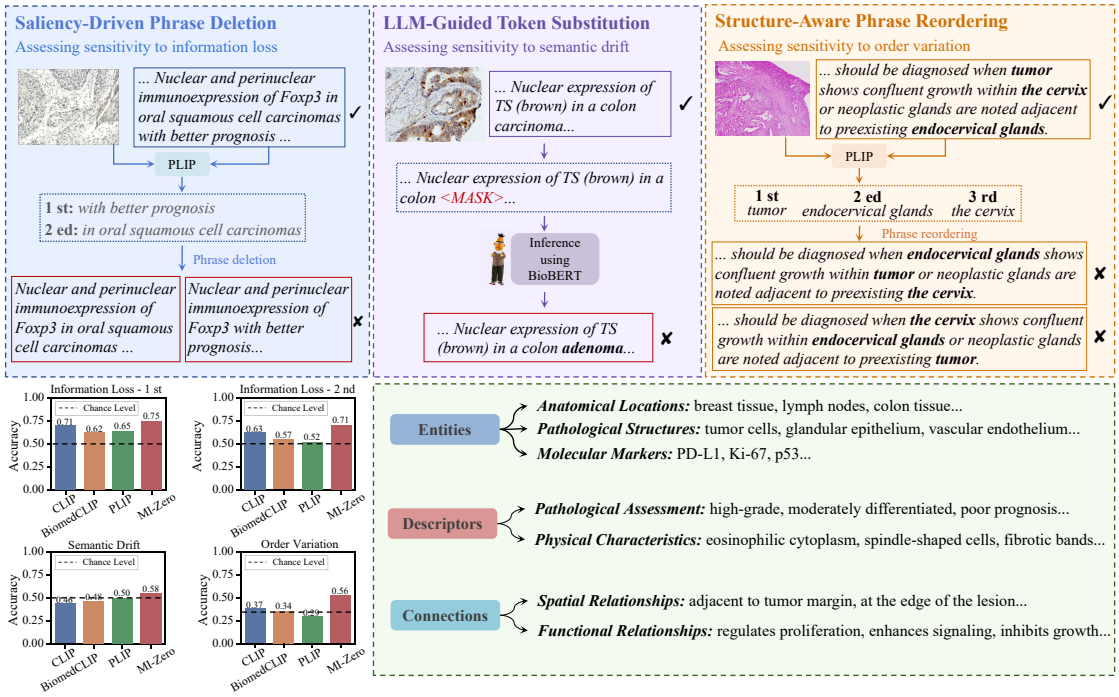


Figure 3: Overview of proposed PathoHR-Bench, comprising three sensitivity tests (top row) with performances of existing VL models (bottom left). Bottom right shows further semantic perturbation levels.

ples test the model's robustness to semantic drifts and fine-grained semantic comprehension. It evaluates whether the VL models can accurately differentiate between subtly different but diagnostically significant textual descriptions.

iii). Assessing sensitivity to order variation: We employ structure-aware phrase reordering to examine the model's sensitivity to order variation. Specifically, we use PLIP (Huang et al., 2023) to identify the top three phrases with the highest image-text similarity and cyclically reorder them within the sentence to generate two adversarial variants. This strategy assesses whether the VL model responds to sentence structure changes, offering insights into its structural awareness.

As shown in the bottom-left of Figure 3, current pathological VL models perform poorly on perturbation-level evaluation tasks, often at or below random chance which highlighting a lack of structural awareness and fine-grained compositional reasoning abilities. Although VL models can leverage shortcut strategies (Geirhos et al., 2020) during contrastive pretraining to excel at coarse-grained classification and matching, such heuristics are insufficient for pathology-specific tasks. The image-text relationships in pathology are more complex than in general vision tasks, requiring hierarchical semantic understanding and the modeling of nuanced diagnostic relations.

To further investigate these limitations, we extend the perturbation-level evaluation by introducing three major semantic roles derived from the structural patterns of pathological texts.

i). Entities: Including anatomical locations, pathological structures, and molecular markers, which form the foundation of pathology reports. VL models must accurately recognize these entities to achieve precise image-text alignment.

ii). Descriptors: Including terms that modify pathological entities, such as diagnostic assessment terms and physical characteristics. These descriptors are crucial for precise disease classification.

iii). Connections: This category consists of textual elements describing spatial relationships and functional interactions between entities. Such information facilitates complex pathological reasoning and disease progression analysis.

Examples of these semantic role classifications are illustrated in Figure 3. By combining the semantic role hierarchy with text perturbation tasks, we construct a systematic, cross-dimensional evaluation framework. The key advantage of this framework is its ability to assess VL models not only for robustness under various text transformations but also for their understanding of different pathology-related semantic concepts. This allows for a more

precise identification of model weaknesses and developing effective optimisation strategies. A detailed evaluation of various models on PathoHR-Bench is presented in Section 5.

4 Pathology Hierarchical Reasoning Training Scheme

We propose a data-driven training scheme for pathological VL model, shown in Figure 4. It consists of four branches for generating positive and negative samples for both text and imagery. The following subsections discuss these four branches and the losses incorporated into VL contrastive learning.

4.1 Pathology-guided textual perturbation

A simple but effective approach of negative text generation is to apply a collection of pre-defined linguistic rules to match and replace words associated with specific entity types or patterns (Doveh et al., 2023). We pre-defined seven pathological attribute dimensions (pathological states and grading, morphological features, histochemical characteristics, staining methods, anatomical structures and organs, biomolecular features, color information) via PubMed (Namata et al., 2012) and MeSH (Lipscomb, 2000), followed by random substitution at the level of pathological attributes to generate controlled perturbations.

To account for the specificity of pathological language, we further introduce a refinement stage using BioGPT (Luo et al., 2022), which ensures that the generated negative texts retain pathological plausibility and clinical relevance. This pathology-guided textual perturbation applies controlled semantic transformations and plausibility correction, enabling VL models to learn subtle yet critical differences in pathological semantics.

