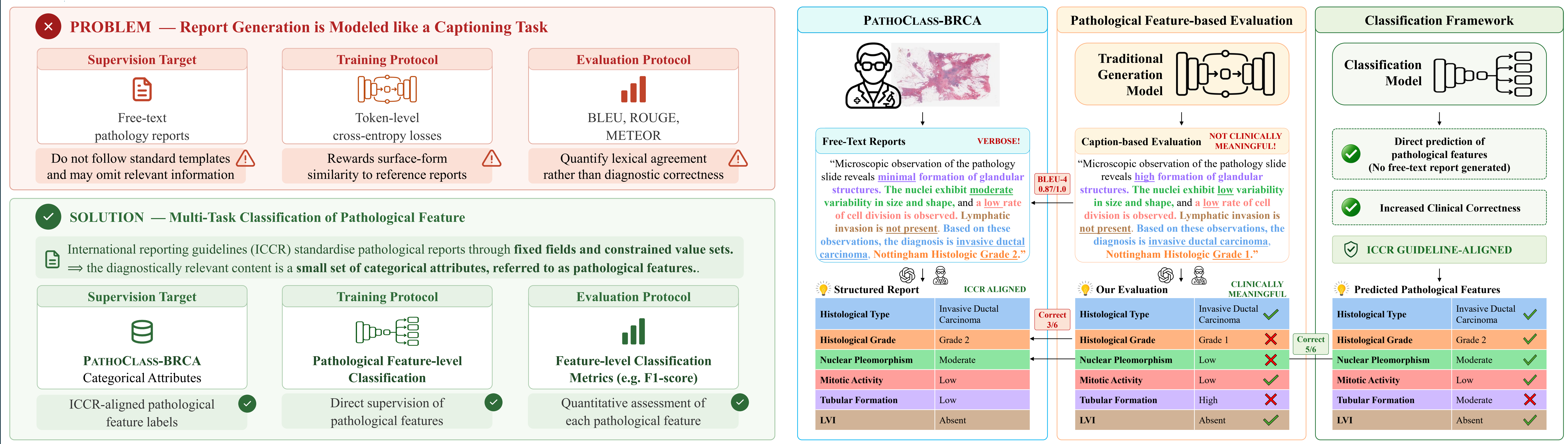
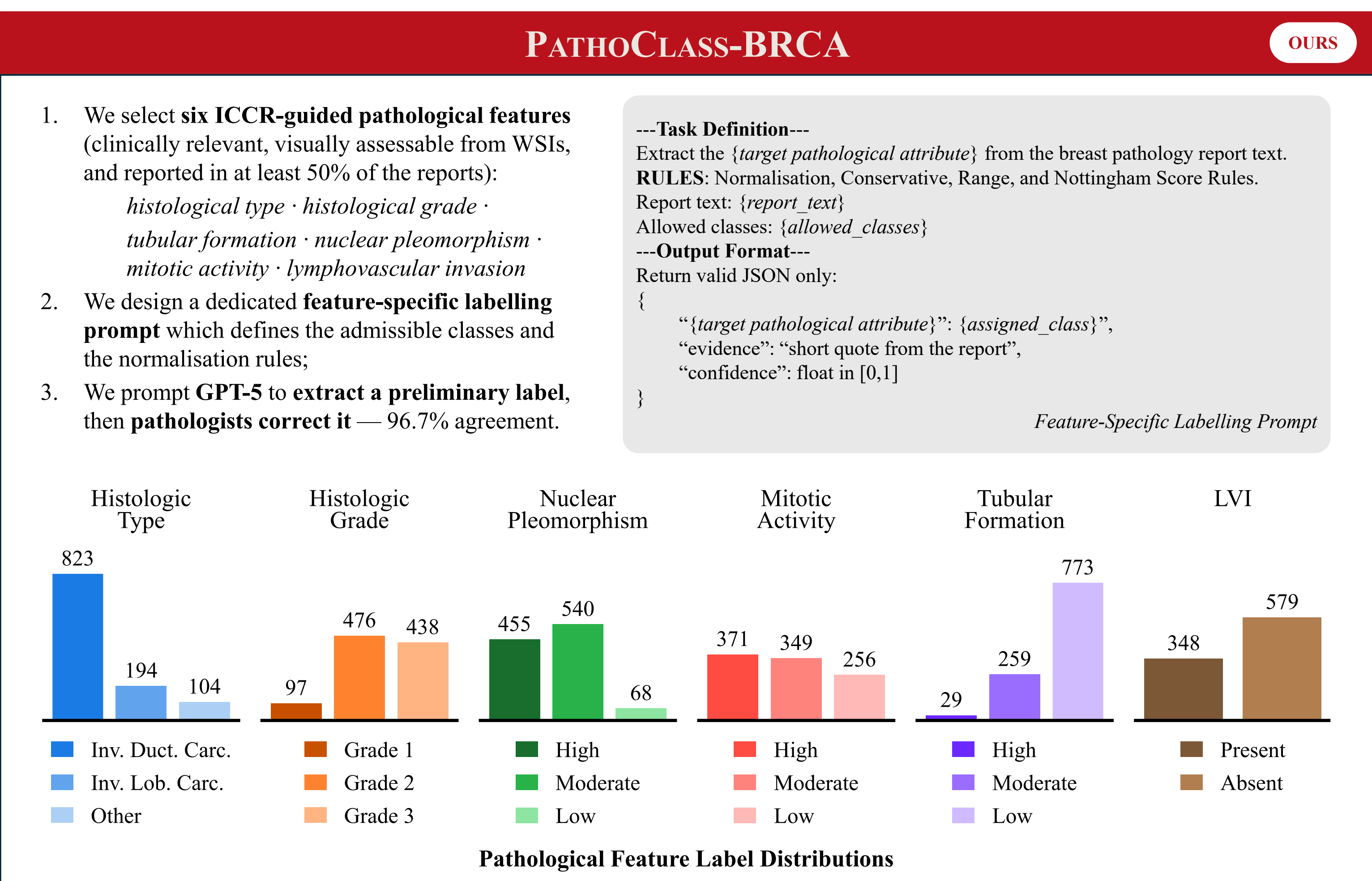
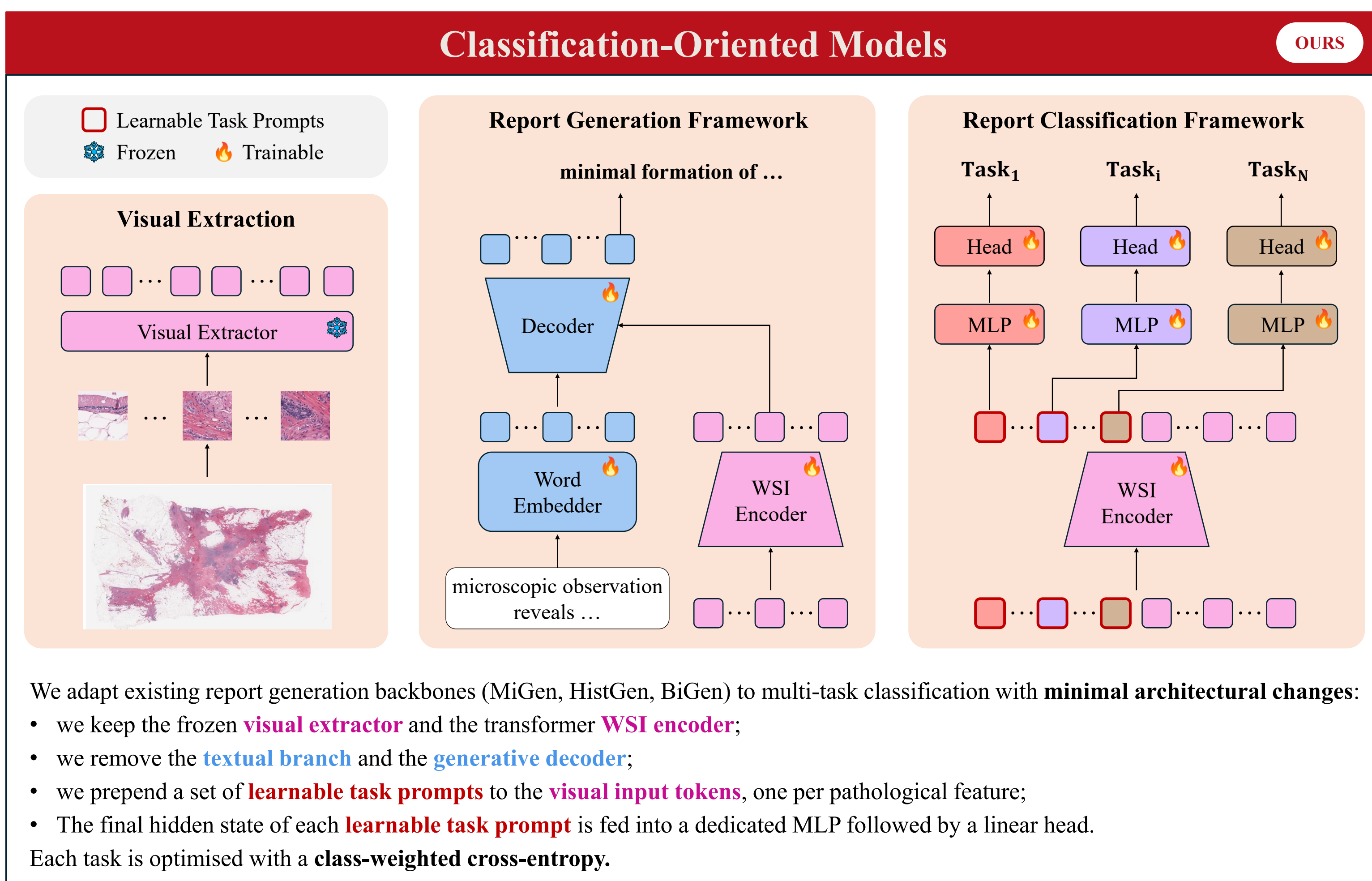
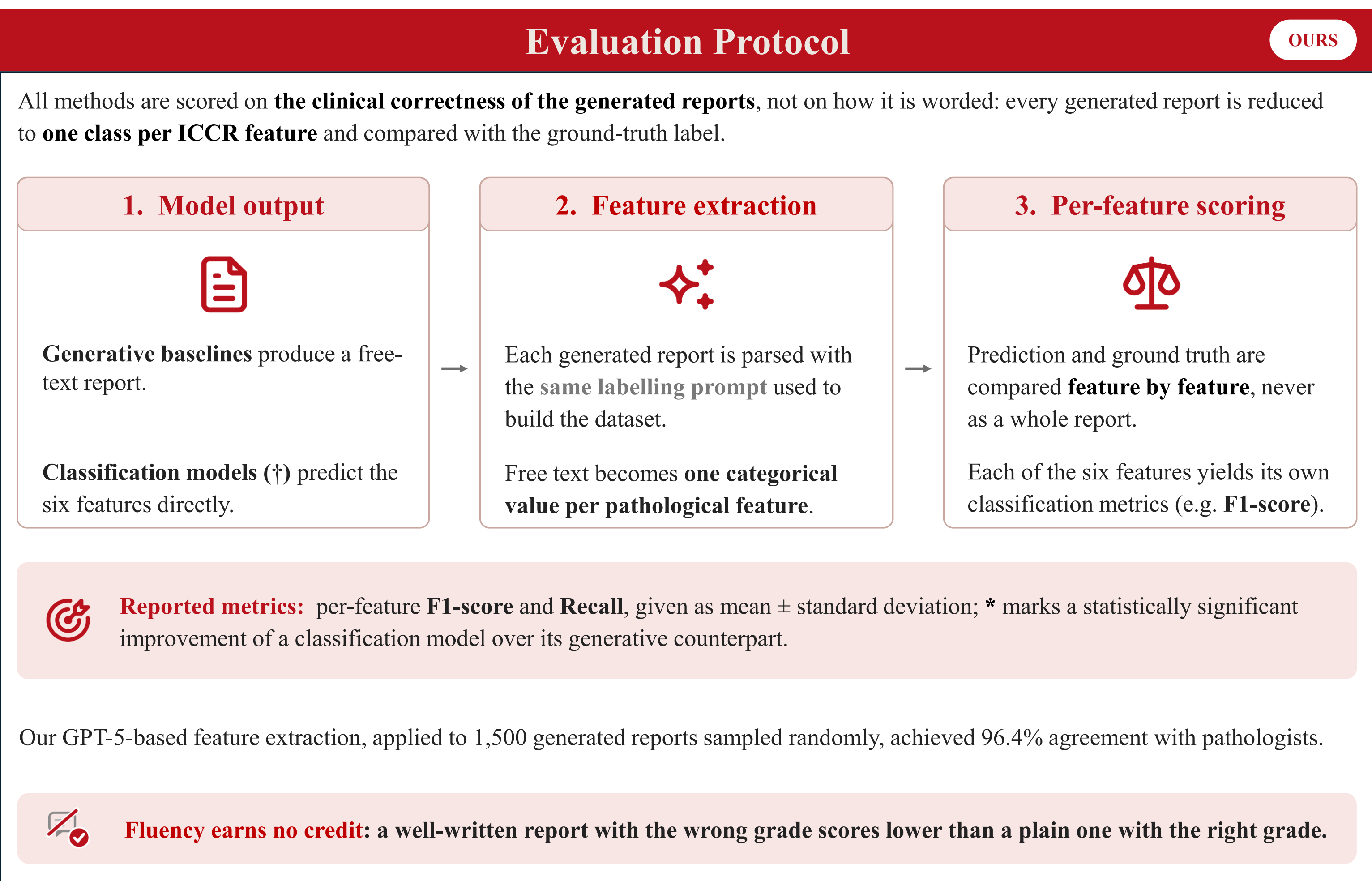


Captioning Falls Short in Pathology: Towards Multi-Task Classification



Main Results

F1-macro (%) and Recall-Macro (%) are reported as mean $\pm$ 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 $p < 0.05$ (one-sided paired t -tests) 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†	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
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†	75.2* $\pm$ 11.5	83.1* $\pm$ 15.1	64.6* $\pm$ 5.3	64.8* $\pm$ 5.6	58.1* $\pm$ 5.0	58.1* $\pm$ 5.1	62.5* $\pm$ 7.4	63.1* $\pm$ 7.1	52.2* $\pm$ 4.2	54.3* $\pm$ 4.6	51.9* $\pm$ 9.7	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	68.9* $\pm$ 6.6	67.9* $\pm$ 6.2	65.4* $\pm$ 5.3	64.5* $\pm$ 5.4	62.9* $\pm$ 6.3	64.3* $\pm$ 5.6	51.7* $\pm$ 3.5	53.5* $\pm$ 4.3	64.7* $\pm$ 12.0	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

- Classification-oriented models consistently outperform their generative counterparts.**
- Classification-oriented models can compete and often outperform large-scale MLLM like WSI-LLaVA** despite its substantially larger parameter count.
- 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.

