PATHOCLASS-BRCA: Reframing Pathology Report Generation as Guideline-Aligned Multi-Task Classification

Alessia Saporita[1, 2, *]
alessia.saporita@unimore.it
Vittorio Pipoli[1, *]
vittorio.pipoli@unimore.it
Federico Bolelli[1, ✉]
federico.bolelli@unimore.it
Andrea Acquaviva[2]
andrea.acquaviva@unibo.it
Elisa Ficarra[1]
elisa.ficarra@unimore.it

[1] "Enzo Ferrari" Engineering Department, University of Modena and Reggio Emilia, Modena, Italy

[2] Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi," University of Bologna, Bologna, Italy

Abstract

Histopathology is the gold standard for cancer diagnosis, and pathology reports directly inform treatment decisions. Although recent deep learning approaches aim to automate report generation from whole-slide images, most methods inherit training objectives and evaluation protocols from image captioning, adopting token-level losses and n-gram metrics that prioritize lexical similarity over diagnostic correctness. To address these limitations, we reformulate pathology report generation as a guideline-aligned multi-task classification problem. Our approach predicts clinically relevant pathological features defined by the International Collaboration on Cancer Reporting guidelines, thereby enabling structured and diagnostically meaningful automated reporting. Our contributions are threefold: *i)* we introduce PATHOCLASS-BRCA, the first breast cancer benchmark that reframes pathology report generation as guideline-aligned multi-task classification of clinically relevant pathological features; *ii)* we propose a pathological feature-level semantic evaluation protocol that maps free-text reports to guideline-defined pathological labels, enabling clinically meaningful evaluation of conventional report generation models using classification metrics; and *iii)* we demonstrate that classification-oriented adaptation of report generation backbones significantly improves the recognition of clinically meaningful histopathological features. The source code and dataset are publicly released at https://github.com/AImageLab-zip/PathoClassBRCA.

1 Introduction

In oncology, microscopic examination of tissue specimens remains fundamental for establishing diagnosis, assessing tumour morphology, and identifying clinically relevant features.

* Equal contribution. Authors may list their names first in their respective CVs. ✉ Corresponding author.

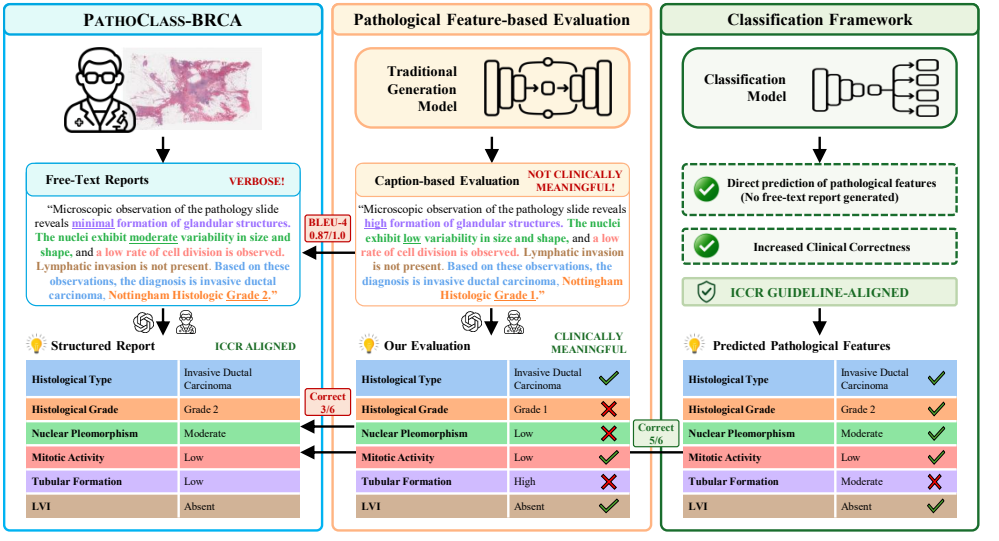

Figure 1: **Overview of our three proposals. Left:** We convert free-text pathology reports derived from WSIs into structured labels for six clinically relevant pathological features, aligned with ICCR reporting guidelines, to construct the PATHOCLASS-BRCA dataset. Preliminary labels are extracted using GPT-5 and verified under pathologist supervision. Coloured text highlights the report evidence used by GPT-5 to assign each pathological-feature label. **Centre:** Conventional free-text report generation models are typically evaluated with captioning metrics such as BLEU-4, which measure lexical overlap rather than clinical correctness. As shown, a report may achieve a high BLEU-4 despite differing in only a few clinically relevant terms. Our evaluation protocol, instead, maps generated reports to the six pathological labels and compares them with our annotations, enabling a quantitative feature-level assessment of clinical correctness. **Right:** We reformulate automated pathology reporting as a guideline-aligned multi-task classification problem. Specifically, we derive a classification-oriented version of free-text report generation models to directly predict the six guideline-aligned pathological features, improving clinical correctness.

Pathology reports translate these findings into diagnostic evidence that supports treatment planning and prognostic assessment [14, 27, 34]. To improve report quality and standardization, cancer reporting guidelines—the International Collaboration on Cancer Reporting (ICCR), the College of American Pathologists (CAP), and the Royal College of Pathologists of Australasia (RCPA) [10, 17]—promote structured reporting, enhancing completeness, consistency, and comparability across cases and institutions. Nevertheless, pathology report drafting remains labour-intensive and time-consuming, motivating automated methods that generate diagnostic reports directly from whole-slide images (WSIs) [5, 13, 20, 39].

Chen *et al*. [5] and Guo *et al*. [13] laid the foundation for this line of work by introducing the first publicly available datasets derived from pathology reports in The Cancer Genome Atlas (TCGA) [4], along with task-specific report generation models, namely MI-Gen and HistGen. More recently, Multimodal Large Language Models (MLLMs) have been introduced in computational pathology for a broad range of tasks, including visual question answering, morphological reasoning, and pathology report generation [1, 20, 25, 31].

Despite these advances, most existing approaches still formulate pathology report generation as open-ended free-text generation, adopting training objectives and evaluation protocols [3, 21, 29] inherited from image captioning [2, 36]. Throughout this paper, we refer to models that generate free-text pathology reports as "report generation models." These models are typically trained with token-level cross-entropy losses that reward surface-form similarity to reference reports, while evaluation relies on captioning metrics that quantify lexical agreement rather than diagnostic correctness. Recent work attempts to address the limitations of captioning metrics by introducing semantic evaluation strategies based on named entity recognition (NER) [28] or LLM-based claim extraction [20]. Although these approaches move beyond lexical overlap, they typically reduce agreement over extracted entities or statements to a global semantic score. As a result, they remain disconnected from cancer reporting guidelines and fail to identify which clinically relevant pathological features have been correctly predicted, limiting their interpretability and clinical utility.

In this work, we address these limitations by reformulating pathology reporting as a clinically grounded multi-task classification problem. We build on the observation that international cancer reporting guidelines, *e.g.* ICCR, define structured reporting templates composed of standardised fields, each corresponding to a clinically relevant pathological feature. This structure enables us to move beyond free-text generation and reformulate automated reporting as the prediction of guideline-defined pathological features directly from WSIs. Specifically, each feature is treated as a distinct classification task, whose predicted value can populate the corresponding reporting field. This formulation provides a training objective explicitly optimised for the extraction of relevant pathological features and enables diagnostic performance to be evaluated at the level of individual pathological features using classification metrics, rather than relying on textual similarity to reference reports.

Relatedly, Fan *et al.* [12] show in radiology that auxiliary classification tasks improve clinically meaningful finding detection and factual correctness in report generation. However, free-text generation remains their final objective and does not enable guideline-level diagnostic evaluation. In contrast, we directly predict guideline-defined pathological features from WSIs, enabling feature-level evaluation of diagnostic correctness.

To this end, we introduce **PATHOCLASS-BRCA**, which, to the best of our knowledge, is the first breast cancer benchmark for multi-task classification of guideline-relevant pathological features. We focus on TCGA-BRCA [5, 39], as it represents the largest TCGA cohort by number of cases [4] and a primary benchmark for existing state-of-the-art report generation models. Guided by the ICCR invasive breast cancer reporting guidelines [11, 17], we identify six clinically relevant pathological features: histological type, histological grade, its constituent components, namely tubular formation, nuclear pleomorphism, and mitotic activity, and lymphovascular invasion (LVI). For each pathological feature, we design a *feature-specific labelling prompt* in consultation with pathologists, specifying the target feature, standardised output labels, and label-assignment rules. These prompts guide GPT-5 in extracting preliminary labels from pathological reports, which are subsequently verified by pathologists to ensure clinical correctness.

Building on this formulation, we introduce the first **pathological feature-level semantic evaluation protocol** for report generation models [1, 5, 13, 20, 39]. For each free-text generated report, we leverage the aforementioned feature-specific labelling prompts to extract standardised labels for the predefined pathological features, which are then compared with PATHOCLASS-BRCA ground-truth labels using classification metrics such as F1-score [32] and recall [8]. Furthermore, we show that deriving a **classification-oriented adaptation** of report generation backbones [5, 13, 39] substantially improves the prediction of guideline-

relevant pathological features across most tasks. These results demonstrate that directly optimising for pathological feature recognition enhances diagnostic correctness compared with image-captioning training objectives. An overview of our proposal is depicted in Figure 1.

Overall, this work reframes pathology report generation as a guideline-aligned prediction problem, moving from unstructured language generation toward explicit recognition of clinically meaningful pathological features. This formulation promotes structured reporting, improves diagnostic interpretability, and enables consistent feature-level evaluation of automated pathology reports.

Contributions. Our contributions are threefold: ***i)*** we introduce **PATHOCLASS-BRCA**, the first breast cancer benchmark that reformulates pathology report generation as six-task classification of guideline-aligned pathological features; ***ii)*** we propose a **pathological feature-level semantic evaluation protocol** that assesses diagnostic correctness by mapping free-text reports to standardised pathological labels and evaluating them with classification metrics; and ***iii)*** we show that **classification-oriented adaptations** of traditional report generation backbones substantially improve the recognition of clinically meaningful histopathological features.