4.2 Hierarchical diagnostic reasoning-based text expansion

This branch provides multi-level structured diagnostic insights, guiding the model to engage in hierarchical compositional reasoning and allowing it to grasp diagnostic logic rather than relying on superficial text-image correlations. Traditional VL models tend to rely heavily on shallow semantic matching, ignoring the reasoning process and hierarchical context of diagnosis. To address this, we employed GPT-4 (Achiam et al., 2023) to generate structured positive texts across four pathological analysis perspectives: pathological description, causes analysis, symptoms identification, and diagnostic basis. This approach enables VL models to learn from multi-perspective and multi-level diagnostic information, enhancing their compositional reasoning in medical knowledge. By incorporating hierarchical diagnostic reasoning, the model shifts from data-driven medical text generation to more explainable diagnostic reasoning, improving its contextual reasoning capabilities.

4.3 Dual-constraint negative image mining

To enhance the fine-grained structural reasoning ability of VL models in pathology, we propose a dual-constraint negative sample mining strategy. This approach generates negative images under two complementary mechanisms: semantic inconsistency and distributional ambiguity, yielding both easy and hard negatives.

i. Text-guided visual editing: To simulate explicit semantic inconsistencies, we employ text-guided image editing using Stable Diffusion (Rombach et al., 2022) to generate easy negative images conditioned on deliberately corrupted diagnostic texts. These generated samples are not expected to conform to natural image distributions. Instead, they serve as visual exaggerations of erroneous semantics, providing strongly misaligned image-text pairs. The goal is not to simulate realistic pathology, but to reinforce the model's ability to identify and reject pathology-irrelevant or structurally invalid cues. This mechanism enhances fine-grained semantic-to-structural reasoning by anchoring the model's attention to diagnostically critical patterns.

ii. Adversarial distribution-aware perturbation: In parallel, we introduce an adversarial distribution-aware sampling strategy that generates hard negatives without relying on textual input. Operating purely in the visual domain, this module seeks perturbations near the real distribution of pathological images that would cause the most confusion to the model. We consider the worst-case scenario (Qiao et al., 2020) around the original distribution M_0:

$$\min_{\theta} \sup_{M_N : D(M_N, M_0) \leq m} E_{M_N} \left[L\left(\theta, I\right) \right] \quad (1)$$

where θ denotes the pretrained model weights, I is the original image, and L represents the task-specific loss. D denotes the distance metric, and m is the maximum distributional variability between M_0 and the generated distribution M_N.

The solution to worst-case problem aims to generate negative samples that remain visually and

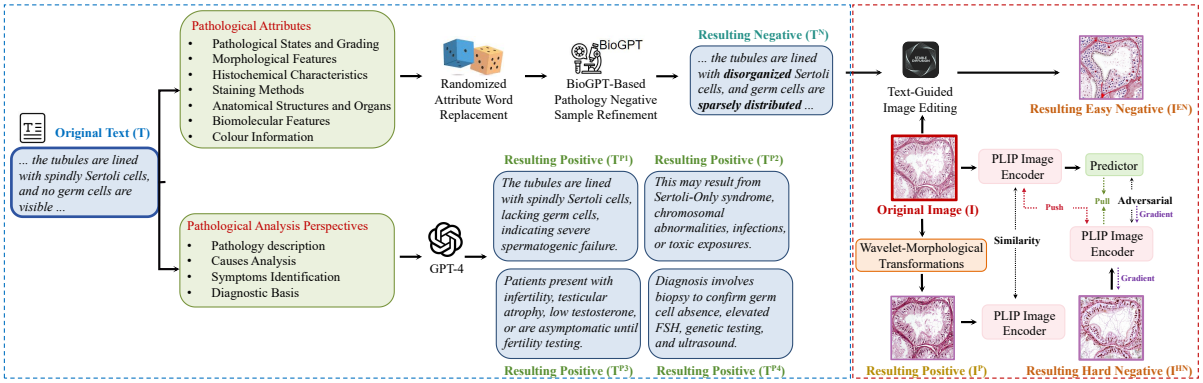


Figure 4: Structured textual and visual data manipulation for pathological VL training.

statistically plausible within the pathology domain, while being maximally separated from the original distribution under the constraint. These samples serve as structurally valid but distributionally challenging examples, exposing the model to difficult decision-boundary cases without deviating from the natural appearance of real pathological images. Detailed formulation and optimization procedures are provided in the Appendix A.

4.4 Wavelet-morphology-guided consistency refinement

This branch enhances the multi-scale representation of pathology images while preserving semantic consistency of tissue structures. A wavelet transform (Othman and Zeebaree, 2020) decomposes each image into frequency components, with high-frequency bands capturing microstructural details. Morphological operations are then applied to these components: Top-Hat enhances bright structures (e.g., nuclei, fibers), Black-Hat emphasizes low-density regions (e.g., necrosis, inflammation), and Morphological Gradient highlights cellular boundaries. The enhanced image is reconstructed via inverse wavelet transform. To ensure semantic alignment with the original image, we apply a consistency constraint based on feature similarity computed using a pretrained PLIP model. A similarity threshold dynamically adjusts transformation parameters, generating structurally enriched positive samples that preserve pathological integrity.

4.5 Loss Function

In our validation experiments, the dual-encoder vision-language models (e.g., CLIP (Radford et al., 2021), PLIP (Huang et al., 2023)) admit text-image pairs (T, I) through a text encoder $f_T(T)$ and an image encoder $f_I(I)$. The text-image similarity score is then computed as:

$$S(T, I) = exp\left(\frac{\alpha f_T(T)^T f_I(I)}{\|f_T(T)\|^2 \|f_I(I)\|^2}\right) \quad (2)$$

where α is a learned temperature parameter.

