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

A Directional cross-modal retrieval

Table S1 reports the full directional retrieval metrics underlying the bidirectional R@10 results in the main manuscript. Overall, performance was largely symmetric with respect to the query modality.

Table S1. Directional cross-modal retrieval. Values are mean ± standard deviation across 31 held-out `Lung` patients or ten `Prostate` spatial folds. MedR is the mean of the per-unit median ranks; lower is better.

Dataset	Model	Direction	R@1	R@5	R@10	MedR
`Lung`	Bimodal	Image→RNA	.135±.077	.332±.159	.444±.187	31.0±55.7
`Lung`	Bimodal	RNA→Image	.134±.074	.334±.155	.442±.184	31.0±58.1
`Lung`	QUERY-ST	Image→RNA	.127±.072	.318±.149	.427±.173	31.4±55.5
`Lung`	QUERY-ST	RNA→Image	.132±.075	.330±.154	.439±.179	30.6±54.6
`Lung`	QUERY-ST	Image→Text	.066±.036	.194±.091	.292±.125	48.5±54.4
`Lung`	QUERY-ST	Text→Image	.073±.040	.213±.097	.312±.129	44.4±55.8
`Lung`	QUERY-ST	RNA→Text	.724±.112	.916±.071	.954±.050	1.1±0.4
`Lung`	QUERY-ST	Text→RNA	.763±.098	.935±.059	.967±.041	1.0±0.2
`Prostate`	Bimodal	Image→RNA	.158±.019	.394±.038	.523±.043	9.6±2.2
`Prostate`	Bimodal	RNA→Image	.149±.018	.375±.037	.504±.042	10.6±2.4
`Prostate`	QUERY-ST	Image→RNA	.154±.021	.385±.043	.513±.050	10.3±2.6
`Prostate`	QUERY-ST	RNA→Image	.161±.022	.387±.043	.514±.049	10.2±2.6
`Prostate`	QUERY-ST	Image→Text	.047±.007	.156±.019	.243±.028	42.5±8.8
`Prostate`	QUERY-ST	Text→Image	.048±.007	.159±.019	.246±.029	41.2±8.5
`Prostate`	QUERY-ST	RNA→Text	.425±.049	.694±.054	.788±.050	2.2±0.4
`Prostate`	QUERY-ST	Text→RNA	.422±.042	.696±.051	.792±.049	2.2±0.4

B HESCAPE Comparison

We used HESCAPE to contextualise the Image–RNA backbone before its extension to trimodal alignment.

Table S2. Lung LOPO Image↔RNA comparison with HESCAPE. Both models use gene-only RNA inputs and a $d = 128$ retrieval space under identical LOPO candidate sets. Parameter counts exclude the frozen UNI2-h encoder. Values are mean ± standard deviation across 31 held-out patients; directional metrics are averaged within each patient. Lower MedR is better.

Framework	Width	Parameters	R@1	R@5	R@10	MedR
HESCAPE	128	0.538M	0.038 ± 0.024	0.134 ± 0.079	0.209 ± 0.117	80.9 ± 77.8
QUERY-ST	128	0.341M	0.063 ± 0.038	0.194 ± 0.103	0.288 ± 0.139	53.9 ± 62.7

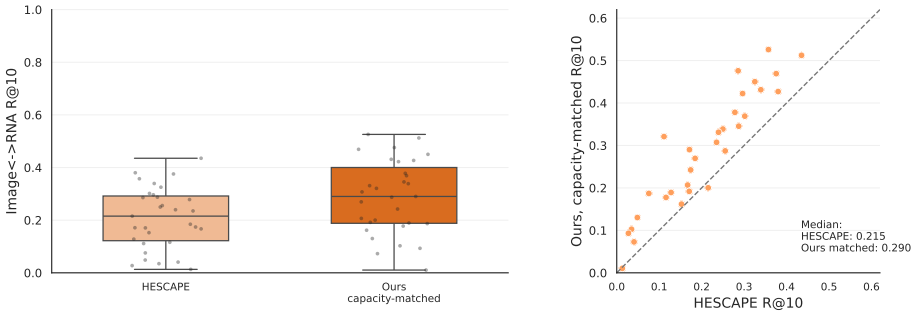


Figure S1. Patient-level Image–RNA backbone comparison with HESCAPE on the `Lung` LOPO benchmark. (**A**) Distribution of bidirectional Image↔RNA R@10 across held-out patients for HESCAPE and the reduced-width, gene-only QUERY-ST model. (**B**) Paired patient-level R@10 under identical LOPO candidate sets; the dashed line denotes equal performance.

C Validation of language-guided retrieval

For the `Prostate` Text→Image example in Fig. 2C, we inspected the paired (ground-truth) ST measurements of the five top-ranked patches after retrieval. All five patches fell within the 95th–100th cellularity percentiles, with an effector CD8+ T-cell state among the four most abundant annotated cell states in every patch.