2 Related Works

2.1 Breast Cancer Reporting Guidelines

Assessing pathological features is essential for breast cancer staging, prognosis, and treatment planning. Accordingly, several reporting protocols have been developed to support comprehensive and standardised pathology reporting [17], including those proposed by the ICCR, CAP, and RCPA. Among these, the ICCR provides standardised reporting templates for cancer pathology reporting, aiming to harmonise report structure and terminology across institutions and healthcare systems [11]. These templates are developed through international expert consensus involving specialist pathologists and are closely aligned with the World Health Organization (WHO) Classification of Tumours series [7]. They distinguish between core and non-core elements, with core elements representing parameters considered essential for clinical management, staging, prognostic assessment, or treatment decision-making. In this work, we adopt the ICCR guidelines as the reference framework for selecting the pathological features included in our classification dataset, owing to their consensus-based development, explicit identification of clinically relevant reporting elements, and alignment with established WHO tumour classification standards.

2.2 Pathology Report Generation Models

Pathology report generation from WSIs aims to translate histopathological images into diagnostic descriptions. Existing methods can be broadly divided into two lines of research: *task-specific pathology report generation methods* and *pathology MLLMs*, the latter also supporting other diagnostic tasks.

Task-specific pathology report generation methods. Early efforts focused on task-specific frameworks, including MiGen, HistGen, and BiGen, which formulate WSI-based pathology report generation under a multiple instance learning (MIL) paradigm [9, 16, 24]. Under this formulation, each WSI is represented as a bag of patch-level instances, whose features are extracted using pathology foundation models [6, 15, 38] and aggregated into a slide-level

embedding that conditions the generation of a free-text pathology report. MiGen [5] is one of the first works to frame WSI-based pathology report generation as a multiple-instance generation problem. It introduces a large-scale TCGA-derived WSI–report dataset and an encoder–decoder architecture for report generation. HistGen [13] extends this line of work by introducing a new WSI–report dataset and a MIL-based architecture that combines local and global visual information. Specifically, it integrates hierarchical feature encoding with cross-modal context interaction, allowing the model to better capture diagnostically relevant regions. These datasets are generally constructed from digitised pathology reports in TCGA [4], using optical character recognition (OCR) [26] to extract textual content from scanned reports, followed by large language model (LLM)-based processing [22, 30, 33] to reduce noise and remove redundant information. More recently, BiGen [39] introduces a retrieval-augmented generation (RAG) framework for WSI-based report generation that retrieves relevant historical reports as auxiliary textual context to enrich visual representations with information from similar cases.

Pathology MLLMs. More recently, MLLMs have been introduced in computational pathology to support a broader range of vision–language tasks, including visual question answering, morphological reasoning, diagnostic reasoning, treatment planning, and pathology report generation [1, 20, 25, 31]. They typically rely on a visual encoder [6, 15, 38] to extract WSI representations, a projection module to align visual features with the language embedding space [18, 41], and an instruction-tuned language model [22, 23] to generate responses conditioned on both the slide representation and the user instruction, enabling more flexible interaction with histopathological data and supporting diverse pathology-related tasks beyond report generation. Among these, WSI-LLaVA [20] currently represents the state-of-the-art MLLM for WSI understanding. It employs a three-stage training approach: WSI-text alignment, which aligns WSI and report representations; feature space alignment, which maps WSI features into the LLM embedding space; and task-specific instruction tuning, which jointly adapts the projector and LLM to downstream pathology tasks. Notably, WSI-LLaVA builds on WSI-Bench [20], a refined benchmark designed to retain morphologically observable content while filtering out information that cannot be inferred from WSIs alone.

Despite these advances, existing pathology report generation models remain largely formulated as image-captioning systems. Their training targets are unstructured free-text reports, and their outputs are optimised for linguistic similarity with reference texts rather than diagnostic correctness. In contrast, our work reformulates automated pathology reporting as guideline-aligned multi-task classification, replacing unconstrained report generation with the direct prediction of predefined pathological features.

2.3 Pathology Multi-Task Classification

Multi-task learning has gained increasing attention in medical image analysis as a framework for addressing multiple related prediction objectives through shared representations and task-specific heads. For instance, Wang *et al*. [37] proposed a multi-task network for colorectal lymph node metastasis diagnosis, jointly performing metastasis detection and subtyping. However, this approach targets a specific classification problem and does not address pathology report generation or standardised reporting of multiple pathological features. Closer to WSI vision–language learning, Zhou *et al*. introduce PathM3 [40], a multimodal multi-task MIL framework for joint WSI classification and captioning. By leveraging diagnostic captions, PathM3 improves WSI representation learning under limited textual supervision.

Nevertheless, its classification objective is independent of the report content and does not target standardised pathological features, while the reporting task remains formulated as free-text generation. In contrast to these works, we formulate WSI-level breast pathology reporting as the direct prediction of guideline-defined pathological features. This formulation treats each feature as an explicit classification task, enabling standardised reporting and feature-level evaluation of diagnostic correctness.

3 Method

In this work, we rethink pathology report generation as a clinically grounded multi-task classification problem. Our proposal consists of three main components: **PATHOCLASS-BRCA**, a multi-task classification benchmark which provides ground-truth labels for breast-cancer guideline-aligned pathological features; a **pathological feature-level semantic evaluation protocol** that assesses the diagnostic correctness of free-text generated reports by comparing their extracted pathological features with PATHOCLASS-BRCA labels; and a **classification-oriented adaptation** of traditional report generation backbones to the proposed multi-task classification setting.

3.1 PATHOCLASS-BRCA: A Classification-Oriented Dataset

We introduce PATHOCLASS-BRCA, a benchmark for evaluating multi-task classification of guideline-relevant pathological features in breast cancer pathology. We focus on breast invasive carcinoma from TCGA-BRCA, as it represents a widely used benchmark in WSI-based report generation and the largest TCGA cohort by number of cases. To construct our classification dataset, we build on WSI-Bench [20], as it retains only report content morphologically observable from WSIs, while excluding non-visual clinical information such as gross descriptions, metastasis status, tumour size and staging, margin status, and immunohistochemical findings. This curation makes WSI-Bench particularly suitable for constructing a classification-oriented benchmark aimed at learning and evaluating pathological features.

Selection of pathological features. Cancer reporting guidelines organise diagnostic information into structured fields, each associated with a specific pathological feature and a constrained set of admissible values. Guided by the ICCR histopathology reporting guide for invasive carcinoma of the breast [11, 17], we first analyse the standardised reporting template to identify clinically relevant fields and their corresponding values. In collaboration with pathologists, we select six ICCR-guided pathological features that are clinically relevant for breast cancer assessment, visually assessable from WSIs, and reported in at least 50% of available pathology reports: histological tumour type, histological tumour grade, tubular formation, nuclear pleomorphism, mitotic activity, and LVI. Further details on feature selection are provided in the Supplementary Material.

These features capture complementary morphological properties of invasive breast carcinoma: histological tumour type describes the morphological subtype of the tumour; histological tumour grade summarises its degree of differentiation; tubular formation measures the preservation of gland-like structures; nuclear pleomorphism captures abnormalities in tumour cell nuclei; mitotic activity reflects tumour cell proliferation; and LVI indicates the presence of tumour cells within lymphatic or blood vessels. These fields are reported as core diagnostic elements in invasive breast carcinoma reporting and are directly associated with microscopic tumour morphology.

-Goal-
Extract the target pathological attribute from the breast pathology report text.

-Steps-
1. Scope Rule. Apply `{scope_rule}` when the attribute must be restricted to the main carcinoma.
2. Normalization Rules. Map heterogeneous report expressions to the allowed standardized classes using `{normalization_rules}`.
3. Conservative Rule. If multiple or conflicting mentions are present, select the highest class.
4. Nottingham Score Rule. When applicable, compute the Nottingham score from its components if not explicitly stated and map it to standardized labels using `{nottingham_score_rule}`.
5. Range Rule. When ranges or composite labels/scores are present, choose the highest one.
6. Additional Rules and Strict Constraints. Use `{additional_rules}` and `{strict_constraints}` to handle missing, uncertain, or non-assessable information.

-Real Data-
Report text: `{report_text}`

-Output Format-
Return valid JSON only:

```
{
   "{target_attribute}": "one allowed class",
   "evidence":           "short quote from the report",
   "confidence":         float in [0,1]
}
```

Feature-Specific Labelling Prompt

Figure 2: Generic feature-specific labelling prompt used to automatically extract structured pathological features from unstructured breast cancer pathology reports.

Label extraction via feature-specific labelling prompts. For each selected pathological feature, we design a dedicated feature-specific labelling prompt to extract the corresponding label from the pathology reports. An overview of the generic structure shared by all feature-specific labelling prompts is provided in Figure 2. Each prompt defines the extraction task, specifies the admissible label set, and provides feature-specific normalisation rules to map heterogeneous reporting expressions to a standardised class. The prompts also include scope constraints to ensure that labels for the grading components and LVI are extracted only from evidence describing the main invasive carcinoma, while disregarding mentions associated with in situ lesions, benign findings, or secondary pathological entities when applicable.

For histological grade, the prompt incorporates the rule-based computation of the final grade from its constituent components, namely tubular formation, nuclear pleomorphism, and mitotic activity. To reduce ambiguity, conservative decision rules are adopted for cases with multiple or conflicting mentions, such as selecting the highest grade or severity class when ranges or discordant expressions are reported. Finally, each prompt defines how to handle uncertain, missing, or conflicting information, ensuring consistent label extraction across reports. The model is instructed to return a structured JSON output containing the extracted label, a short supporting evidence span from the report, and a confidence score.

Using this prompting framework, each pathology report is processed independently for each selected feature. Specifically, GPT-5 is queried with the corresponding feature-specific labelling prompt to obtain a preliminary standardised label for that feature. When a feature cannot be determined either directly from the report or through the predefined derivation rules, we assign a label of -1 to denote missing information. The extracted labels are then reviewed and corrected by pathologists, yielding the final ground-truth labels used in our proposed dataset PATHOCLASS-BRCA.

