BMVC 2026

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

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

AImage Lab
University of Modena and Reggio Emilia

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

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.

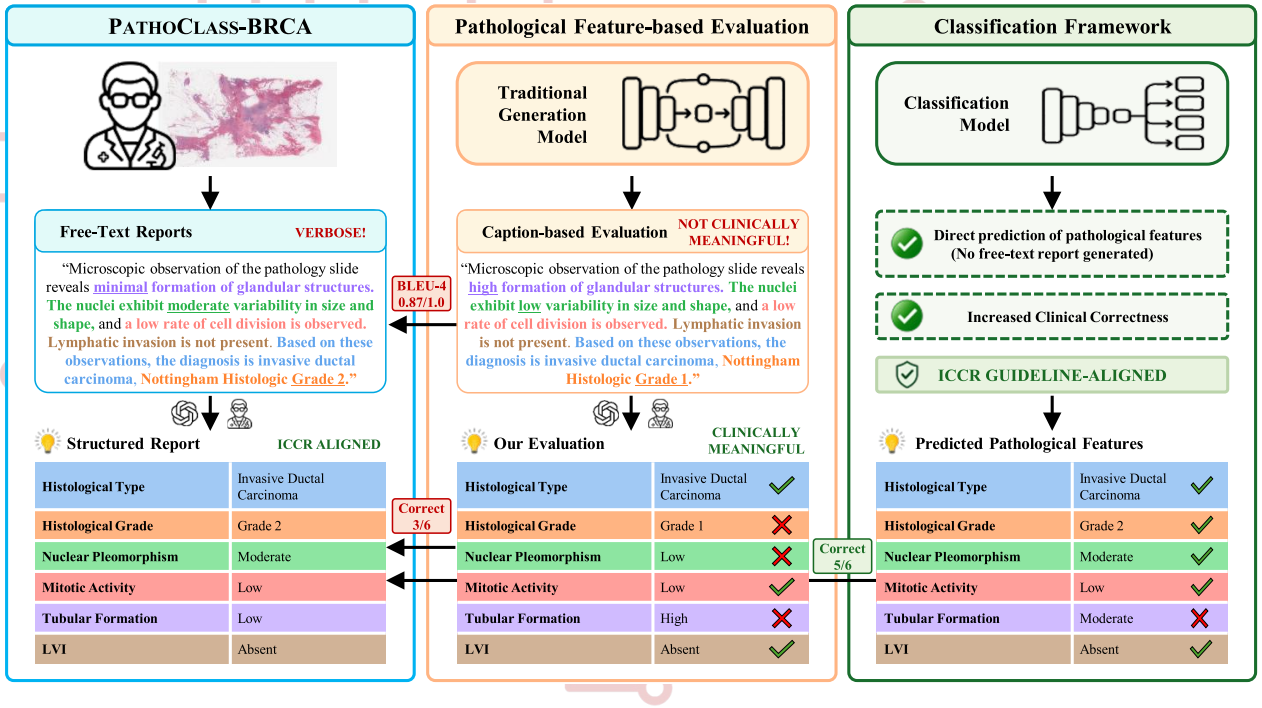
In our figure:

- 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**.