Contrastive Loss. Similar to most contemporary VL models, we employ the contrastive CLIP-loss as one of the losses for each batch $B = \{(T_i, I_i,)\}$:

$$L_{con} = \sum_i \log\left(\frac{S(T_i,I_i)}{\sum_j S(T_i,I_j)}\right) + \log\left(\frac{S(T_i,I_i)}{\sum_k S(T_k,I_i)}\right) \quad (3)$$

Negative Loss. For the negative text T_i^N generated in Section 4.1, we focus on similarity difference between the original text T_i and the generated T_i^N with respect to their corresponding original image I_i, leading to a negative text loss:

$$L_{neg}^{text} = \sum_i -\log\left(\frac{S(T_i,I_i)}{S(T_i,I_i)+S\left(T_i^N,I_i\right)}\right) \quad (4)$$

Similarly, for the negative images I_i^{EN} and I_i^{HN} generated in Section 4.3, we obtain negative image loss $L_{neg}^{img_E}$ and $L_{neg}^{img_N}$ similar to Equation 4. Then the total negative loss L_{neg} is obtained:

$$L_{neg} = L_{neg}^{text} + L_{neg}^{img_E} + L_{neg}^{img_N} \quad (5)$$

Positive Loss. For the positive text, to ensure that the four hierarchical levels of positive text samples $\left(T_i^{P^k}, k = 1, 2, 3, 4\right)$ generated in Section 4.2 are closely aligned in the feature space and matching with the original text T_i, we compute the cosine similarity between their text embeddings using $S(T_1, T_2)$ and define the text-text positive loss as:

$$L_{pos}^{text_T} = \sum_i -\log\left(\frac{\sum_{k=1}^4 S\left(T_i^{P_k}, T_i\right)}{\sum_j \sum_{k=1}^4 S\left(T_i^{P_k}, T_j\right)}\right) \quad (6)$$

To ensure that all four hierarchical levels of positive text samples align with the original image I_i and preserve cross-modal consistency, we introduce a text-image positive loss:

$$L_{pos}^{text_I} = \sum_i -\log\left(\frac{\sum_{k=1}^{4} S\left(T_i^{P_k}, I_i\right)}{\sum_j \sum_{k=1}^{4} S\left(T_i^{P_k}, I_j\right)}\right) \quad (7)$$

For the positive image loss, we omit the image-image positive term, as the generation process described in Section 4.4 ensures feature consistency with the original image. Thus, our goal is to ensure that the positive images I_i^P remain closely aligned with their corresponding text descriptions:

$$L_{pos}^{img} = \sum_i -\log\left(\frac{S\left(T_i, I_i^P\right)}{\sum_j S\left(T_j, I_i^P\right)}\right) \quad (8)$$

The total positive loss L_{pos} is computed as:

$$L_{pos} = L_{pos}^{text_T} + L_{pos}^{text_I} + L_{pos}^{img} \quad (9)$$

Finally, the full fine-tuning loss of our proposed method can be written as:

$$L = L_{con} + \alpha \cdot L_{neg} + \beta \cdot L_{pos} \quad (10)$$

5 Experiments

5.1 Datasets

The ARCH dataset (Gamper and Rajpoot, 2021) is the only widely available pathology-specific paired image–text dataset, comprising 8,617 pairs extracted from pathology textbooks and PubMed research articles. To avoid overlap between training and evaluation, we carefully split ARCH into two disjoint subsets: textbook-derived samples were used to construct PathoHR-Bench, while PubMed-derived samples were reserved for training. Building on this resource, PathoHR-Bench was created by systematically applying three types of perturbations across three semantic roles. This expansion yielded a total of 77,553 image–text pairs for evaluation. The textbook portion provides more structured and hierarchical narratives, making it especially suitable for robust assessment of reasoning capabilities, while the PubMed portion ensures fair training for all baseline models.

We compared the performance of our proposed method with current pathological VL models, including baseline CLIP (Radford et al., 2021), PLIP (Huang et al., 2023), BiomedCLIP (Zhang et al., 2023), CONCH (Lu et al., 2024), QuiltNet (Ikezogwo et al., 2023), and MI-Zero (Lu et al., 2023c) on PathoHR-Bench. Inspired by the fair comparison strategy in CPLIP (Javed et al., 2024), we fine-tuned all baseline models on the ARCH dataset before evaluation, with the exception of CLIP. Notably, while PLIP was originally pre-trained on PubMed captions and various biomedical image–text pairs, it had not been fine-tuned on ARCH in its released version. To ensure consistent data distribution and minimize domain bias, we additionally fine-tuned PLIP and the other baselines on ARCH using their original loss objectives, without introducing our perturbation-based augmentations. CLIP remained a purely zero-shot baseline without fine-tuning on ARCH, serving as a general-purpose reference point. All models were then evaluated under the same testing splits and inference prompts for fair comparison.

For the zero-shot task, we utilized six publicly available datasets covering a range of cancer types, including four datasets at the patch level and two at the whole-slide level. To ensure fairness and consistency across all methods, we adopted a single prompt per class for evaluation, rather than using prompt ensembling. Detailed dataset descriptions are provided in the Appendix B, and implementation details can be found in the Appendix C. To disentangle the effect of ARCH-specific fine-tuning from the inherent representation ability of existing models under our unified single-prompt protocol, we additionally report the performance of BiomedCLIP and PLIP *without* any fine-tuning on ARCH in Appendix D.

5.2 Comparison of existing VL models

To comprehensively evaluate the current pathological VL models, we conducted a detailed assessment using the benchmark proposed in Section 3. Results are shown in Table 1. The results reveal an unexpected fact that some existing VL models that were trained specifically on pathological datasets performed worse in structural awareness and compositional reasoning tests for pathological texts compared to the CLIP baseline trained on natural images.