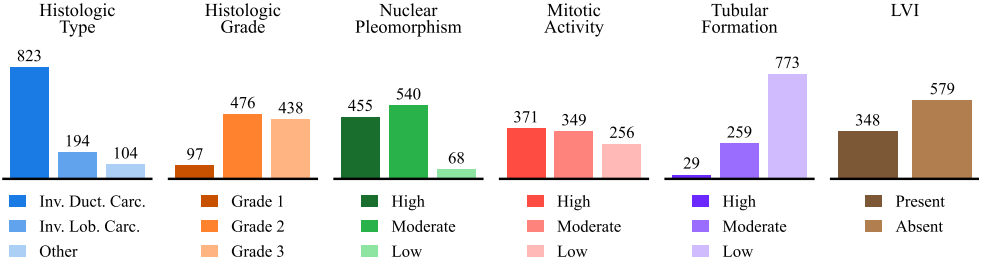

Figure 3: PATHOCLASS-BRCA label distributions for the six selected pathological features.

The left panel of Figure 1 illustrates the proposed label-extraction pipeline, which transforms unstructured pathology reports into structured and standardised labels for the selected guideline-relevant pathological features. The extracted labels are subsequently reviewed by a pathologist to ensure their clinical consistency and reliability. Figure 3 provides a dataset-level overview of the resulting label distributions after this review process, revealing substantial class imbalance across several features, with histological type exhibiting the most pronounced skew. Overall, the automatically extracted labels show strong concordance with the pathologist-reviewed annotations, achieving an overall agreement of 96.7% across the selected pathological features.

3.2 Pathological Feature-level Semantic Evaluation Protocol

A further key contribution is our pathological feature-level semantic evaluation protocol, which evaluates the diagnostic correctness of free-text generated pathology reports by determining whether clinically relevant pathological features are accurately identified. This enables the validation of existing report generation methods beyond textual similarity, shifting their evaluation toward clinical correctness.

Conventional report generation metrics, typically adopted directly from image captioning, provide an inadequate basis for clinical evaluation because they primarily measure lexical similarity to reference reports through n-gram overlap rather than diagnostic agreement. High textual similarity does not necessarily imply clinical correctness: a generated report may closely match the reference in sentence structure, wording, and lexical content while incorrectly describing clinically relevant pathological features. In contrast, our protocol converts free-text reports into structured pathological-feature predictions, enabling a quantitative assessment of pathological features rather than measuring text similarity.

Our protocol consists of three main steps. First, we apply the previously described label extraction pipeline to each generated report to derive feature-specific labels for the six selected pathological features. Second, the extracted labels are matched against the corresponding ground-truth labels in PATHOCLASS-BRCA. Third, these feature-level comparisons are aggregated across samples to compute classification metrics, *e.g.* F1-score [32] and recall [8]. Features not specified in the ground-truth report, corresponding to a ground-truth label of -1, are excluded from the evaluation for that report. To further validate the proposed protocol, pathologists review labels extracted from $1,500$ reports generated by HistGen, MiGen, BiGen, and WSI-LLaVA. Our extraction protocol achieves 96.4% overall concordance in identifying the intended feature values in generated text.

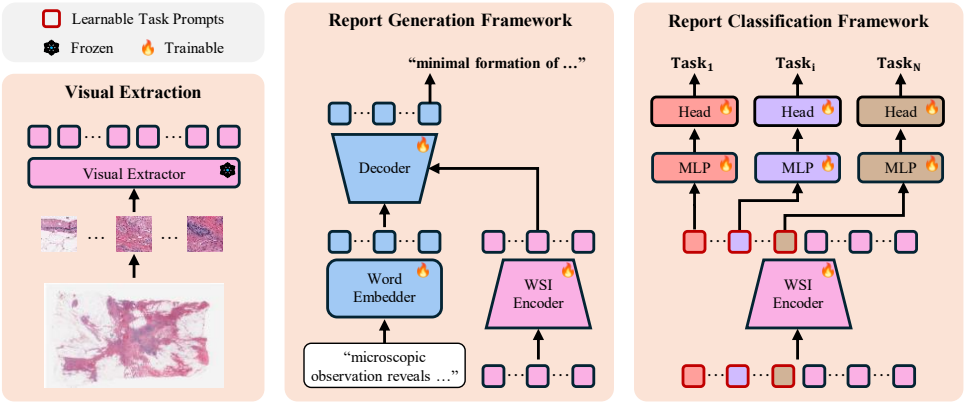

Figure 4: Comparison between the conventional report generation framework and our proposed report classification framework.

3.3 Classification-Oriented Models

In this work, we reformulate WSI-based pathology reporting as the prediction of predefined pathological features rather than as free-text report generation. We therefore adapt existing report generation backbones to a multi-task classification setting, in which each model directly predicts a fixed set of guideline-defined pathological features from the WSI. This formulation allows us to assess whether architectures originally developed for report generation can be reformulated to improve pathological feature recognition, thereby supporting more structured, interpretable, and guideline-aligned reporting. Figure 4 compares the conventional generation framework with the proposed classification-oriented formulation.

Model Adaptation. We adapt representative report generation methods to the proposed multi-task classification paradigm with minimal architectural modifications. As shown in Figure 4, pathology report generation models share a common structure with two main branches: a textual branch and a visual branch. In the textual branch, text is tokenised and encoded through a learnable word embedder, then decoded using either a small transformer decoder or an entire LLM, conditioned on the visual branch. The visual branch encodes WSI patches using a pretrained, frozen visual extractor to obtain patch-level embeddings $\mathbf{X} \in \mathbb{R}^{n_{\text{tokens}} \times d_m}$, where n_{tokens} denotes the number of patch tokens and d_m the embedding dimension. These embeddings are then contextualised through a WSI encoder, typically a transformer-based backbone. The only exception is WSI-LLaVA [20], whose WSI encoder consists solely of a visual projector, implemented as a linear layer, that maps visual features into the LLM representation space. In this case, the limited capacity of the WSI encoder is compensated for by the substantially larger LLM decoder.

We derive classification-oriented variants of the considered pathology report generation models by *i)* retaining the visual processing pipeline, including the visual feature extractor and transformer-based WSI encoder; *ii)* removing the textual input branch and word embedding module; and *iii)* replacing the generative decoder with multiple task-specific classification heads, each predicting a distinct pathological feature. Leveraging the transformer-based WSI encoder design [10, 35], we introduce *learnable task prompts* [19] to derive pathological feature-specific representations from the WSI patch embeddings. Given patch-level embeddings $\mathbf{X} \in \mathbb{R}^{n_{\text{tokens}} \times d_m}$, we prepend a set of learnable task prompts $\mathbf{P} \in \mathbb{R}^{hl_{tp} \times d_m}$, orga-

nized as $\mathbf{P} = [\mathbf{P}_1; \ldots; \mathbf{P}_h]$, where $\mathbf{P}_k \in \mathbb{R}^{l_{tp} \times d_m}$ denotes the prompts associated with the k-th task, each corresponding to a distinct pathological feature, h denotes the number of tasks/-pathological features, l_{tp} the number of prompts per task, and d_m the embedding dimension. The input to the transformer-based WSI encoder is:

$$\mathbf{Z}_0 = [\mathbf{P}; \mathbf{X}] \in \mathbb{R}^{(hl_{tp} + n_{\text{tokens}}) \times d_m}. \tag{1}$$

For each transformer layer ℓ, the token sequence is updated as

$$\mathbf{Z}'_\ell = \mathbf{Z}_{\ell-1} + \text{MSA}(\text{LN}(\mathbf{Z}_{\ell-1})), \qquad \mathbf{Z}_\ell = \mathbf{Z}'_\ell + \text{MLP}\big(\text{LN}(\mathbf{Z}'_\ell)\big), \tag{2}$$

where MSA denotes multi-head self-attention and LN denotes layer normalization [10, 35]. Through self-attention, the task prompts interact with the WSI patch tokens and aggregate task-relevant visual information. Denoting the complete WSI encoder as Enc_{WSI}, its output and contextualised prompts associated with the k-th task are:

$$\widetilde{\mathbf{Z}} = \text{Enc}_{\text{WSI}}(\mathbf{Z}_0), \qquad \widetilde{\mathbf{P}}_k = \widetilde{\mathbf{Z}}_{(k-1)l_{tp}+1:kl_{tp}} \in \mathbb{R}^{l_{tp} \times d_m}. \tag{3}$$

The contextualised prompts are mean-pooled as $\mathbf{z}_k = \frac{1}{l_{tp}} \sum_{j=1}^{l_{tp}} \widetilde{\mathbf{P}}_{k,j} \in \mathbb{R}^{d_m}$ and passed to a task-specific head comprising a two-layer MLP with a ReLU non-linearity, followed by a linear classifier:

$$\mathbf{o}_k = \mathbf{W}_k \, \text{MLP}_k(\mathbf{z}_k) + \mathbf{b}_k, \qquad k = 1, \ldots, h, \tag{4}$$

where $\mathbf{W}_k \in \mathbb{R}^{c_k \times d_m}$, $\mathbf{b}_k \in \mathbb{R}^{c_k}$, and $\mathbf{o}_k \in \mathbb{R}^{c_k}$ denotes the logits for the k-th pathological feature. We do not derive a classification-oriented variant of WSI-LLaVA [20], as the only learnable visual module is the visual projector, which is a simple linear layer. Since the model capacity is therefore primarily determined by the underlying LLM, a classification-oriented comparison would be uninformative. Details about the specific adaptation of each backbone are provided in the Supplementary Material.