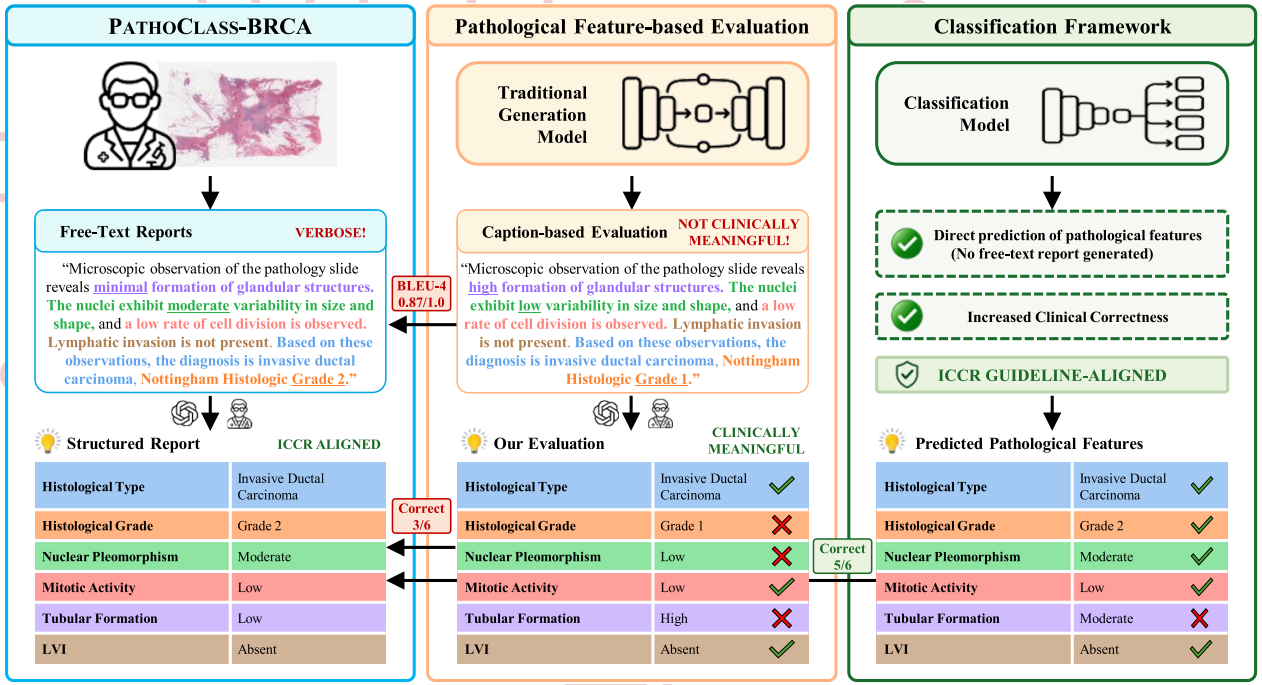
BMVC 2026

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.



BMVC 2026

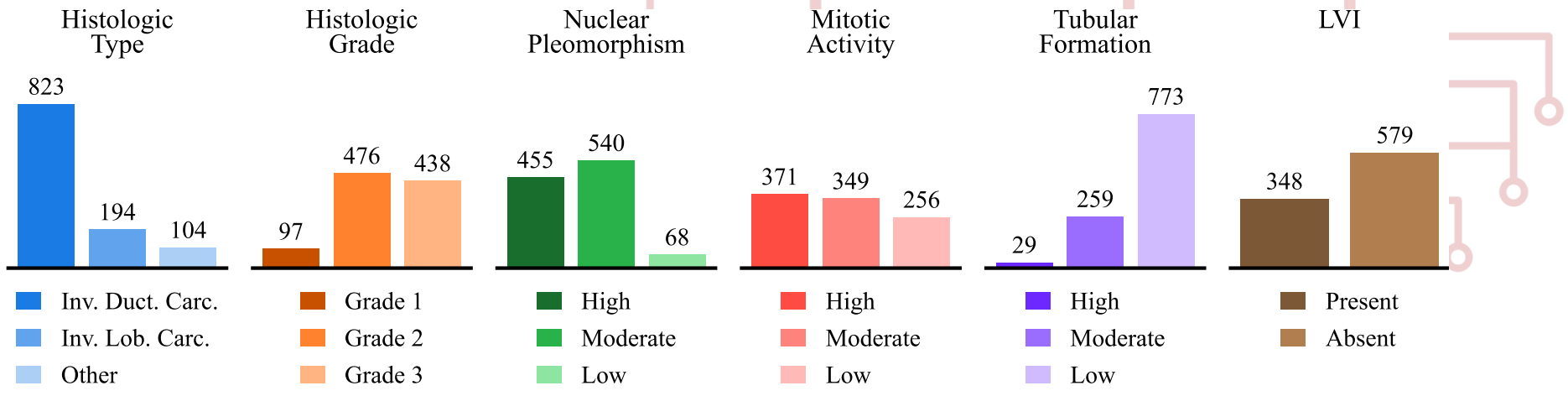
PathoClass-BRCA

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.