This discrepancy may be attributed to the limited scale of pathological datasets and the rich and diverse open-world concepts in natural image-text pairs, which may inherently promote compositional reasoning capability. However, CLIP performed poorly in pathology-specific tasks, demonstrated

Model	Information Loss-1st			Information Loss-2nd			Semantic Drift			Order Variation		
	Entities	Descriptors	Connections	Entities	Descriptors	Connections	Entities	Descriptors	Connections	Entities	Descriptors	Connections
CLIP	0.7072	0.6643	0.5851	0.6431	0.6129	0.5237	0.4915	0.5220	0.4655	0.3831	0.3026	0.3409
BiomedCLIP	0.6325	0.5201	0.4138	0.5924	0.5088	0.4152	0.5037	0.4904	0.4326	0.3142	0.2874	0.2703
PLIP	0.6679	0.5472	0.4664	0.5210	0.4936	0.4427	0.4720	0.4843	0.4473	0.2926	0.3079	0.2931
CONCH	0.6531	0.5496	0.4389	0.6031	0.5142	0.4195	0.5028	0.5288	0.4736	0.3574	0.4397	0.4052
QuiltNet	0.7354	0.6992	0.7158	0.6985	0.6079	0.6423	0.5529	0.5706	0.5230	0.4572	0.5831	0.4590
MI-Zero	0.7623	0.6236	0.6089	0.7114	0.6033	0.5796	0.5789	0.6083	0.5519	0.5748	0.6139	0.5327
Ours	**0.9134**	**0.8780**	**0.8205**	**0.8901**	**0.8427**	**0.7945**	**0.6853**	**0.6744**	**0.6520**	**0.7932**	**0.7819**	**0.7813**

Table 1: Performance comparison across different compositional reasoning aspects on PathoHR-Bench. Best results are highlighted in **bold** and second best in underline.

Model	CRC100K		UHU		PanNuke		DigestPath		TCGA-BRCA		TCGA-RCC	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
CLIP	0.2593	0.1971	0.3349	0.1836	0.3220	0.3517	0.2056	0.1432	0.5021	0.3490	0.3348	0.1956
BiomedCLIP	0.4461	0.3549	0.3538	0.2281	0.5632	0.5926	0.6149	0.5974	0.5164	0.4209	0.6533	0.6273
PLIP	0.5361	0.4603	0.3735	0.2538	0.6275	0.6438	0.7955	0.8031	0.4508	0.3502	0.6987	0.7055
CONCH	0.5482	0.5219	0.3574	0.2306	0.5074	0.5128	0.8386	0.8470	0.6327	0.6152	0.7899	0.7740
QuiltNet	0.5097	0.4390	0.3413	0.2451	0.6377	0.5932	0.8059	0.7814	0.7235	0.6943	0.8053	0.7905
MI-Zero	0.5637	0.5626	0.3492	0.2870	0.6527	0.6779	0.8253	0.8214	0.7729	0.7006	0.7986	0.7260
Ours	**0.6985**	**0.6841**	**0.4937**	**0.4612**	**0.7386**	**0.7609**	**0.8517**	**0.8638**	**0.8129**	**0.7746**	**0.8327**	**0.7996**

Table 2: Zero-shot classification performance using a single prompt per class, reported as balanced accuracy and weighted F1 across **six** datasets. Best and second results are highlighted with **bold** and underline.

by near-zero effectiveness as shown in Table 2, as expected. These results highlight a fundamental limitation of current pathological VL models that domain adaptation may help improve zero-shot performance in pathological classification tasks, it remains insufficient to address the complexity of pathological reasoning.

In contrast, our proposed method effectively leverages limited pathology-specific data to enable fine-grained compositional reasoning while maintaining strong performance on conventional classification tasks. As shown in Tables 1 and 2, the model achieves consistent improvements across nine textual perturbation settings and six zero-shot classification benchmarks. Notably, substantial gains are observed on CRC100K and UHU (with fine-grained cancer subtypes), while more moderate gains are seen on PanNuke and DigestPath (with coarser tumor/normal classification). These results further support PathoHR-Bench as a robust benchmark for evaluating a VL model's ability to capture fine-grained pathological semantics. The case studies provided in the appendix E further illustrate the model's performance on fine-grained diagnosis.

5.3 Ablation studies

We performed a series of ablation studies to validate the effectiveness of the proposed key components in Section 4, including negative text (Text Neg) generated via textual perturbation, positive text (Text Pos) derived from diagnostic reasoning-based expansion, easy and hard negative images (Img Neg_{Easy} and Img Neg_{Hard}) obtained through a dual-constraint negative sample mining strategy, and positive images (Img Pos) refined by wavelet-morphology-guided consistency.

Results are presented in the Table 3, which summarizes the individual contributions to model performance on PathoHR-Bench, along with average zero-shot accuracy across six pathology datasets. We observe that incorporating either negative text or negative images enhances structural awareness, but alone is insufficient to boost zero-shot performance. As detailed in Appendix F, the generation of negative samples introduces contrastive, parallel, and inclusion relationships; among them, inclusion relationships often lead to semantic confusion due to high similarity, which undermines contrastive learning effectiveness.

In contrast, positive text and images improve classification accuracy but may reduce the model's sensitivity to structural variations. Our full scheme achieves optimal results by jointly enhancing structural understanding, compositional reasoning, and generalization ability. To further assess the clinical plausibility of the generated samples, we conducted a qualitative evaluation with a certified pathologist. The evaluation protocol and results are presented in Appendix G.

Model	Information Loss			Semantic Drift			Order Variation			6 Zero-shot Tasks Average
	Ent.	Desc.	Conn.	Ent.	Desc.	Conn.	Ent.	Desc.	Conn.	
CLIP	0.6752	0.6386	0.5544	0.4915	0.5220	0.4655	0.3831	0.3026	0.3409	0.3725
PLIP	0.5945	0.5204	0.4546	0.4720	0.4843	0.4473	0.2926	0.3079	0.2931	0.5246
Ours Text Neg	0.8896	0.8637	0.7833	**0.7032**	0.6651	0.6509	0.7863	0.7640	0.7698	0.5892
Ours Text+Img Neg	0.9005	0.8720	0.7926	0.6924	0.6704	0.6395	0.7914	0.7729	0.7520	0.5804
Ours Text Pos	0.8032	0.8154	0.6941	0.6538	0.6370	0.5786	0.5249	0.5462	0.5033	0.6497
Ours Text+Img Pos	0.7976	0.8203	0.6859	0.6249	0.6459	0.5842	0.5087	0.4576	0.4725	0.6718
Ours w/o Img Neg_{Easy}	0.8529	0.8342	0.7803	0.6625	0.6431	0.6319	0.7635	0.7796	0.7446	0.6728
Ours w/o Img Neg_{Hard}	0.8973	0.8675	0.8018	0.6949	0.6724	0.6507	0.7911	0.7804	0.7801	0.6537
Ours w/o Text Pos	0.8742	0.8631	0.7746	0.6731	0.6522	0.6425	0.7806	0.7649	0.7537	0.6293
Ours w/o Img Pos	0.9022	**0.8791**	0.7854	0.6780	0.6638	0.6409	0.7824	0.7725	0.7679	0.6358
Ours Combined	**0.9018**	0.8604	**0.8075**	0.6853	**0.6744**	**0.6520**	**0.7932**	**0.7819**	**0.7813**	**0.7380**

Table 3: Ablation study on the contribution of different components. Best and second results are highlighted with **bold** and underline.