Classification Training Objective. Owing to the imbalanced label distributions across tasks and the presence of missing annotations, we optimise each task with a separate weighted cross-entropy loss, with class weights estimated from the training set. These weights compensate for class imbalance by assigning higher penalties to under-represented classes. During training, labels may be available only for a subset of tasks for each sample. Therefore, for each mini-batch and for each task, the corresponding loss is computed only over the subset of samples for which the label of that task is available. The resulting valid task-specific losses are then averaged to obtain the final training objective. This strategy jointly optimises all task heads while ensuring that gradients are derived only from observed annotations and that class imbalance is explicitly accounted for.

4 Experiments

4.1 Implementation Details

We implement experiments in PyTorch v2.7.1 and run them on NVIDIA L40S GPUs with 48GB of memory. For MiGen [5], HistGen [13], and BiGen [39], we use UNI [6] as the frozen visual feature extractor. For WSI-LLaVA [20], we follow its original feature extraction protocols. In all settings, we keep the visual feature extractors frozen. For report generation, we evaluate WSI-LLaVA using its released weights, as it is trained on WSI-Bench.

Table 1: **Experiments conducted on the "WSI-Bench split"**. F1-macro (%) and Recall-macro (%) reported as mean ± std over 10 random seeds except for WSI-LLaVA, which is evaluated once using the released checkpoint. † indicates training with our classification strategy and * denotes $p < 0.05$ against the corresponding generative counterpart. Best results in **bold**.

Model	Histological Type		Histological Grade		Nuclear Pleomorphism		Mitotic Activity		Tubular Formation		LVI	
	F1-score	Recall	F1-score	Recall	F1-score	Recall	F1-score	Recall	F1-score	Recall	F1-score	Recall
MiGen	$60.4_{\pm 4.1}$	$59.7_{\pm 5.3}$	$42.0_{\pm 8.7}$	$42.1_{\pm 6.6}$	$47.4_{\pm 7.5}$	$48.9_{\pm 7.2}$	$43.2_{\pm 10.9}$	$42.9_{\pm 11.6}$	$35.7_{\pm 10.3}$	$34.0_{\pm 10.3}$	$38.8_{\pm 9.7}$	$40.7_{\pm 11.2}$
MiGen†	$\mathbf{60.9}^{*}_{\pm 8.7}$	$\mathbf{62.0}^{*}_{\pm 13.1}$	$\mathbf{65.2}^{*}_{\pm 7.5}$	$\mathbf{65.0}^{*}_{\pm 7.3}$	$\mathbf{62.4}^{*}_{\pm 10.2}$	$\mathbf{61.4}^{*}_{\pm 8.5}$	$\mathbf{56.1}^{*}_{\pm 5.7}$	$\mathbf{57.8}^{*}_{\pm 5.7}$	$\mathbf{49.7}^{*}_{\pm 3.4}$	$\mathbf{52.5}^{*}_{\pm 3.3}$	$\mathbf{67.4}^{*}_{\pm 13.4}$	$\mathbf{70.8}^{*}_{\pm 10.2}$
HistGen	$60.6_{\pm 2.2}$	$59.9_{\pm 2.5}$	$40.8_{\pm 6.2}$	$40.7_{\pm 6.5}$	$44.9_{\pm 7.4}$	$49.8_{\pm 7.8}$	$52.6_{\pm 8.2}$	$53.5_{\pm 8.1}$	$33.6_{\pm 8.1}$	$32.1_{\pm 7.9}$	$38.7_{\pm 12.2}$	$42.9_{\pm 14.5}$
HistGen†	$\mathbf{75.2}^{*}_{\pm 11.5}$	$\mathbf{83.1}^{*}_{\pm 15.1}$	$\mathbf{64.6}^{*}_{\pm 5.3}$	$\mathbf{64.8}^{*}_{\pm 5.6}$	$\mathbf{58.1}^{*}_{\pm 5.0}$	$\mathbf{58.1}^{*}_{\pm 5.1}$	$\mathbf{62.5}^{*}_{\pm 7.4}$	$\mathbf{63.1}^{*}_{\pm 7.1}$	$\mathbf{52.2}^{*}_{\pm 4.2}$	$\mathbf{54.3}^{*}_{\pm 4.6}$	$\mathbf{51.9}^{*}_{\pm 9.7}$	$\mathbf{52.9}^{*}_{\pm 9.7}$
BiGen	$62.9_{\pm 3.3}$	$63.8_{\pm 4.0}$	$40.2_{\pm 7.6}$	$38.0_{\pm 5.9}$	$41.5_{\pm 9.4}$	$43.5_{\pm 8.5}$	$42.4_{\pm 10.8}$	$42.7_{\pm 9.9}$	$34.0_{\pm 3.8}$	$32.6_{\pm 3.6}$	$60.1_{\pm 12.6}$	$62.8_{\pm 12.9}$
BiGen†	$62.4_{\pm 2.3}$	$64.0_{\pm 1.8}$	$\mathbf{68.9}^{*}_{\pm 6.6}$	$\mathbf{67.9}^{*}_{\pm 6.2}$	$\mathbf{65.4}^{*}_{\pm 5.3}$	$\mathbf{64.5}^{*}_{\pm 5.4}$	$\mathbf{62.9}^{*}_{\pm 6.3}$	$\mathbf{64.3}^{*}_{\pm 5.6}$	$\mathbf{51.7}^{*}_{\pm 3.5}$	$\mathbf{53.5}^{*}_{\pm 4.3}$	$\mathbf{64.7}_{\pm 12.0}$	$\mathbf{67.3}_{\pm 10.1}$
$\text{WSI}_{\text{LLaVA}}$	59.1	75.0	48.4	47.9	42.5	41.6	35.1	33.3	63.5	61.9	66.1	70.0

We instead train the encoder–decoder models MiGen, HistGen, and BiGen from scratch on the BRCA subset of WSI-Bench under the conventional next-token prediction objective. For classification, we train the classification-oriented versions of MiGen, HistGen, and BiGen from scratch on PATHOCLASS-BRCA within our multi-task framework. We do not adapt WSI-LLaVA to classification, as discussed in Section 3.3.

For MiGen, HistGen, and BiGen, we strictly follow the original report-generation training protocol, including the optimiser, learning rate, scheduler, and number of epochs, and use the same experimental settings and hyperparameters for both generative and classification training. Specifically, optimisation is performed with Adam using a learning rate of 1×10^{-4} and weight decay of 5×10^{-5}, together with a StepLR scheduler with step size 50 and $\gamma = 0.1$. We set the model dimensionality to $d_{\text{m}} = 512$. For classification, we set $l_{tp} = 1$ unless otherwise stated. Results are averaged over 10 random seeds, from 9233 to 9242, and statistical significance is assessed using one-sided paired t-tests.

4.2 Report Generation vs. Report Classification

We evaluate the clinical accuracy of existing report generation models and compare them with their classification-oriented counterparts. Evaluation is conducted on the original WSI-Bench test split restricted to BRCA, enabling direct comparison with WSI-LLaVA.

Results are reported in Table 1, where MiGen, HistGen, and BiGen denote models trained under the conventional report-generation paradigm and evaluated using our pathological feature-level semantic evaluation protocol described in Section 3.2. Meanwhile, their classification-oriented counterparts—MiGen†, HistGen†, and BiGen†—are trained and evaluated directly within our multi-task classification framework. We also provide conventional captioning metrics for all models in the Supplementary Material.

Under the generation paradigm, all examined report generation models achieve comparable captioning performance on WSI-Bench BRCA, with average scores of BLEU-4 = 0.39, ROUGE-L = 0.55, and METEOR = 0.30. However, our feature-level evaluation shows that lexical similarity does not consistently translate into clinical correctness. The task-specific generative models, MiGen, HistGen, and BiGen, achieve moderate performance on histolog-

Table 2: **Ablation studies on the "WSI-Bench split" with MiGen†.** F1-macro (%) and Recall-macro (%) reported as mean ± std over 10 random seeds. d_m denotes the model dimensionality and l_{tp} the number of learnable task prompts for each task.

Model	d_m	l_{tp}	Histological Type		Histological Grade		Nuclear Pleomorphism		Mitotic Activity		Tubular Formation		LVI	
			F1-score	Recall	F1-score	Recall	F1-score	Recall	F1-score	Recall	F1-score	Recall	F1-score	Recall
MiGen†	512	1	$60.9_{\pm 8.7}$	$62.0_{\pm 13.1}$	$65.2_{\pm 7.5}$	$65.0_{\pm 7.3}$	$62.4_{\pm 10.2}$	$61.4_{\pm 8.5}$	$56.1_{\pm 5.7}$	$57.8_{\pm 5.7}$	$49.7_{\pm 3.4}$	$52.5_{\pm 3.3}$	$67.4_{\pm 13.4}$	$70.8_{\pm 10.2}$
MiGen†	256	1	$68.3_{\pm 8.7}$	$67.0_{\pm 11.4}$	$62.1_{\pm 4.0}$	$61.9_{\pm 3.9}$	$59.7_{\pm 5.9}$	$59.9_{\pm 6.0}$	$60.0_{\pm 8.0}$	$57.4_{\pm 9.1}$	$52.5_{\pm 5.8}$	$52.9_{\pm 5.6}$	$74.1_{\pm 7.2}$	$73.5_{\pm 7.6}$
MiGen†	1024	1	$63.5_{\pm 2.2}$	$62.3_{\pm 2.8}$	$53.9_{\pm 9.5}$	$50.7_{\pm 11.9}$	$60.2_{\pm 4.3}$	$58.7_{\pm 5.3}$	$49.4_{\pm 7.1}$	$50.3_{\pm 6.5}$	$49.1_{\pm 6.9}$	$48.5_{\pm 7.7}$	$73.3_{\pm 9.9}$	$72.1_{\pm 10.5}$
MiGen†	512	2	$60.7_{\pm 3.0}$	$58.4_{\pm 2.3}$	$61.5_{\pm 4.5}$	$61.8_{\pm 4.5}$	$58.4_{\pm 8.7}$	$58.5_{\pm 8.7}$	$53.1_{\pm 5.5}$	$50.2_{\pm 6.8}$	$54.0_{\pm 4.2}$	$53.4_{\pm 1.5}$	$68.3_{\pm 10.8}$	$65.7_{\pm 12.7}$
MiGen†	512	4	$62.5_{\pm 5.5}$	$62.6_{\pm 3.2}$	$56.5_{\pm 3.8}$	$55.4_{\pm 6.1}$	$59.0_{\pm 5.9}$	$55.5_{\pm 9.7}$	$56.7_{\pm 8.5}$	$55.7_{\pm 7.3}$	$52.6_{\pm 2.0}$	$51.4_{\pm 1.5}$	$71.5_{\pm 7.4}$	$68.9_{\pm 10.6}$

ical type but exhibit marked degradation across most remaining pathological features, including histological grade, its constituent components, and LVI. The main exception is BiGen on LVI, where performance is comparatively stronger. Similarly, WSI-LLaVA performs reasonably on histological type, tubular formation and LVI but remains limited on histological grade, nuclear pleomorphism, and mitotic activity. Despite its substantially larger parameter count, WSI-LLaVA does not consistently outperform smaller task-specific generative models in identifying clinically relevant pathological features.

In contrast, the classification-oriented models consistently improve over their corresponding generative counterparts. Averaged across the six tasks, MiGen†, HistGen†, and BiGen† improve F1-score by 15.7, 15.6, and 15.8 percentage points, respectively, and recall by 16.9, 16.2, and 16.4 points. Across models, the largest gains are observed for histological grade (+25.2 F1, +25.6 recall), nuclear pleomorphism (+17.4 F1, +13.9 recall) and mitotic activity (+14.4 F1, +15.4 recall).

Overall, these results provide strong evidence that feature-level evaluation exposes clinically relevant errors that conventional captioning metrics fail to capture, and that directly optimising models for guideline-aligned pathological feature prediction yields more reliable clinical performance than free-text report generation alone.

4.3 Ablation Study

As no backbone consistently dominates across all pathological features, we use MiGen† for ablation studies. Its simpler architecture makes the impact of hyperparameter changes easier to interpret, enabling a clearer assessment of model dimensionality, d_m, and the number of learnable task prompts per task, l_{tp}. This analysis isolates how model capacity and task-conditioning affect pathological feature recognition within the proposed classification framework. Results are reported in Table 2.

We adopt $d_m = 512$ as the default to match the original MiGen report generation setting. Although $d_m = 256$ performs better on several features, $d_m = 512$ achieves the strongest results on histological grade and nuclear pleomorphism while remaining competitive elsewhere, supporting its use as a balanced configuration for fine-grained pathological feature recognition. Regarding the number of task prompts, performance varies across pathological features, indicating that increasing l_{tp} does not provide uniform benefits. We therefore adopt $l_{tp} = 1$ as a simple and effective default configuration.

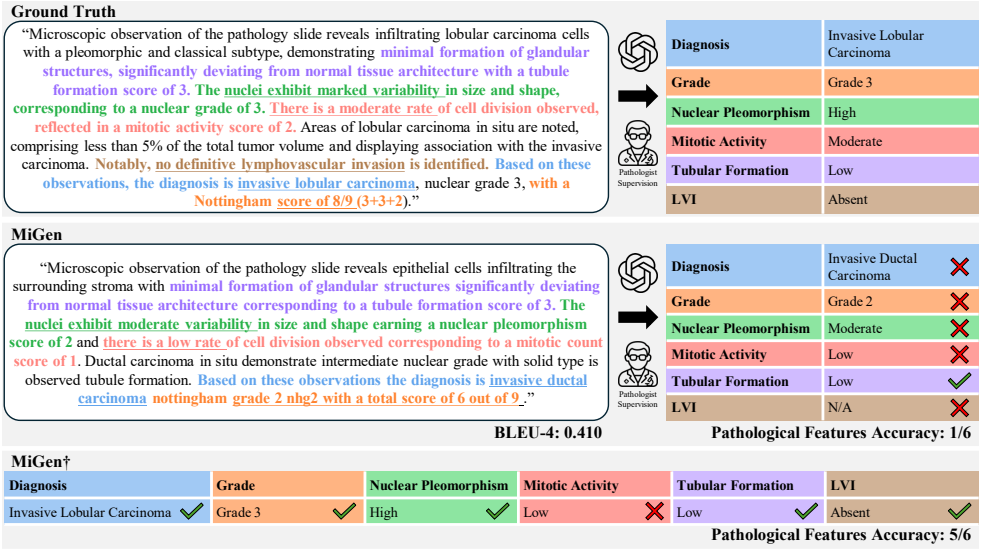

Figure 5: Qualitative example showing the mismatch between lexical overlap and clinical correctness. Although MiGen's report obtains a BLEU-4 score close to the model's average performance, clinically relevant term substitutions and omitted findings result in only $1/6$ correct pathological features. In contrast, the classification-based MiGen† directly predicts structured labels and correctly identifies $5/6$ features for the same WSI.

4.4 Qualitative Analysis and Visualisation

Qualitative Analysis. We provide a qualitative analysis to examine whether captioning metrics reliably capture clinical correctness in pathology report generation. We focus on BLEU-4, a widely used metric based on higher-order n-gram overlap, and analyse a representative report generated by MiGen with a score of 0.41, close to the average BLEU-4 performance of MiGen, HistGen, and BiGen. An overview is shown in Figure 5. Although the generated report largely overlaps with the reference at the lexical level, preserving the same introductory context and differing only in a few underlined terms, these differences are clinically critical. In particular, "invasive lobular carcinoma" is changed to "invasive ductal carcinoma", and a "moderate rate of cell division" is changed to a "low rate of cell division." The generated report also omits relevant findings, including the reported absence of lymphovascular invasion. Thus, small lexical substitutions and omissions can substantially alter the diagnostic interpretation while having limited impact on BLEU-4.

Notably, applying our pathological feature-level extraction pipeline confirms this discrepancy. Compared with the pathologist-verified annotations, the generated report correctly matches only $1/6$ pathological features, indicating that most clinically relevant findings are omitted or incorrectly described despite a BLEU-4 score close to the average performance of current state-of-the-art report generation models. In contrast, the classification-oriented MiGen† correctly predicts $5/6$ features for the same WSI. Together with the quantitative results, this case shows that structured pathological-feature prediction reduces clinically relevant omissions and misclassifications, while feature-level semantic evaluation captures errors that lexical-overlap metrics may miss.

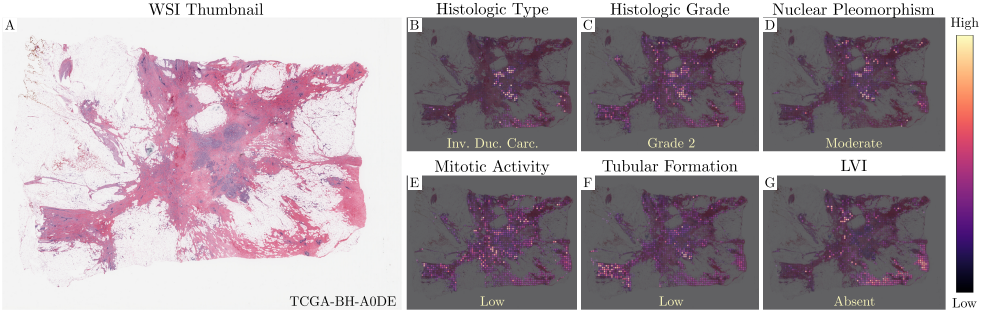

Figure 6: **Visualization of MiGen†. A**: original WSI thumbnail of a test case of the "WSI-Bench split" for which the model correctly predicted all six task-specific labels. **B–G**: attention maps from the learnable task prompts to WSI patches. Correct labels denoted in beige.

This case also highlights the role of clinically grounded prompts in the extraction pipeline. In the reference report, the histological tumour grade is not explicitly stated as grade 1, 2, or 3, but encoded through the Nottingham score and its component breakdown, *e.g.* 8/9 $(3+3+2)$. Incorporating this clinical knowledge enables the extraction pipeline to derive the corresponding structured label rather than relying solely on explicit grade mentions.

Attention Map Visualisation. Task-specific prompts improve interpretability by associating each prompt with a distinct pathological feature, enabling inspection of the WSI regions that contribute to each feature prediction. We visualise task-dependent attention maps over WSI patches to assess whether different prompts attend to distinct slide regions. As shown in Figure 6 for MiGen†, distinct prompts focus on different regions in a case where all six task-specific labels are correctly predicted, indicating specialisation toward feature-relevant morphological patterns.

We further quantify prompt specialisation by measuring cosine similarity between learned task prompts across all models and random seeds. The average similarity remains below 0.18, indicating limited redundancy among prompts and supporting the emergence of distinct task-conditioned representations within the multi-task classification framework.

5 Conclusion

We release PATHOCLASS-BRCA, the first classification-oriented breast cancer WSI dataset that reframes pathology report generation as a clinically grounded multi-task prediction problem. By structuring supervision around six ICCR-guided pathological features, our framework enables objective evaluation of diagnostic correctness using classification metrics rather than linguistic similarity scores. Through systematic comparison of generative and classification-based training paradigms, we quantitatively assess the clinical correctness of conventional report generation models, and we demonstrate that explicit supervision of clinically meaningful features within our proposed classification framework improves the recognition of diagnostically relevant features. Beyond quantitative improvements, our formulation promotes structured reporting aligned with clinical standards and represents a principled step toward clinically reliable AI systems for computational pathology. In future work, we will extend PATHOCLASS-BRCA to additional tumour types, establishing a standardised multi-cancer benchmark for clinically grounded evaluation in computational pathology.

Acknowledgements. This work was supported by the European Union-funded project "IRIS Africa: Intelligent and Responsible Innovation with GenAI for Societal Impact" (IRIS Africa, CUP E93C25003670006).

Disclosure of Interests. The authors have no conflicts of interest to declare.

References

[1] Basit Alawode, Arif Mahmood, Muaz Khalifa Al Radi, et al. MLLM-HWSI: A Multimodal Large Language Model for Hierarchical Whole Slide Image Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13732–13743, 2026.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[3] Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *IEEvaluation@ACL*, 2005.

