This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material

PathPool: Decoupled and Pooled Feature Serving for Pathology Foundation Models

Table A1. M-to-N scaling on 120 WSIs (401,233 patches) under arrival-order FIFO for both encoders (mean over three seeds; utilization averaged over active GPUs). Each encoder uses its own calibrated batch size (16 for Virchow2, 32 for UNI), so absolute throughputs are not comparable across blocks. A single producer supplies about 640 patches/s in both, visible at $(1, 2)$ and $(1, 4)$: patch production does not depend on the encoder, consumption does. With Virchow2 the workers are the limit and throughput scales nearly linearly; with UNI the producers are, so utilization stays low until more are added. No single producer-to-worker ratio suits both encoders, which is why PathPool exposes M and N as separate parameters.

Encoder	(M, N)	Throughput (patches/s)	GPU util. (%)	System CPU (%)	Makespan (s)
Virchow2	(1,1)	285.0	95.1	20.9	1406
	(1,2)	507.0	84.0	32.8	792
	(1,4)	633.3	53.0	38.3	633
	(2,4)	983.7	82.6	59.8	408
	(4,4)	1067.3	89.1	67.9	376
UNI	(1,1)	521.7	66.0	28.9	769
	(1,2)	629.7	39.7	34.5	637
	(1,4)	645.0	20.8	34.0	622
	(2,4)	1163.0	36.7	59.9	346
	(4,4)	1385.7	44.4	75.6	290

Table A2. Per-cohort workload statistics (patches per slide) and small/medium/large tier counts. Tiers are assigned by file size and tissue area within each cohort.

Cohort	Slides	Min	Median	Max	Mean	CV	Small	Med.	Large
BRCA	20	305	2903	9538	3144	0.626	4	10	6
CESC	20	232	2418	7151	2783	0.772	4	10	6
COAD	20	367	3574	7623	3310	0.517	4	10	6
LUAD	20	371	3319	8342	3499	0.686	4	10	6
PRAD	20	167	3702	10621	3736	0.649	4	10	6
STAD	20	260	3794	6235	3587	0.412	4	10	6
All	120	167	3515	10621	3343	0.621	24	60	36

Table A3. Embedding correctness on all 120 slides (20 per cohort). PathPool at fp16 is bit-identical to the independent fp16 reference (cosine ≈ 1, $\max|\Delta| = 0$) and every output is index- and coordinate-complete. Reported cosine minima are 1 to within 1.1×10^{-15}, the rounding of the similarity computation. The per-cohort rows are Virchow2; the final row repeats the whole check with UNI, whose own precision gap was not measured. The last two columns report the encoder's own fp16-vs-fp32 precision gap, which is unrelated to PathPool and is measured on a stratified subset of three slides per cohort.

		PathPool vs. ref (fp16)		ref fp16 vs. fp32	
Cohort	Complete	cosine (min)	$\max\|\Delta\|$	cosine (min)	$\max\|\Delta\|$
BRCA	20/20	1.0000000	0.0	0.99924	1.012
CESC	20/20	1.0000000	0.0	0.99985	0.297
COAD	20/20	1.0000000	0.0	0.99991	0.261
LUAD	20/20	1.0000000	0.0	0.99992	0.251
PRAD	20/20	1.0000000	0.0	0.99964	0.517
STAD	20/20	1.0000000	0.0	0.99979	0.504
All	120/120	1.0000000	0.0	0.99924	1.012
All (UNI)	120/120	1.0000000	0.0	—	—

Table A4. Per-workflow flow time (mean and p95, in seconds) under static partitioning and shared pooling, with Virchow2. For the balanced and skewed mixes, wf0–wf3 are BRCA, LUAD, COAD, and PRAD; under skew, wf0 (BRCA) is the heavy workflow. Lower is better.

		Static		Pooled	
Scenario	WF	mean	p95	mean	p95
Balanced	wf0	123.7	216.1	131.2	251.8
	wf1	138.9	242.4	137.7	239.1
	wf2	124.9	235.2	137.9	234.1
	wf3	146.3	265.1	142.6	242.0
Skewed	wf0	215.2	397.2	93.8	179.2
	wf1	95.8	169.1	105.8	171.1
	wf2	36.1	57.2	107.4	162.8
	wf3	54.1	76.7	114.6	172.4
Scale-out	wf0	266.1	430.3	206.3	344.4
	wf1	219.5	381.4	168.6	331.3
	wf2	128.8	236.3	188.5	341.8
	wf3	140.0	251.7	203.1	357.0
	wf4	284.8	488.5	204.2	335.0
	wf5	250.9	446.4	210.2	347.5