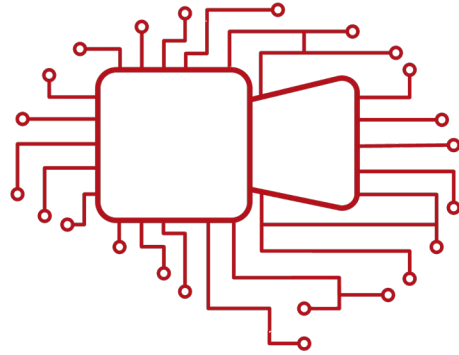




BMVC
2026

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

A. Saporita, V. Pipoli, F. Bolelli, A. Acquaviva, E. Ficarra

Motivation

Existing pathology report generation approaches inherit training objectives and evaluation protocols from **image captioning** which measure and reward **lexical agreement**:

- *Training: token-level cross-entropy* → rewards surface-form similarity to reference reports.
- *Evaluation: captioning metrics (BLEU, ROUGE, METEOR)* → quantify lexical agreement.

However, lexical overlap does not reflect clinical correctness.

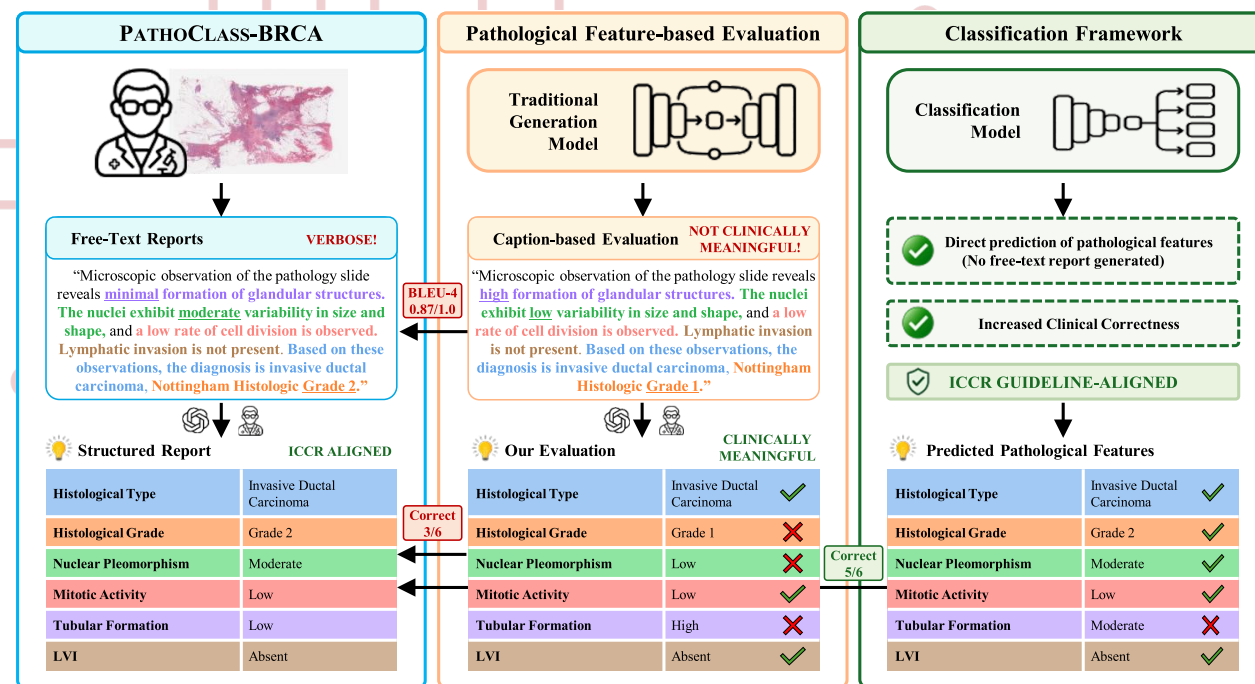
In our figure:

- a report with high BLEU-4 correctly identifies only 3/6 pathological features;
- “invasive *ductal*” → “invasive *lobular*”: negligible BLEU impact, opposite diagnosis.

Moreover, the target output is not open-ended text.

International reporting standards (ICCR) standardize reporting: fixed fields, constrained value sets.

⇒ the diagnostically relevant content = a small set of **categorical attributes**.