Table S3. ST-derived validation of the `Prostate` retrieval example. Paired ST measurements for the five top-ranked patches retrieved for "High-cellularity immune region enriched for effector CD8+ T cells." Cellularity denotes the percentile rank of patch cell count. Effector CD8+ rank is the abundance rank of the highest-ranked effector CD8+ T-cell state among annotated cell states; lower values indicate greater abundance. Effector CD8+ cells denotes the total number of cells assigned to effector CD8+ states.

Retrieval rank	Cellularity	CD8+ rank	CD8+ cells
1	0.99	2	16
2	0.95	4	3
3	0.99	1	9
4	0.98	3	5
5	1.00	2	9

D Contribution of text report levels

We examined how the three report levels contribute to cross-modal alignment using the fixed `Lung` spatial ablation split. Seven model variants were trained using

Level 1 (*Composition*), Level 2 (*Microenvironment interactions*), Level 3 (*Density*), their pairwise combinations, or all three levels. Architecture, optimisation, RNA features, and data partitions were held fixed; only the text representation was varied.

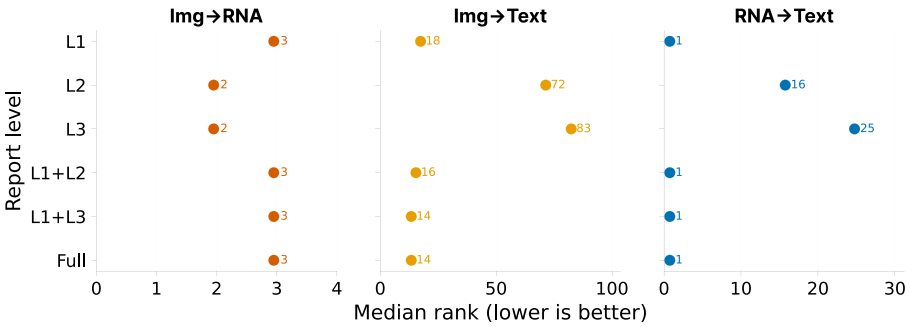


Figure S2. Text report-level ablation. Median retrieval rank for models trained using different ST-derived report levels on the fixed `Lung` ablation split; lower is better. Level 1 composition provides the strongest standalone anchor, while Level 2 interaction and Level 3 density provide complementary contextual information when combined with composition.

Among the individual report levels, Composition provided the strongest standalone alignment to the text modality (Fig. S2). This is expected because Level 1 is derived from cell-type abundances that are represented directly in the RNA branch and may also have visible histomorphological correlates. Interaction and density were weaker in isolation, consistent with these properties providing more specific contextual information that is not always uniquely identifiable from local gene-expression or H&E morphology.

Combining interaction and density information with composition preserved strong RNA→Text alignment while improving Image→Text retrieval relative to the contextual levels alone. Thus, Level 1 provides the principal alignment signal, whereas Levels 2 and 3 extend the semantic representation with relational and density information used by more specific queries.

E Computational scalability

We evaluated the computational scaling of the trainable alignment stage using cached UNI2-h, gene-expression, and text embeddings from the primary `Lung` dataset. Experiments used $d = 128$, $w = 2048$, $\lambda_1 = 0.4$, and $\lambda_2 = 0.1$ on a single NVIDIA A100 40 GB GPU, with each configuration repeated five times using different random seeds.

Runtime increased from approximately 0.42 to 1.48 s per epoch across 1,000–28,362 patches, showing near-linear scaling with dataset size (Fig. S3A). Across

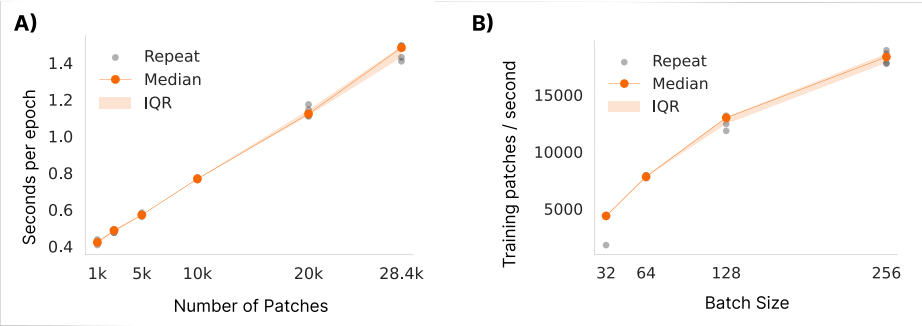


Figure S3. Computational scalability of QUERY-ST. (**A**) Time per epoch as the number of aligned `Lung` patches increases at batch size 256. (**B**) Effective training throughput on all 28,362 patches across batch sizes 32–256. Points show individual repeats; markers and shading denote the median and interquartile range.

batch sizes 32–256, throughput increased from approximately 4,400 to 18,000 patches/s, indicating more efficient GPU utilisation at larger batch sizes (Fig. S3B).