6 Conclusions

In this study, we investigated the limitations of current VL models in pathological image analysis and introduced a new benchmark to evaluate their structure awareness and compositional reasoning on pathology-specific text-image pairs. To the best of our knowledge, this is the first work to explicitly focus on compositional reasoning in pathological VL models. By targeting fundamental reasoning capabilities, PathoHR-Bench provides deep insights into current model shortcomings and can foster methodological advancements in pathology-specific VL understanding. In addition, we proposed a data-driven training scheme to enhance the fine-grained learning capacity of existing pathological VL models. Through diverse augmentations and perturbations across four branches, our approach not only improves the model's fine-grained reasoning ability but also achieves notable gains in zero-shot pathological diagnosis tasks.

Potential Limitations

Despite the improvements achieved with PathoHR-Bench and our proposed training scheme, several limitations remain. First, while the proposed benchmark systematically evaluates structural awareness and compositional reasoning, it does not yet incorporate external medical knowledge, which could further enhance model interpretability and diagnostic accuracy. Second, although our negative sample construction was carefully designed to simulate realistic yet diagnostically challenging perturbations with both clinical plausibility and semantic precision, its scalability remains limited. Future work could explore more automated approaches, such as leveraging large language models to simplify perturbation generation and improve adaptability. Third, the perturbation strategies used in the framework, though effective, may not fully capture the diverse and nuanced variations present in real-world pathology reports.

Additionally, our approach to model training is data-driven, meaning that its performance is still constrained by the availability and quality of existing pathology datasets. Fourth, although we conducted an expert evaluation to assess the clinical plausibility of generated samples, it was based on a limited sample set and a single pathologist's judgment. Broader expert participation, inter-rater agreement analysis, and context-aware evaluations are needed to fully validate the diagnostic reliability and safety of generated content.

Future work should explore larger-scale multimodal datasets, integrate explicit medical reasoning modules, and refine adaptive augmentation techniques to further improve the robustness and clinical applicability of pathology-focused VL models.

Ethics Statement

The data utilized in our study was sourced from public repositories, and does not pose any privacy concerns. We are confident that our research adheres to the ethical standards set forth by ACL.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Eirini Arvaniti, Kim S Fricker, Michael Moret, Niels Rupp, Thomas Hermanns, Christian Fankhauser, Norbert Wey, Peter J Wild, Jan H Rueschoff, and Manfred Claassen. 2018. Automated gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific reports*, 8(1):12054.

Kaustav Bera, Kurt A Schalper, David L Rimm, Vamsidhar Velcheti, and Anant Madabhushi. 2019. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature reviews Clinical oncology*, 16(11):703–715.

Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester Van Boven, Robert Vink, et al. 2022. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, 28(1):154–163.

Qian Da, Xiaodi Huang, Zhongyu Li, Yanfei Zuo, Chenbin Zhang, Jingxin Liu, Wen Chen, Jiahui Li, Dou Xu, Zhiqiang Hu, et al. 2022. Digestpath: A benchmark dataset with challenge review for the pathological detection and segmentation of digestive-system. *Medical Image Analysis*, 80:102485.

Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. *arXiv preprint arXiv:2403.10378*.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. 2022. Why is winoground hard? investigating failures in visuolinguistic compositionality. *arXiv preprint arXiv:2211.00768*.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. 2023. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668.

Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. 2019. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *Digital Pathology: 15th European Congress, ECDP 2019, Warwick, UK, April 10–13, 2019, Proceedings 15*, pages 11–19. Springer.

Jevgenij Gamper and Nasir Rajpoot. 2021. Multiple instance captioning: Learning representations from histopathology textbooks and articles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16549–16559.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. 2022. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719.

Federica Grillo, Luca Mastracci, Luca Saragoni, Alessandro Vanoli, Francesco Limarzi, Irene Gullo, Jacopo Ferro, Michele Paudice, Paola Parente, and Matteo Fassan. 2020. Neoplastic and pre-neoplastic lesions of the oesophagus and gastro-oesophageal junction. *Pathologica*, 112(3):138.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. 2023. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316.

Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. 2023. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36:37995–38017.

Sajid Javed, Arif Mahmood, Iyyakutti Iyappan Ganapathi, Fayaz Ali Dharejo, Naoufel Werghi, and Mohammed Bennamoun. 2024. Cplip: Zero-shot learning for histopathology with comprehensive vision-language alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11450–11459.

Kishlay Jha, Guangxu Xun, Vishrawas Gopalakrishnan, and Aidong Zhang. 2017. Augmenting word embeddings through external knowledge-base for biomedical application. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1965–1974. IEEE.

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. 2019. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1):e1002730.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ming Y Lu, Bowen Chen, and Faisal Mahmood. 2023a. Harnessing medical twitter data for pathology ai. *Nature Medicine*, 29(9):2181–2182.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. 2024. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Andrew Zhang, Long Phi Le, et al. 2023b. Towards a visual-language foundation model for computational pathology. *arXiv preprint arXiv:2307.12914*.

Ming Y Lu, Bowen Chen, Andrew Zhang, Drew FK Williamson, Richard J Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. 2023c. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19764–19775.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Galileo Mark Namata, Ben London, Lise Getoor, and Bert Huang. 2012. Query-driven active surveying for collective classification. In *International Workshop on Mining and Learning with Graphs*, Edinburgh, Scotland.

Gheyath Othman and Diyar Qader Zeebaree. 2020. The applications of discrete wavelet transform in image processing: A review. *Journal of Soft Computing and Data Mining*, 1(2):31–43.

Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron Van den Oord, Sergey Levine, and Pierre Sermanet. 2019. Wasserstein dependency measure for representation learning. *Advances in Neural Information Processing Systems*, 32.

Pushpak Pati, Guillaume Jaume, Zeineb Ayadi, Kevin Thandiackal, Behzad Bozorgtabar, Maria Gabrani, and Orcun Goksel. 2023. Weakly supervised joint whole-slide segmentation and classification in prostate cancer. *Medical Image Analysis*, 89:102915.

Vu Minh Hieu Phan, Yutong Xie, Yuankai Qi, Lingqiao Liu, Liyang Liu, Bowen Zhang, Zhibin Liao, Qi Wu, Minh-Son To, and Johan W Verjans. 2024. Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11492–11501.

Fengchun Qiao, Long Zhao, and Xi Peng. 2020. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12556–12565.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Ana Isabel Ruiz López, Juan Carlos Pérez Mesa, Yanelis Cruz Batista, and Lienny Eliza González Lorenzo. 2017. Actualización sobre cáncer de próstata. *Correo científico médico*, 21(3):876–887.

Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. 2015. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77.

Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024. Vary: Scaling up the vision vocabulary for large vision-language model. In *European Conference on Computer Vision*, pages 408–424. Springer.

Yutong Xie, Qi Chen, Sinuo Wang, Minh-Son To, Iris Lee, Ee Win Khoo, Kerolos Hendy, Daniel Koh, Yong Xia, and Qi Wu. 2024. Pairaug: What can augmented image-text pairs do for radiology? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11652–11661.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh

Rao, Mu Wei, Naveen Valluri, et al. 2023. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2022a. An explainable toolbox for evaluating pre-trained vision-language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 30–37.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2022b. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*.

A Adversarial distribution-aware perturbation

This section provides the mathematical formulation and optimization process of the adversarial distribution-aware perturbation branch used for hard negative sample generation in our method.

To expose the model to visually plausible yet structurally challenging examples, we aim to generate hard negative images that deviate from the source domain distribution M_0 but still remain within the manifold of valid pathological appearances. We consider the worst-case scenario around the source distribution M_0 as Equation (1), its solution guarantees good performance against data distributions that are distance m away from M_0. The Wasserstein distance (Ozair et al., 2019) is used as the distance metric D, which is defined by calculating the minimum transport cost between two distributions in the representation space. Assuming that X_N and X_0 are obtained by sampling from the generated distributions M_N and original distribution M_0, we can measure the distance in the representation space as follows:

$$D_\theta\left(M_N, M_0\right) = inf_{\gamma \in \Gamma\left(M_N, M_0\right)} E_\gamma\left[\|\|f\left(\theta, X_N\right) - f\left(\theta, X_0\right)\|\|_2^2\right] \tag{11}$$

where $f\left(\theta, X_N\right)$ is the encoder of the contrastive learning model, γ is the joint probability distribution that satisfies the marginal distribution for M_N and M_0, and Γ is the set of all joint probability distributions that satisfy the marginal distribution.

By substituting the formula for the distance metric defined in Equation (11) into the worst-case objective Equation (1), the generated distribution M_N that satisfies $D\left(M_N, M_0\right) \leq m$ achieves maximal divergence in the feature space while preserving pathological plausibility and structural coherence with respect to the source domain. However, directly solving for the supremum over the constrained distributional space is intractable. To address this, we introduce a Lagrangian relaxation with a penalty term λ that softly enforces the adversarial constraint. The final formulation for adversarial distribution-aware hard negative generation becomes:

$$\min_\theta \sup_{M_N} E_{M_N}\left[L\left(\theta, I\right) - \lambda D_\theta\left(M_N, M_0\right)\right] \tag{12}$$

B Datasets

We used six independent publicly available computational pathology datasets, including four at the patch level and two at the whole-slide level:

i). CRC100K (Kather et al., 2019): is a colorectal cancer dataset comprising 224×224 pixel image patches acquired at $0.5\mu m$ per pixel resolution from 50 patients. It includes nine tissue categories: colorectal adenocarcinoma epithelium, normal mucosa, smooth muscle, lymphocytes, mucus, cancer-associated stroma, adipose tissue, background, and debris. The official split consists of 100,000 training and 7,180 testing images. For zero-shot tile-level classification, we directly evaluated on the testing split without any fine-tuning.

ii). UHU (Arvaniti et al., 2018): is a prostate cancer dataset comprising five tissue microarrays with a total of 886 tissue cores, each sized at $3,100 \times 3,100$ pixels. All slides were scanned at $40\times$ magnification using a NanoZoomer scanner. An experienced pathologist annotated benign (BN) regions and three cancer grades (Gleason grade 3, 4, and 5) with pixel-level segmentation masks. We adopted the official preprocessing pipeline from (Arvaniti et al., 2018), generating a total of 22,022 image patches (750×750 pixels) after excluding patches dominated by luminal or unannotated areas. The training set contains 2,076 BN, 6,303 grade 3, 4,541 grade 4, and 2,383 grade 5 patches. The test set includes 127 BN, 1,602 grade 3, 2,121 grade 4, and 387 grade 5 patches. Only the patch-level test set was used for zero-shot evaluation.

iii). PanNuke (Gamper et al., 2019): is a multi-organ dataset designed for nuclei segmentation and classification, encompassing 19 tissue types across various pathological conditions. It contains 4,346 training and 1,888 testing images, each sized at 256×256 pixels. Following the evaluation protocol used in PLIP, we assess zero-shot performance on

the testing split for a binary classification task that distinguishes Tumor vs. Normal Benign.

iv). DigestPath (Da et al., 2022): is a dataset of H&E-stained colonoscopy tissue sections, comprising 660 whole-slide images. Following the protocol used in PLIP, we conducted patch-level zero-shot classification (Tumor vs. Normal) on the official testing set, which includes 18,814 image patches.

v). TCGA-BRCA (Tomczak et al., 2015): is a whole-slide image (WSI) dataset of invasive breast carcinoma derived from The Cancer Genome Atlas (TCGA), comprising two subtypes: Invasive Ductal Carcinoma (IDC) and Invasive Lobular Carcinoma (ILC). It includes a total of 1,048 WSIs, with 837 IDC and 211 ILC slides. For zero-shot WSI-level classification, following the protocol in MI-Zero, we used a test set consisting of 75 WSIs from each class, ensuring no patient-level overlap between training and testing splits.

vi). TCGA-RCC (Tomczak et al., 2015): is a renal cell carcinoma WSI dataset from TCGA consisting of three subtypes: Clear Cell RCC (CCRCC, 519 slides), Papillary RCC (PRCC, 294 slides), and Chromophobe RCC (CHRCC, 109 slides), totaling 922 WSIs. For zero-shot classification, we use 75 WSIs from each subtype as the test set, with no patient overlap between training and testing, following the MI-Zero setup.