[4] Cancer Genome Atlas Research Network, John N. Weinstein, Eric A. Collisson, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45:1113–1120, 2013.

[5] Pingyi Chen, Honglin Li, Chenglu Zhu, et al. WsiCaption: Multiple Instance Generation of Pathology Reports for Gigapixel Whole-Slide Images. In *Proceedings of Medical Image Computing and Computer Assisted Intervention*, volume LNCS 15004, 2024.

[6] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

[7] Gábor Cserni. Histological type and typing of breast carcinomas and the WHO classification changes over time. *Pathologica*, 112:25–41, 2020.

[8] Jesse Davis and Mark Goadrich. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, 2006.

[9] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2): 31–71, 1997.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

[11] I Ellis, F Webster, KH Allison, et al. Dataset for reporting of the invasive carcinoma of the breast: recommendations from the International Collaboration on Cancer Reporting (ICCR). *Histopathology*, 85:418–436, 2024.

[12] Yijian Fan, Zhenbang Yang, Rui Liu, Mingjie Li, and Xiaojun Chang. Medical report generation is a multi-label classification problem. *arXiv preprint arXiv:2409.00250*, 2024.

[13] Zhengrui Guo, Jiabo Ma, Yingxue Xu, et al. Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 189–199, 2024.

[14] Lei He, L. Rodney Long, Sameer Antani, et al. Histology Image Analysis for Carcinoma Detection and Grading. *Computer Methods and Programs in Biomedicine*, 107: 538–556, 2012.

[15] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, et al. A visual–language foundation model for pathology image analysis using medical Twitter. *Nature Medicine*, pages 1–10, 2023.

[16] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[17] Natthawadee Laokulrath, Mihir Ananta Gudi, Rahul Deb, Ian O Ellis, and Puay Hoon Tan. Invasive breast cancer reporting guidelines: ICCR, CAP, RCPath, RCPA datasets and future directions. *Diagnostic Histopathology*, 30:87–99, 2024.

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[19] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing*, pages 4582–4597, 2021.

[20] Yuci Liang, Xinheng Lyu, Wenting Chen, et al. WSI-LLaVA: A Multimodal Large Language Model for Whole Slide Image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22718–22727, October 2025.

[21] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Annual Meeting of the Association for Computational Linguistics*, 2004.

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, et al. Visual Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[24] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.

[25] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, et al. A multimodal generative AI copilot for human pathology. *Nature*, 634(8033), 2024.

[26] Jamshed Memon, Maira Sami, Rizwan Khan, et al. Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR). *IEEE Access*, pages 1–1, 2020.

[27] Stacey E Mills. *Histology for pathologists*. 2012.

[28] Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, 2021.

[29] Kishore Papineni, Salim Roukos, Todd Ward, et al. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Annual Meeting of the Association for Computational Linguistics*, 2002.

[30] Qwen et al. Qwen2.5 Technical Report, 2025.

[31] Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. 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*, 2024.

[32] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.*, 45, 2009.

[33] Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. LLaMA: Open and Efficient Foundation Language Models, 2023.

[34] L. J. Tseng, A. Matsuyama, and V. MacDonald-Dickinson. Histology: The Gold Standard for Diagnosis? *Canadian Veterinary Journal*, 64:389–391, 2023.

[35] Ashish Vaswani et al. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30, 2017.

[36] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[37] Tong Wang, Su-Jin Shin, Mingkang Wang, Qi Xu, Guiyang Jiang, Fengyu Cong, Jeonghyun Kang, and Hongming Xu. Multi-task adaptive resolution network for lymph node metastasis diagnosis from whole slide images of colorectal cancer. *IEEE Journal of Biomedical and Health Informatics*, 29(1):420–432, 2024.

[38] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 2024.

[39] Ling Zhang, Boxiang Yun, Qingli Li, and Yan Wang. Historical Report Guided Bimodal Concurrent Learning for Pathology Report Generation. In *Proceedings of Medical Image Computing and Computer Assisted Intervention*, 2025.

[40] Qifeng Zhou, Wenliang Zhong, Yuzhi Guo, Michael Xiao, Hehuan Ma, and Junzhou Huang. Pathm3: A multimodal multi-task multiple instance learning framework for whole slide image classification and captioning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 373–383, 2024.

[41] Deyao Zhu, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pages 18378–18394, 2024.

Supplementary Material for PathoClass-BRCA: Reframing Pathology Report Generation as Guideline-Aligned Multi-Task Classification

Alessia Saporita[1, 2, *]
alessia.saporita@unimore.it
Vittorio Pipoli[1, *]
vittorio.pipoli@unimore.it
Federico Bolelli[1, ✉]
federico.bolelli@unimore.it
Andrea Acquaviva[2]
andrea.acquaviva@unibo.it
Elisa Ficarra[1]
elisa.ficarra@unimore.it

[1] "Enzo Ferrari" Engineering Department, University of Modena and Reggio Emilia, Modena, Italy
[2] Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi," University of Bologna, Bologna, Italy

1 Pathological Feature Selection and Annotation

1.1 ICCR-Guided Feature Selection

In consultation with pathologists, we selected breast cancer pathological features that are defined as core elements in the ICCR guidelines and are both clinically relevant and visually assessable from WSIs. We then applied a minimum coverage threshold of 50%, excluding features documented in fewer than half of the pathology reports. The coverage of all candidate features identified in consultation with pathologists is reported in Table 3. Specifically, ductal carcinoma *in situ*, calcifications, and necrosis were excluded because their coverage was below 50%, whereas the retained features exhibited coverage ranging from 83% to 100%. This selection strategy preserves a set of clinically meaningful features that is sufficient to assess the limitations of free-text report generation, while reducing the risk that the evaluation is disproportionately influenced by sparse label availability.

1.2 Feature-Specific Labeling Prompts

We provide the complete prompts used for feature-specific label extraction from free-text pathology reports below. Each prompt is designed to convert unstructured report text into a standardized label for a predefined pathological feature. All prompts follow a shared structure defining the target feature to be extracted, the set of admissible output classes, and the

* Equal contribution. Authors may list their names first in their respective CVs. ✉ Corresponding author.

normalization rules used to map heterogeneous report terminology to standardized labels. They also specify conservative decision criteria for label assignment, including rules for resolving conflicting mentions, and explicitly handle missing evidence, unsupported statements, and ambiguous descriptions, typically by assigning an `Unspecified` label. When required, task-specific scope rules restrict extraction to findings associated with the main invasive carcinoma. Each prompt returns a standardized JSON object containing the extracted label, supporting textual evidence, and a confidence score. These outputs were used as preliminary annotations and were subsequently reviewed by pathologists to obtain the final ground-truth labels used in PATHOCLASS-BRCA.

2 Classification-Oriented Model Adaptation

We provide model-specific details on the classification-oriented adaptations, focusing on how learnable task prompts are incorporated into each architecture to extract pathological feature representations for the corresponding classification heads.

MiGen. For MiGen [1], the adaptation directly instantiates the general formulation described in the main paper. Learnable task prompts are prepended to the WSI patch embeddings and processed by the original transformer encoder. The resulting task-specific token representations are then passed to the corresponding classification heads.

HistGen. For HistGen [2], which combines regional and global attention mechanisms, we apply the same adaptation principle at the regional level. Specifically, each region is augmented with learnable task prompts, enabling the model to capture feature-relevant evidence within local tissue regions. The resulting regional representations are aggregated through the original attention-pooling module to obtain one representation per pathological feature, which is then passed to the corresponding classification heads.

BiGen. For BiGen [4], we preserve the original bi-modal visual–textual encoding strategy while adapting it for classification. In the original architecture, a visual latent token attends to WSI patch embeddings, ranks patches according to the resulting attention scores, and guides the retrieval of semantically related textual embeddings from an external text bank. A textual latent token subsequently attends to the retrieved embeddings, and the resulting visual and textual representations are concatenated to condition autoregressive report decoding. In our classification-oriented variant, we replace the single visual and textual latent tokens with task-specific visual and textual query tokens. For each pathological feature, the visual query tokens aggregate relevant evidence from the WSI, whereas the textual query tokens incorporate contextual information from the retrieved text embeddings. The corresponding visual and textual outputs are concatenated to form a bi-modal feature representation, which is forwarded to the associated classification head.

3 Additional Experimental Results

3.1 Captioning Performance on BRCA

Table 1 reports captioning performance on the BRCA subset of WSI-Bench [3]. Overall, the three task-specific report generation models achieve comparable performance, with relatively small differences across the evaluated metrics. HistGen obtains the highest BLEU-1 and BLEU-2 scores, as well as the highest METEOR score, indicating stronger lexical and

Table 1: **Captioning performance on WSI-Bench BRCA [3]**. Results are reported using BLEU-1–4, METEOR, and ROUGE-L. MiGen [1], HistGen [2], and BiGen [4] are evaluated over 10 seeds (9233 to 9242), with results reported as mean ± standard deviation, while the pathology MLLM WSI-LLaVA is evaluated once. Best results are highlighted in bold.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
MiGen	$0.632_{\pm 0.02}$	$0.526_{\pm 0.03}$	$0.456_{\pm 0.02}$	$0.403_{\pm 0.03}$	$0.302_{\pm 0.01}$	$0.562_{\pm 0.02}$
HistGen	$\mathbf{0.640}_{\pm 0.01}$	$\mathbf{0.535}_{\pm 0.01}$	$0.464_{\pm 0.01}$	$0.410_{\pm 0.01}$	$\mathbf{0.306}_{\pm 0.01}$	$0.577_{\pm 0.01}$
BiGen	$0.625_{\pm 0.01}$	$0.530_{\pm 0.01}$	$\mathbf{0.465}_{\pm 0.01}$	$\mathbf{0.414}_{\pm 0.01}$	$0.305_{\pm 0.01}$	$\mathbf{0.582}_{\pm 0.01}$
WSI-LLaVA	0.537	0.437	0.373	0.324	0.291	0.484

Table 2: **Classification performance on WSI-Bench LUNG**. Macro F1-score (%) is reported as mean ± standard deviation over 10 seeds (9233 to 9242) for MiGen [1], HistGen [2], and BiGen [4]. † denotes our classification version, and * indicates $p < 0.05$. Best results are highlighted in bold.

Model	Histological Type	Histological Grade	Nuclear Pleomorphism	LVI	Necrosis	Mean
MiGen	$88.4_{\pm 6.5}$	$39.1_{\pm 4.6}$	$38.7_{\pm 10.0}$	$38.5_{\pm 12.1}$	$31.9_{\pm 5.8}$	47.3
MiGen†	$\mathbf{95.0}^{*}_{\pm 4.7}$	$\mathbf{60.6}^{*}_{\pm 14.4}$	$\mathbf{67.0}^{*}_{\pm 7.9}$	$\mathbf{45.5}_{\pm 10.9}$	$\mathbf{64.3}^{*}_{\pm 12.6}$	**66.5**
HistGen	$73.3_{\pm 16.1}$	$53.4_{\pm 15.6}$	$44.1_{\pm 5.3}$	$35.2_{\pm 9.0}$	$25.8_{\pm 6.1}$	46.3
HistGen†	$\mathbf{95.5}^{*}_{\pm 3.7}$	$\mathbf{56.5}_{\pm 13.5}$	$\mathbf{70.6}^{*}_{\pm 10.4}$	$\mathbf{46.5}^{*}_{\pm 10.2}$	$\mathbf{63.7}^{*}_{\pm 7.5}$	**66.6**
BiGen	$87.9_{\pm 4.2}$	$58.1_{\pm 15.2}$	$41.5_{\pm 8.0}$	$34.0_{\pm 7.0}$	$24.6_{\pm 8.2}$	49.2
BiGen†	$\mathbf{95.5}^{*}_{\pm 2.4}$	$\mathbf{63.6}_{\pm 16.2}$	$\mathbf{70.6}^{*}_{\pm 5.6}$	$\mathbf{49.7}^{*}_{\pm 7.5}$	$\mathbf{47.2}^{*}_{\pm 10.5}$	**65.2**
$\text{WSI}_{\text{LLaVA}}$	95.0	62.5	36.8	65.9	20.8	56.2

short-range phrase overlap with the reference reports. Conversely, BiGen achieves the best BLEU-3, BLEU-4, and ROUGE-L scores, suggesting a slight advantage in preserving longer textual patterns and report-level sequence similarity. In contrast, WSI-LLaVA achieves lower performance than the task-specific report generation models across all captioning metrics. This highlights that specialized architectures can achieve competitive or superior captioning performance with substantially fewer parameters.

3.2 Generalisation to Lung Cancer

Our framework is designed to generalize across tumour types through a standardized procedure: for each cancer, pathologists identify clinically relevant and WSI-assessable features according to the corresponding ICCR dataset or other authoritative pathology reporting guidelines. Our pipeline then supports the annotation process, and pathologists ultimately review and validate the resulting labels. To provide preliminary evidence of such generalizability, we extend our analysis beyond breast cancer to lung cancer by combining the TCGA-LUAD and TCGA-LUSC cohorts and selecting the corresponding guideline-aligned pathological features in consultation with pathologists. Specifically, we evaluate histologic type, histologic grade, nuclear pleomorphism, lymphovascular invasion (LVI), and necrosis. Results are reported in Table 2. They demonstrate a consistent benefit of adapting the report-generation architectures to our classification-oriented setting. For all three task-specific models, the classification-oriented variants substantially improve the average macro

F1-score over their original report-generation counterparts across most pathological features, with particularly large gains for histologic type, nuclear pleomorphism, and necrosis. The classification-oriented variants also remain competitive with the pathology MLLM WSI-LLaVA, although WSI-LLaVA achieves the highest performance on LVI. Overall, these experiments provide preliminary empirical evidence that the proposed classification framework can be transferred across tumour types, supporting its broader applicability across cancer types and pathology reporting tasks.

histological grade labelling prompt

Role.
You are a pathology information extraction assistant.

Task Definition.
Extract the histological tumour grade from a breast pathology report.

Output Format.
You MUST return valid JSON only, with keys:
- `grade`: one string from the allowed classes;
- `evidence`: a short quote copied from the report;
- `confidence`: a float in `[0,1]`.

Conservative Rule.
If multiple grade expressions are present, including Nottingham grade, total score, differentiation wording, or grade ranges, select the highest grade mentioned.

Normalization Rules.
Map grade expressions as follows.
- `Grade 1` includes Grade 1, Grade I, NHG1, Nottingham grade I, Nottingham histological Grade 1, well differentiated, and low grade.
- `Grade 2` includes Grade 2, Grade II, NHG2, Nottingham grade II, Nottingham histological Grade 2, moderately differentiated, and intermediate grade.
- `Grade 3` includes Grade 3, Grade III, NHG3, Nottingham grade III, Nottingham histological Grade 3, poorly differentiated, high grade, and poor histological grade.

Nottingham Score Rule.
If a Nottingham total score is given, map score 3–5 to `Grade 1`, score 6–7 to `Grade 2`, and score 8–9 to `Grade 3`. If an explicit grade and a Nottingham score conflict, choose the highest grade implied.

Range Rule.
For ranges or composite grades, such as I/II, II/III, 1–2, 2–3, or 2/3, choose the highest grade mentioned.

Additional Rules.
- Choose exactly one grade.
- Prefer explicit histological grade statements when present.
- If the text states that grade cannot be assessed, or if no grade information is present, return `grade="Unspecified"` and `confidence=0.0`.

Strict Constraints.
- Do not invent a grade.
- Do not infer grade from unrelated tumour features.
- Return JSON only.

Report text: `{report_text}`

Allowed Classes. `Grade 1; Grade 2; Grade 3; Unspecified`

Diagnosis labelling prompt

Role.
You are a pathology information extraction assistant.

Task Definition.
Classify the primary invasive diagnosis from a breast pathology report.

Output Format.
You MUST return valid JSON only, with keys:
- `diagnosis`: one string from the allowed classes;
- `components`: an array of strings; MUST be empty unless `diagnosis` is `"mixed carcinoma"`;
- `evidence`: a short quote copied from the report;
- `confidence`: a float in `[0,1]`.

Conservative Rule.
If the report describes an invasive carcinoma without explicitly naming a subtype, but mentions features of a specific subtype, classify the tumour according to that subtype. For example, `"invasive carcinoma with lobular features"` should be classified as `"invasive lobular carcinoma"`.

Mixed Carcinoma Rule.
Use `"mixed carcinoma"` only when the report explicitly states more than one invasive subtype or mixed invasive differentiation, such as ductal and lobular carcinoma, mixed ductal and lobular features, mixed differentiation, or tubulolobular carcinoma. If `diagnosis` is `"mixed carcinoma"`, `components` must contain only the invasive subtypes explicitly stated in the report and selected from the allowed components. Otherwise, `components` must be empty.

Rare/Special Subtype Rule.
If a rare or special invasive subtype is named and is not one of the listed diagnoses, return `"rare carcinoma subtype"`, unless it is explicitly mixed with another listed invasive subtype.

Additional Rules.
- Choose exactly one diagnosis.
- Prefer the final diagnosis statement when present.
- Do not classify DCIS, LCIS, benign lesions, or precursor lesions as the primary invasive diagnosis.

Strict Constraints.
- Do not invent diagnoses or components.
- Do not infer subtype from grade, receptor status, or treatment information.
- Return JSON only.

Report text: `{report_text}`

Allowed Classes. `invasive ductal carcinoma; invasive lobular carcinoma; mucinous carcinoma; metaplastic carcinoma; tubular carcinoma; invasive papillary carcinoma; medullary carcinoma; rare carcinoma subtype; invasive carcinoma (unspecified); mixed carcinoma`

Allowed Components. `invasive ductal carcinoma; invasive lobular carcinoma; mucinous carcinoma; metaplastic carcinoma; tubular carcinoma; invasive papillary carcinoma; medullary carcinoma; rare carcinoma subtype; invasive carcinoma (unspecified)`

Nuclear pleomorphism labelling prompt

Role.
You are a pathology information extraction assistant.

Task Definition.
Extract the nuclear pleomorphism class associated exclusively with the main invasive carcinoma diagnosis from a breast pathology report.

Scope Rule.
Ignore nuclear pleomorphism related to DCIS, LCIS, other in situ lesions, benign lesions, precursor lesions, or secondary findings.

Output Format.
You MUST return valid JSON only, with keys:
- `nuclear_pleomorphism`: one string from the allowed classes;
- `evidence`: a short quote copied from the report;
- `confidence`: a float in `[0,1]`.

Conservative Rule.
If multiple nuclear pleomorphism expressions are present for the main invasive diagnosis, select the highest nuclear pleomorphism class.

Normalization Rules.
- `nuclear pleomorphism (high)` includes nuclear pleomorphism score 3, nuclear score 3, nuclear grade 3, marked nuclear pleomorphism, marked variability in nuclear size and shape, and high nuclear grade.
- `nuclear pleomorphism (moderate)` includes nuclear pleomorphism score 2, nuclear score 2, nuclear grade 2, moderate pleomorphism, and intermediate nuclear variability.
- `nuclear pleomorphism (low)` includes nuclear pleomorphism score 1, nuclear score 1, nuclear grade 1, mild pleomorphism, and uniform nuclei.
- `nuclear pleomorphism (unspecified)` is used only when nuclear features are mentioned without severity or score.

Nottingham Score Rule.
If a Nottingham component explicitly reports a nuclear score, map score 1 to `nuclear pleomorphism (low)`, score 2 to `nuclear pleomorphism (moderate)`, and score 3 to `nuclear pleomorphism (high)`.

Range Rule.
For ranges or composite scores/classes, such as 1–2, 2–3, 1/2, 2/3, low-to-moderate, or moderate-to-high, choose the highest class implied.

Additional Rules.
- Choose exactly one class.
- Prefer explicit nuclear scoring statements linked to the main invasive diagnosis.
- If nuclear features cannot be assessed, or if no valid nuclear pleomorphism information is present, return `nuclear pleomorphism (unspecified)` with `confidence=0.0`.

Strict Constraints.
- Do not invent information.
- Do not infer nuclear pleomorphism from tumour grade unless a nuclear score or explicit pleomorphism statement is provided.
- Return JSON only.

Report text: `{report_text}`

Allowed Classes. `nuclear pleomorphism (high); nuclear pleomorphism (moderate); nuclear pleomorphism (low); nuclear pleomorphism (unspecified)`

Mitotic activity labelling prompt

Role.
You are a pathology information extraction assistant.

Task Definition.
Extract the mitotic activity class associated exclusively with the main invasive carcinoma diagnosis from a breast pathology report.

Scope Rule.
Ignore mitotic activity related to DCIS, LCIS, other in situ lesions, benign or precursor lesions, and secondary or incidental findings.

Output Format.
You MUST return valid JSON only, with keys:
- `mitotic_activity`: one string from the allowed classes;
- `evidence`: a short quote copied from the report;
- `confidence`: a float in `[0,1]`.

Conservative Rule.
If multiple mitotic activity expressions are present for the main invasive diagnosis, select the highest mitotic activity class.

Normalization Rules.
- `mitotic activity (high)` includes mitotic score 3, high mitotic rate, marked mitotic activity, brisk mitoses, numerous mitoses, many mitoses, frequent mitoses, and high rate of cell division.
- `mitotic activity (moderate)` includes mitotic score 2, moderate mitotic rate, intermediate mitotic activity, some mitoses, and occasional mitoses.
- `mitotic activity (low)` includes mitotic score 1, low mitotic rate, rare mitoses, few mitoses, scant mitoses, and minimal mitotic activity.
- `mitotic activity (unspecified)` is used only when mitotic activity is mentioned but cannot be mapped to high, moderate, or low.

Nottingham Score Rule.
If a Nottingham component explicitly states a mitotic score, map score 1 to `mitotic activity (low)`, score 2 to `mitotic activity (moderate)`, and score 3 to `mitotic activity (high)`.

Range Rule.
For ranges or composite scores/classes, such as 1–2, 2–3, 1/2, 2/3, low-to-moderate, or moderate-to-high, choose the highest class implied.

Additional Rules.
- Choose exactly one class.
- Prefer mitotic activity statements clearly linked to the main invasive diagnosis.
- If mitotic activity cannot be assessed, or if no valid mitotic activity information is present, return `mitotic activity (unspecified)` with `confidence=0.0`.

Strict Constraints.
- Do not invent information.
- Do not infer mitotic activity from tumour grade unless a mitotic score or explicit mitotic activity statement is provided.
- Return JSON only.

Report text: `{report_text}`

Allowed Classes. `mitotic activity (high); mitotic activity (moderate); mitotic activity (low); mitotic activity (unspecified)`

Tubular formation labelling prompt

Role.
You are a pathology information extraction assistant.

Task Definition.
Extract the tubular formation class associated exclusively with the main invasive carcinoma diagnosis from a breast pathology report.

Scope Rule.
Ignore tubular formation related to DCIS, LCIS, other in situ lesions, benign glands, normal tissue, precursor lesions, or secondary findings.

Output Format.
You MUST return valid JSON only, with keys:
- `tubular_formation`: one string from the allowed classes;
- `evidence`: a short quote copied from the report;
- `confidence`: a float in `[0,1]`.

Conservative Rule.
If multiple tubule or gland formation expressions are present for the main invasive diagnosis, select the lowest degree of formation, corresponding to the worst differentiation class.

Normalization Rules.
- `tubular formation (high)` includes tubule formation score 1, predominant gland formation, well-formed glands, prominent tubule formation, and well differentiated.
- `tubular formation (moderate)` includes tubule formation score 2, moderate gland formation, partially formed glands, and intermediate differentiation.
- `tubular formation (low/absent)` includes tubule formation score 3, minimal glandular structures, lack of gland formation, solid sheets of tumour cells, no tubules, and poorly differentiated.
- `tubular formation (unspecified)` is used only when gland/tubule formation is not mentioned, unclear, or mentioned without a clear degree or score.

Nottingham Score Rule.
If a Nottingham component explicitly reports a tubular score, map score 1 to `tubular formation (high)`, score 2 to `tubular formation (moderate)`, and score 3 to `tubular formation (low/absent)`.

Range Rule.
For ranges or composite scores/classes, such as score 2–3, or moderate-to-low/absent, choose the lowest formation class, corresponding to the worst differentiation.

Additional Rules.
- Choose exactly one class.
- Prefer explicit tubular formation statements linked to the main invasive diagnosis.
- If gland/tubule formation cannot be assessed, or if no valid information is present, return `tubular formation (unspecified)` with `confidence=0.0`.

Strict Constraints.
- Do not invent information.
- Do not infer tubular formation from tumour grade unless a tubule formation score or explicit gland/tubule formation statement is provided.
- Return JSON only.

Report text: `{report_text}`

Allowed Classes. `tubular formation (high); tubular formation (moderate); tubular formation (low/absent); tubular formation (unspecified)`

Lymphovascular invasion labelling prompt

Role.
You are a pathology information extraction assistant.

Task Definition.
Extract the lymphovascular invasion class associated exclusively with the main invasive carcinoma diagnosis from a breast pathology report.

Scope Rule.
Ignore lymphovascular invasion statements associated with DCIS, LCIS, other in situ lesions, benign lesions, precursor lesions, or secondary findings.

Output Format.
You MUST return valid JSON only, with keys:
- `lymphovascular_invasion`: one string from the allowed classes;
- `evidence`: a short quote copied from the report;
- `confidence`: a float in `[0,1]`.

Conservative Rule.
If both present and absent statements are found for the main invasive diagnosis, or if there are mixed expressions, select "`lymphovascular invasion (present)`".

Normalization Rules.
- `lymphovascular invasion (present)` includes lymphovascular invasion present, LVI present, vascular invasion present, tumour emboli in lymphovascular spaces, tumour cells within lymphatic or vascular spaces, angioinvasion present, lymphatic invasion, vascular permeation, and intravasation identified.
- `lymphovascular invasion (absent)` includes lymphovascular invasion not present, no lymphovascular invasion, LVI absent, LVI negative, no definite lymphovascular invasion, vascular invasion not present, no vascular invasion, and no angioinvasion.
- `lymphovascular invasion (unspecified)` is used when there is no valid statement for the main invasive diagnosis, when the mention is vague, or when the report states uncertainty.

Ambiguity Rule.
If the report states uncertainty, such as cannot be assessed, indeterminate, suspicious for, or equivocal, return `lymphovascular invasion (unspecified)` with `confidence <= 0.4`.

Additional Rules.
- Choose exactly one class. Prefer explicit lymphovascular invasion statements linked to the main invasive diagnosis.
- If no valid lymphovascular invasion information is present, return `lymphovascular invasion (unspecified)` with `confidence=0.0`.

Strict Constraints.
- Do not use DCIS-associated information.
- Do not invent information.
- Do not infer lymphovascular invasion from tumour type, grade, stage, lymph-node status, or metastasis.
- Return JSON only.

Report text: `{report_text}`

Allowed Classes. `lymphovascular invasion (absent); lymphovascular invasion (present); lymphovascular invasion (unspecified)`

Table 3: **Label distribution and feature coverage in WSI-Bench BRCA [3]**. For each pathological feature, the table reports the number and percentage of samples assigned to each class, together with the overall feature coverage. For histologic type, invasive carcinoma (unspecified) is grouped with invasive ductal carcinoma, while less frequent histologic subtypes are grouped into the *Other* category. Unspecified labels correspond to features not explicitly documented in the pathology report.

Feature	Label	Count	Percentage (%)	Coverage (%)
Histological type	Invasive ductal / unspecified carcinoma	823	73.42	100.00
	Invasive lobular carcinoma	194	17.31	
	Other	104	9.28	
Histological grade	Grade 1	97	8.65	90.19
	Grade 2	476	42.46	
	Grade 3	438	39.07	
	Unspecified	110	9.81	
Nuclear pleomorphism	High	455	40.59	94.83
	Moderate	540	48.17	
	Low	68	6.07	
	Unspecified	58	5.17	
Mitotic activity	High	371	33.10	87.07
	Moderate	349	31.13	
	Low	256	22.84	
	Unspecified	145	12.93	
Tubular formation	High	29	2.59	94.65
	Moderate	259	23.10	
	Low/absent	773	68.96	
	Unspecified	60	5.35	
LVI	Present	348	31.04	82.69
	Absent	579	51.65	
	Unspecified	194	17.31	
Ductal carcinoma in situ	Present	465	41.48	44.25
	Absent	31	2.77	
	Unspecified	625	55.75	
Necrosis	Present	186	16.59	33.81
	Absent	193	17.22	
	Unspecified	742	66.19	
Calcifications	Present	135	12.04	26.85
	Absent	166	14.81	
	Unspecified	820	73.15	

References

[1] Pingyi Chen, Honglin Li, Chenglu Zhu, et al. WsiCaption: Multiple Instance Generation of Pathology Reports for Gigapixel Whole-Slide Images. In *Proceedings of Medical Image Computing and Computer Assisted Intervention*, volume LNCS 15004, 2024.

[2] Zhengrui Guo, Jiabo Ma, Yingxue Xu, et al. Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 189–199, 2024.

[3] Yuci Liang, Xinheng Lyu, Wenting Chen, et al. WSI-LLaVA: A Multimodal Large Language Model for Whole Slide Image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22718–22727, October 2025.

[4] Ling Zhang, Boxiang Yun, Qingli Li, and Yan Wang. Historical Report Guided Bi-modal Concurrent Learning for Pathology Report Generation. In *Proceedings of Medical Image Computing and Computer Assisted Intervention*, 2025.