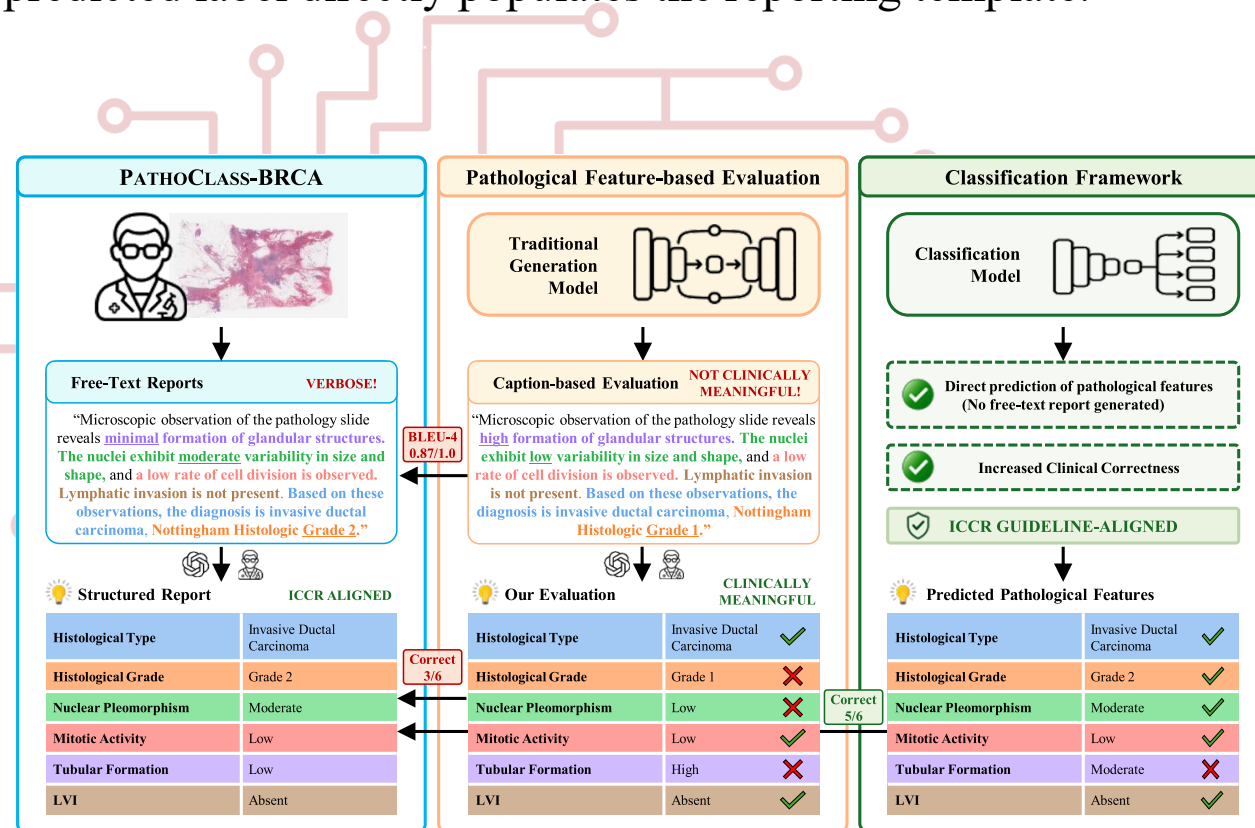


Our Proposal

We reformulate pathology reporting as a clinically grounded **multi-task classification problem**: each ICCR field becomes a classification task whose predicted label directly populates the reporting template.

Our contributions:

1. **PathoClass-BRCA**: the first breast cancer benchmark of guideline-aligned pathological features;
2. **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;
3. **Classification-oriented adaptations** of traditional report generation backbones.



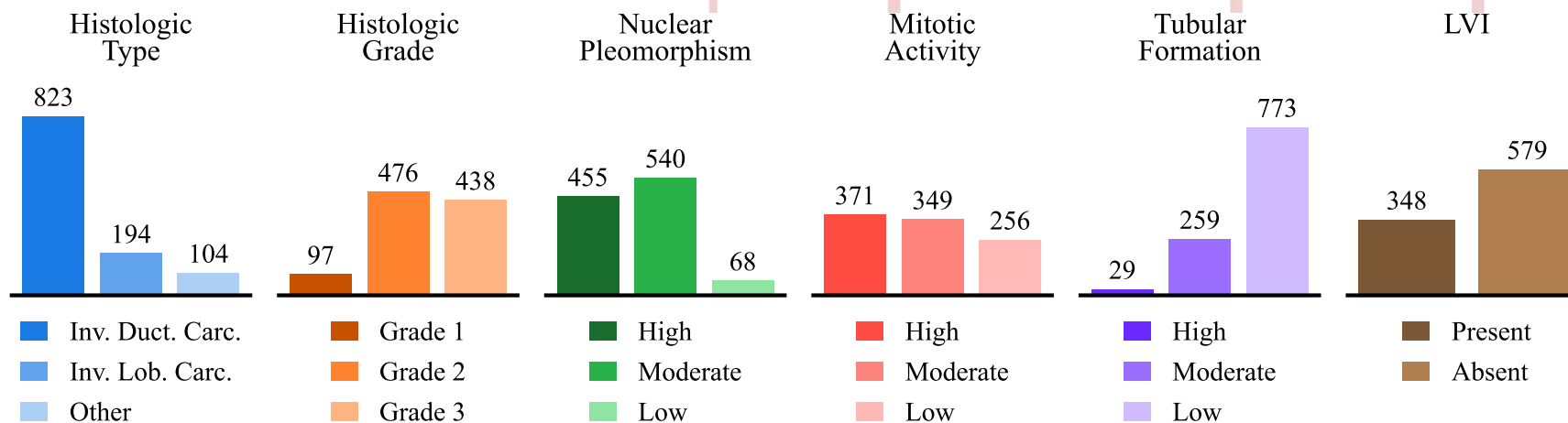
PathoClass-BRCA

Built on the TCGA-BRCA pathology reports of the WSI-Bench split.

- We select **six ICCR-guided pathological features** — clinically relevant, visually assessable from WSIs, reported in at least 50% of the reports:

histological type · histological grade · tubular formation · nuclear pleomorphism · mitotic activity · LVI

- We design a dedicated **feature-specific labelling prompt** which defines the admissible classes and the normalisation rules;
- We prompt GPT-5 to extract a preliminary label, then **pathologists review and correct it** — 96.7% agreement.



Our Semantic Evaluation Protocol

It evaluates generated reports on **clinical content** rather than on lexical agreement:

- apply the same feature-specific labelling prompts to each generated report;
- match the extracted labels against the PathoClass-BRCA ground truth;
- aggregate them into classification metrics (F1-macro, Recall-macro).

Validated by pathologists on 1,500 generated reports: **96.4% concordance** with the intended feature values

Ground Truth

“Microscopic observation of the pathology slide reveals infiltrating lobular carcinoma cells with a pleomorphic and classical subtype, demonstrating **minimal formation of glandular structures, significantly deviating from normal tissue architecture with a tubule formation score of 3. The nuclei exhibit marked variability in size and shape, corresponding to a nuclear grade of 3. There is a moderate rate of cell division observed, reflected in a mitotic activity score of 2.** Areas of lobular carcinoma in situ are noted, comprising less than 5% of the total tumor volume and displaying association with the invasive carcinoma. **Notably, no definitive lymphovascular invasion is identified. Based on these observations, the diagnosis is invasive lobular carcinoma,** nuclear grade 3, **with a Nottingham score of 8/9 (3+3+2).**”







Pathologist Supervision

Diagnosis	Invasive Lobular Carcinoma
Grade	Grade 3
Nuclear Pleomorphism	High
Mitotic Activity	Moderate
Tubular Formation	Low
LVI	Absent

MiGen

“Microscopic observation of the pathology slide reveals epithelial cells infiltrating the surrounding stroma with **minimal formation of glandular structures significantly deviating from normal tissue architecture corresponding to a tubule formation score of 3. The nuclei exhibit moderate variability in size and shape earning a nuclear pleomorphism score of 2 and there is a low rate of cell division observed corresponding to a mitotic count score of 1.** Ductal carcinoma in situ demonstrate intermediate nuclear grade with solid type is observed tubule formation. **Based on these observations the diagnosis is invasive ductal carcinoma nottingham grade 2 nhg2 with a total score of 6 out of 9.**”







Pathologist Supervision

Diagnosis	Invasive Ductal Carcinoma	✗
Grade	Grade 2	✗
Nuclear Pleomorphism	Moderate	✗
Mitotic Activity	Low	✗
Tubular Formation	Low	✓
LVI	N/A	✗

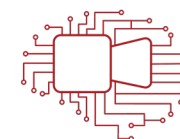
BLEU-4: 0.410

Pathological Features Accuracy: 1/6

MiGen†

Diagnosis	Grade	Nuclear Pleomorphism	Mitotic Activity	Tubular Formation	LVI
Invasive Lobular Carcinoma ✓	Grade 3 ✓	High ✓	Low ✗	Low ✓	Absent ✓

Pathological Features Accuracy: 5/6

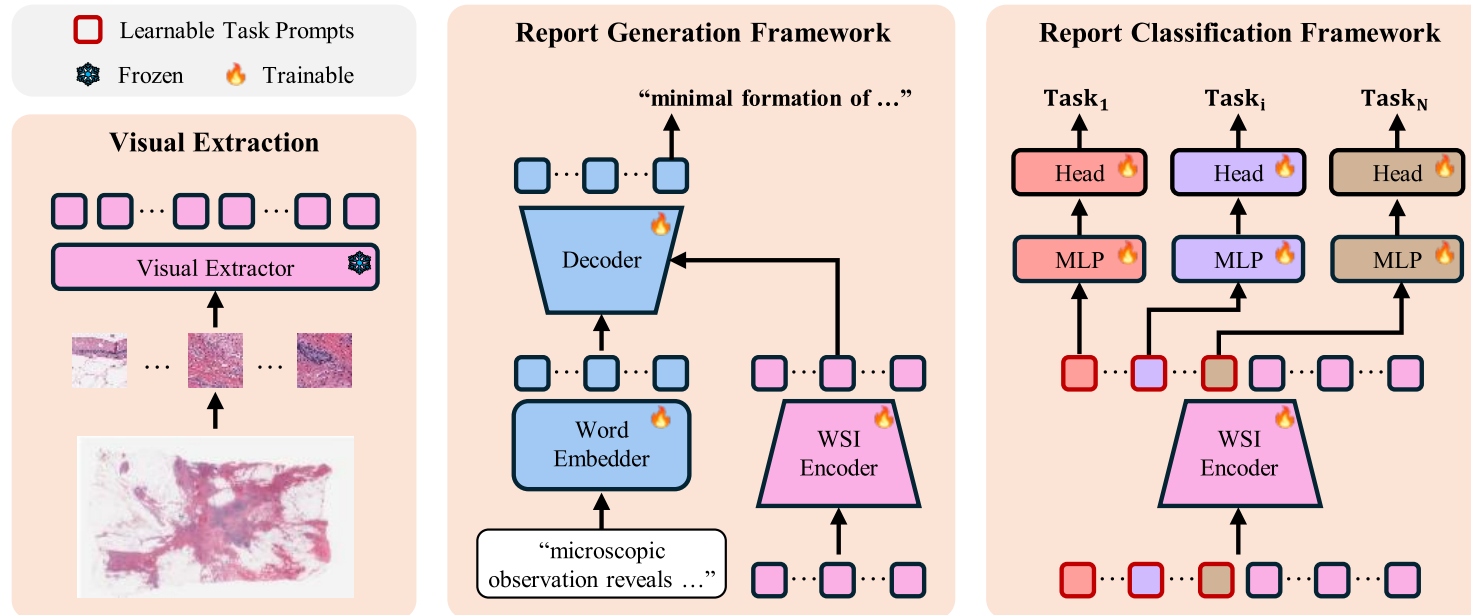


Our Classification-Oriented Adaptations

We adapt existing report generation backbones (MiGen, HistGen, BiGen) to the multi-task setting with **minimal architectural changes**:

- we keep the frozen patch extractor and the transformer WSI encoder;
- we remove the textual branch and the generative decoder;
- we prepend a set of **learnable task prompts**, one per pathological feature, each fed into its own MLP + linear head.

Each task is optimised with a **class-weighted cross-entropy** computed only on the samples for which that label is available.

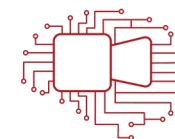


Quantitative Results

Classification-oriented variants outperform their generative counterparts on almost every pathological feature:

- **+15.7** (MiGen), **+15.6** (HistGen), **+15.8** (BiGen) F1 points on average, and $\approx +16$ recall points;
- all generative models reach comparable captioning scores (BLEU-4 0.39, ROUGE-L 0.55, METEOR 0.30): lexical metrics do not separate them, feature-level evaluation does.