The first four datasets are at the patch level, focusing on localized morphological patterns, while the latter two are whole-slide level datasets that require global diagnostic reasoning across multiple cancer subtypes.

C Implementation Details

We implemented all models using PyTorch and conducted training on 2 NVIDIA A100 GPUs. Across all vision-language pretraining variants, we used a temperature parameter of 0.02 for the contrastive loss and optimized using AdamW (Loshchilov and Hutter, 2017) with an initial learning rate of 5×10^{-6}. A cosine decay scheduler was applied throughout training. All models were trained for 50 epochs with a batch size of 256. During fine-tuning, we adopted **BioClinicalBERT** (with a maximum token length of 512) as the text encoder, and **PLIP-ViT-B/32-224** as the visual encoder. All images were resized to 224×224, and standard data augmentations including random cropping, horizontal flipping, and color jittering were applied during training. For zero-shot classification, we used single prompts per class for evaluation, following the standard setup in prior works. Following a unified protocol, we finetune all pathology-focused baselines on the ARCH training split using their *original* objectives and *without* our perturbation-based augmentations, and then evaluate them on PathoHR-Bench and public datasets with a single prompt per class. Although PLIP is pretrained on large biomedical corpora, it is not originally fine-tuned on ARCH; we therefore additionally finetune PLIP on ARCH to reduce distribution shift. To preserve a general-purpose reference, CLIP remains a purely zero-shot baseline (no ARCH fine-tuning).

Model (no ARCH finetune)	UHU		PanNuke		DigestPath	
	Acc	F1	Acc	F1	Acc	F1
BiomedCLIP (w/o finetune)	0.3371	0.1804	0.5019	0.5073	0.5481	0.5226
PLIP (w/o finetune)	0.3618	0.2085	0.6139	0.6014	0.7890	0.8027

Table A1: Zero-shot results *without* ARCH fine-tuning for two widely used baselines. Numbers are reported with a single prompt per class under our unified protocol.

D Zero-shot without ARCH fine-tuning.

In Table A1, we observe that the performance differences between the settings with and without fine-tuning on ARCH are not very large, especially for PLIP. This is likely because PLIP was already pretrained on a sufficiently large number of pathology-specific image–text pairs, so plain fine-tuning on paired data may approach an upper limit for representation learning on coarser-grained tasks such as PanNuke and DigestPath. In contrast, our method yields clear improvements on fine-grained subtype classification (CRC100K and UHU) and on PathoHR-Bench even when trained with only limited ARCH data.

We attribute these gains to the use of carefully designed structured positives and targeted hard or soft negatives that explicitly supervise compositional reasoning over entities, descriptors, and relations, rather than relying solely on generic image-text alignment. Taken together, these results indicate that our framework can more effectively capture the inherent logic and hierarchical structure of pathology data beyond what standard fine-tuning can achieve.

E Case Study

In this section, we present additional case studies to further demonstrate the improvement in fine-grained learning capabilities achieved by our proposed training scheme. Figure A1 illustrates examples of semantic drift perturbations at the connections level, information loss perturbations at the descriptors level, and order variation perturbations at the entities level within the PathoHR-Bench. Additionally, comparative results are presented on PLIP and our proposed training framework across these tasks, highlighting their performances in handling different types of perturbations. The results demonstrate that our proposed training scheme significantly enhances compositional reasoning and structural awareness, enabling the model to better understand pathological diagnostic logic and reasoning processes.

To further highlight the advantages of our scheme in real-world diagnostic applications, we take a practical diagnostic task as an example. Prostate Gleason grading is a standard pathology assessment used to evaluate the aggressiveness of prostate cancer, which is classified into four grades (Benign, Grade 3, Grade 4 and Grade 5). Although existing models can effectively distinguish between low-grade and high-grade cancers, they often struggle to differentiate between Grade 3 and Grade 4, where the differences are subtle and accurate classification requires attention to fine-grained pathological features, leading to frequent mis-classifications. Such errors can significantly affect prognosis and treatment decisions. As shown in the figure A2, our model exhibits a notable advantage on these challenging borderline cases, demonstrating its superior capability in capturing subtle pathological differences and providing more reliable diagnostic predictions.

F Pathological Relationships in Negative Sample Generation

In this section, we analyze three types of relationships that may arise when generating negative text samples as shown in Table A2. Contrasting relationships refer to pairs of pathological concepts with opposite pathological characteristics or diagnostic properties, while parallel relationships represent concepts at the same hierarchical level but belonging to different pathological categories. Negative samples generated through these two relationships enhance the model's fine-grained classification capability and category boundary awareness. However, there exists an inclusive relationship, where one pathological concept is a subset of a broader concept, which can result in high semantic similarity between negative samples and the original samples, thereby reducing the effectiveness of contrastive learning.

G Expert Evaluation of Generated Samples

To assess the clinical relevance and diagnostic validity of the generated samples, we conducted a qualitative review with an expert pathologist.

G.1 Evaluation Setup

We randomly selected 200 samples, divided into five categories (40 per type):
i). Text Neg: generated via semantic perturbation.
ii). Img Neg_{Easy}: generated by Stable Diffusion guided by corrupted diagnostic texts.
iii). Img Neg_{Hard}: generated through adversarial distribution-aware perturbation.
iv). Text Pos: generated through hierarchical diagnostic reasoning.
v). Img Pos: generated via wavelet-morphology-guided consistency refinement.