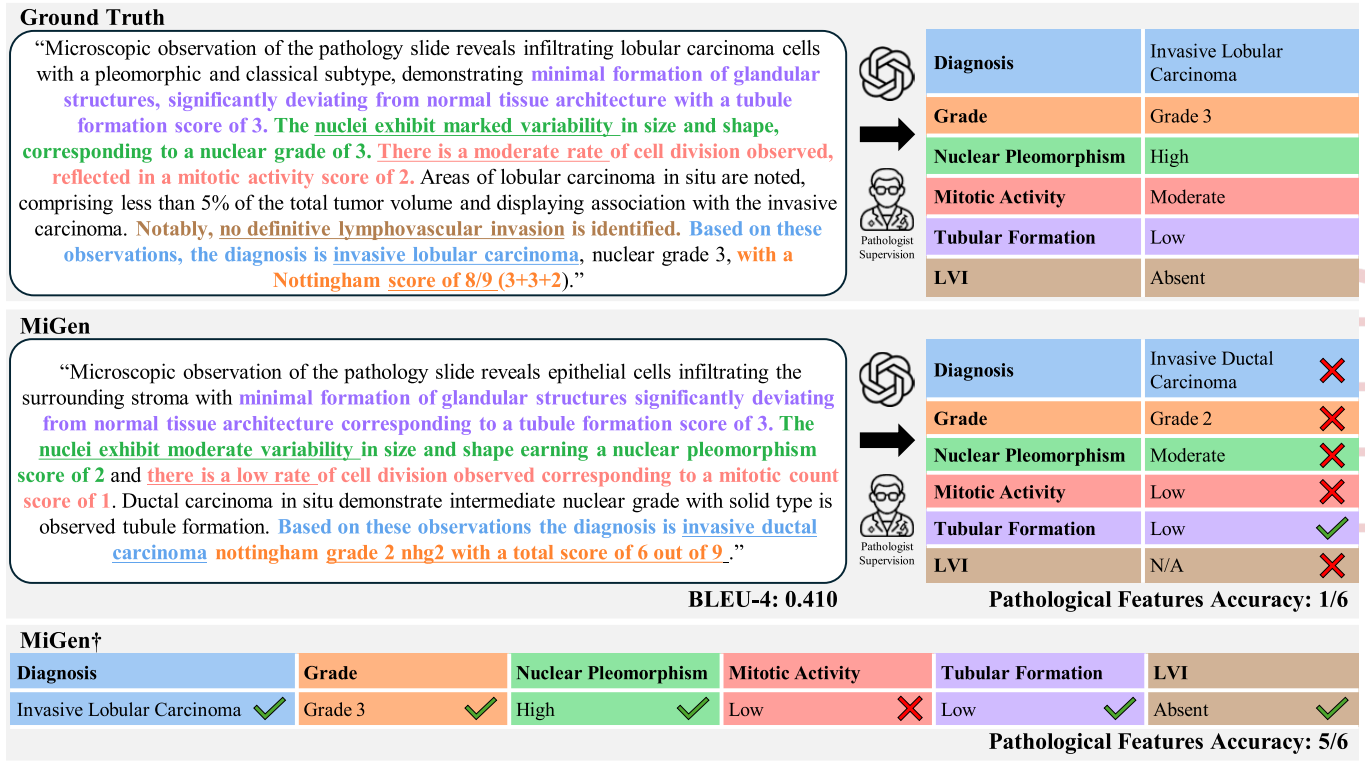
BMVC 2026

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



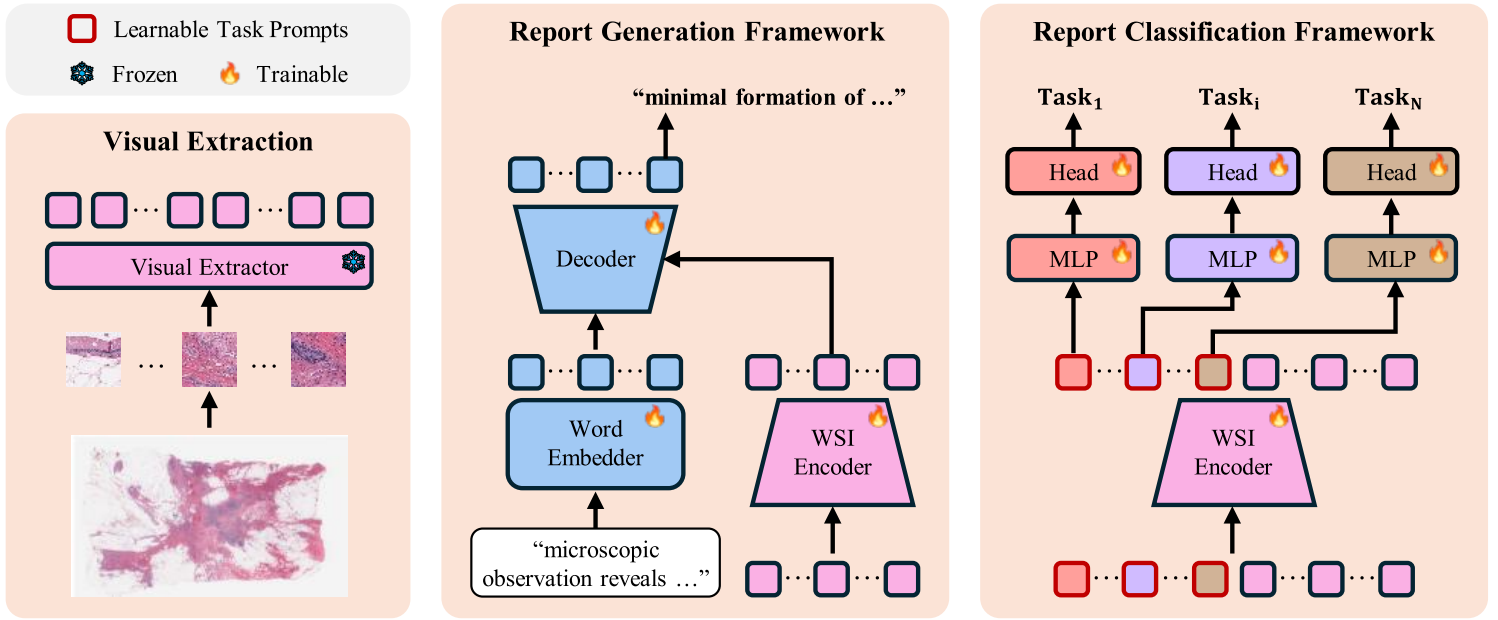
BMVC 2026

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.