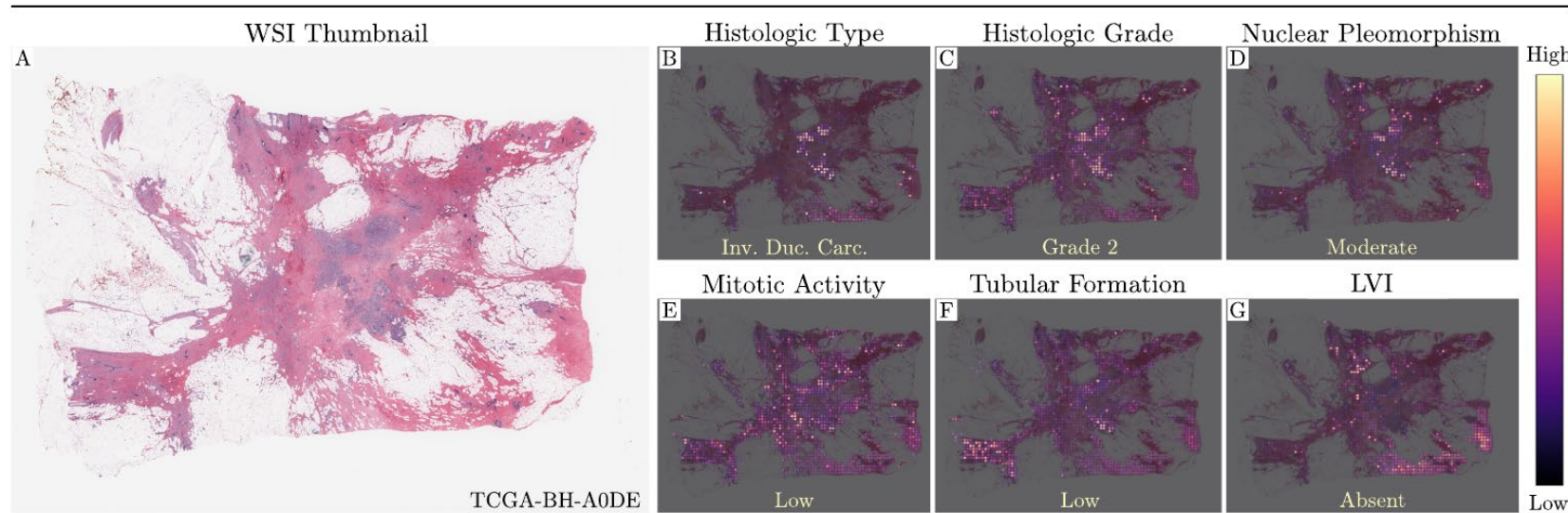
Model	Histological Type		Histological Grade		Nuclear Pleomorphism		Mitotic Activity		Tubular Formation		LVI	
	F1-score	Recall	F1-score	Recall	F1-score	Recall	F1-score	Recall	F1-score	Recall	F1-score	Recall
MiGen	60.4 _{±4.1}	59.7 _{±5.3}	42.0 _{±8.7}	42.1 _{±6.6}	47.4 _{±7.5}	48.9 _{±7.2}	43.2 _{±10.9}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



Conclusions

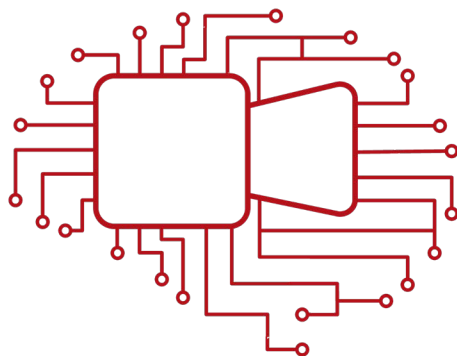
- Pathology reporting can be reformulated as **guideline-aligned multi-task classification**.
- **Feature-level evaluation** exposes clinically relevant errors that captioning metrics fail to capture.
- Task prompts specialise on distinct slide regions → **interpretable** predictions.
- Future work: extension to additional tumour types, toward a multi-cancer benchmark.

Dataset and code: github.com/AImageLab-zip/PathoClassBRCA





Scan Me!



BMVC

2026



A. Saporita



V. Pipoli



F. Bolelli



A. Acquaviva



E. Ficarra