Each sample was rated by the expert using a 1–5 Likert scale on dimensions relevant to its intended purpose:

- For positive samples and hard negative images: clinical realism and structural integrity.
- For text-guided negatives and textual negatives: semantic inconsistency clarity and misleading plausibility.

Free-form comments were also collected.

G.2 Evaluation Summary

Textual Negatives: 90% were judged as plausibly misleading, effectively simulating clinical contradictions. Inclusion-type perturbations were particularly subtle and challenging.
Text-guided Image Negatives: 85% of samples exaggerated incorrect structures in a way that was visually interpretable. While not clinically realistic by design, the expert confirmed their usefulness for training models to reject structurally invalid patterns.
Adversarial Image Negatives: 87.5% retained realistic tissue appearance while introducing subtle,

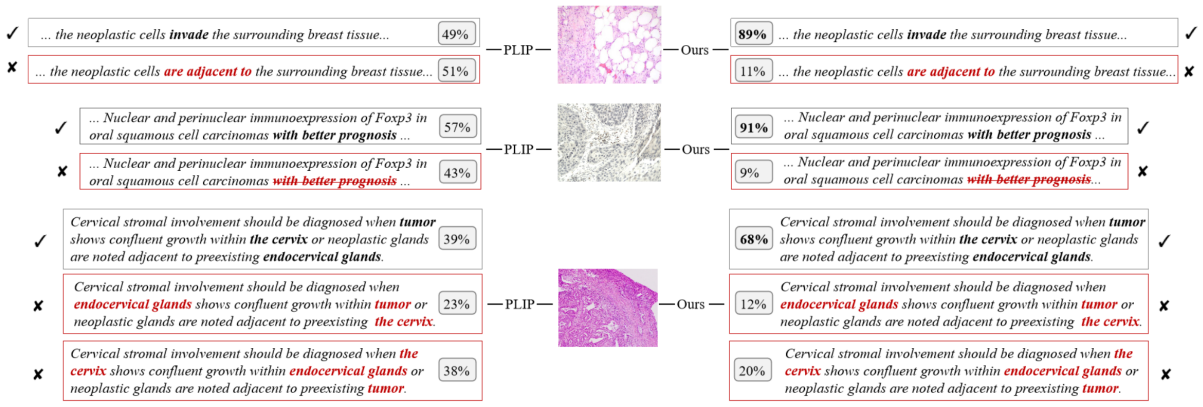


Figure A1: Cases study of semantic drift perturbations at the connections level, information loss perturbations at the descriptors level, and order variation perturbations at the entities level within the PathoHR-Bench.

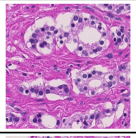
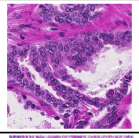
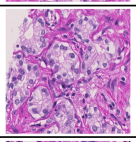
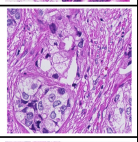
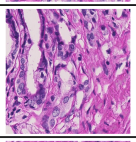
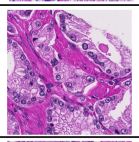
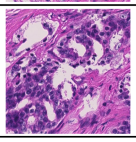
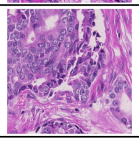
Ground Truth	Image	Model	Grade 3	Grade 4	Ground Truth	Image	Model	Grade 3	Grade 4
Grade 3		CLIP	0.42	0.58	Grade 3		CLIP	0.39	0.61
		PLIP	0.70	0.30			PLIP	0.43	0.57
		MI-Zero	0.51	0.49			MI-Zero	0.59	0.41
		Ours	**0.76**	0.24			Ours	**0.69**	0.31
Grade 4		CLIP	0.47	0.53	Grade 4		CLIP	0.48	0.52
		PLIP	0.41	0.59			PLIP	0.36	0.64
		MI-Zero	0.47	0.53			MI-Zero	0.36	0.64
		Ours	0.37	**0.63**			Ours	0.28	**0.72**
Grade 3		CLIP	0.56	0.44	Grade 3		CLIP	0.59	0.41
		PLIP	0.63	0.37			PLIP	0.75	0.25
		MI-Zero	0.69	0.31			MI-Zero	0.81	0.19
		Ours	**0.82**	0.18			Ours	**0.94**	0.06
Grade 4		CLIP	0.52	0.48	Grade 4		CLIP	0.43	0.57
		PLIP	0.41	0.59			PLIP	0.28	0.72
		MI-Zero	0.40	0.60			MI-Zero	0.20	0.80
		Ours	0.11	**0.89**			Ours	0.08	**0.92**
Prompt	Grade 3	*Photo of a prostate biopsy showing* ***Gleason Grade 3 (well-formed separate glands).***							
	Grade 4	*Photo of a prostate biopsy showing* ***Gleason Grade 4 (fused glands with cribriform pattern).***							

Figure A2: Case Study of fine-grained classification for Gleason grades.

Relationship	Case
Contrasting Relationship	in colon **carcinoma** → in colon **adenoma**
Parallel Relationship	in **colon** carcinoma → in **gastric** carcinoma
Inclusion Relationship	in **colon** carcinoma → in **gastrointestinal** carcinoma

Table A2: Pathological relationship types in negative sample generation.

distribution-aware shifts. Many resembled borderline cases in real practice.

Textual Positives: 80% were rated clinically coherent, capturing diagnostic logic. Some samples were considered too generic or lacking critical context for high-grade interpretations.

Image Positives: 92.5% preserved diagnostic structures and enhanced local detail (e.g., nuclei, fibrous margins) without introducing artifacts.

G.3 Expert-Identified Limitations

During the review, the expert also identified several limitations in the generated samples. Some text-guided image negatives exhibited subtle structural artifacts or biologically implausible tissue combinations. In several cases, semantic perturbations were considered too trivial to mislead a diagnostic system. For positive text expansions, certain samples lacked sufficient diagnostic context, such as staging information or relevant biomarkers. These observations provide valuable feedback for refining the sample generation strategies and improving their clinical fidelity.