BMVC 2026

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 ≈ +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.

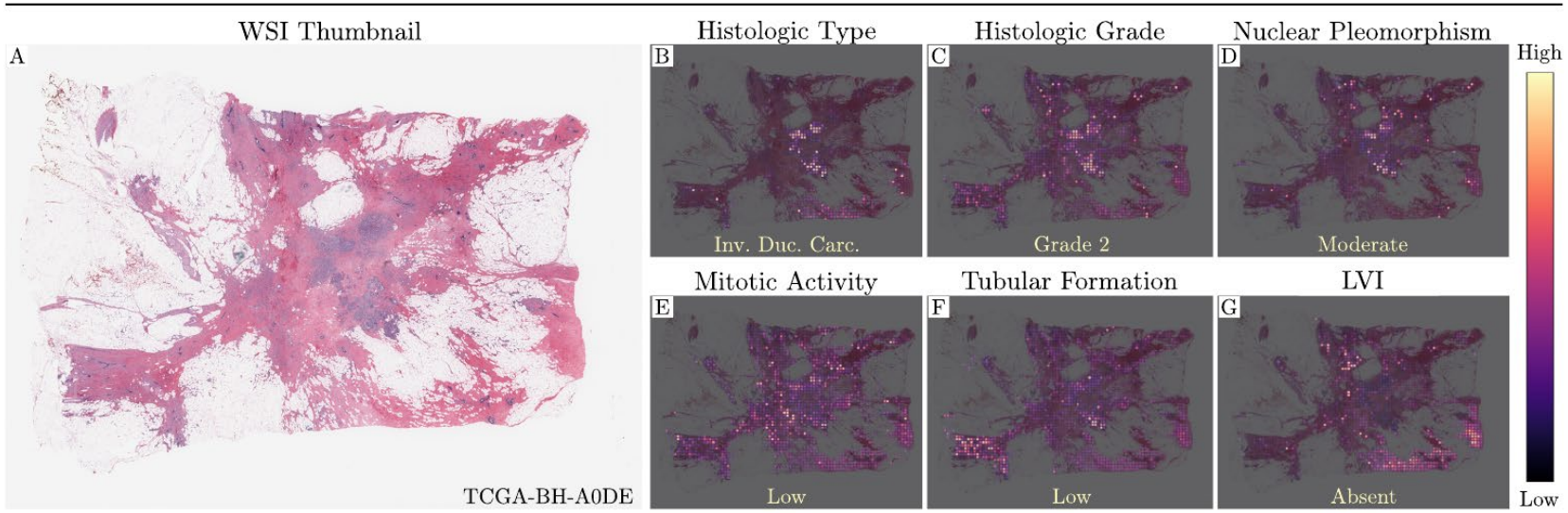
Model	Histological Type		Histological Grade		Nuclear Pleomorphism		Mitotic Activity		Tubular Formation		LVI	
	F1-score	Recall	F1-score	Recall	F1-score	Recall	F1-score	Recall	F1-score	Recall	F1-score	Recall
MiGen	$60.4_{\pm 4.1}$	$59.7_{\pm 5.3}$	$42.0_{\pm 8.7}$	$42.1_{\pm 6.6}$	$47.4_{\pm 7.5}$	$48.9_{\pm 7.2}$	$43.2_{\pm 10.9}$	$42.9_{\pm 11.6}$	$35.7_{\pm 10.3}$	$34.0_{\pm 10.3}$	$38.8_{\pm 9.7}$	$40.7_{\pm 11.2}$
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†	$62.4_{\pm 2.3}$	$64.0_{\pm 1.8}$	$\mathbf{68.9}^{*}_{\pm 6.6}$	$\mathbf{67.9}^{*}_{\pm 6.2}$	$\mathbf{65.4}^{*}_{\pm 5.3}$	$\mathbf{64.5}^{*}_{\pm 5.4}$	$\mathbf{62.9}^{*}_{\pm 6.3}$	$\mathbf{64.3}^{*}_{\pm 5.6}$	$\mathbf{51.7}^{*}_{\pm 3.5}$	$\mathbf{53.5}^{*}_{\pm 4.3}$	$\mathbf{64.7}^{*}_{\pm 12.0}$	$\mathbf{67.3}^{*}_{\pm 10.1}$
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

BMVC 2026

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

AImage Lab
University of Modena and Reggio Emilia

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA