A. Cross-Modal Attention over local representations

The pairwise local interaction for WSIs and RNA-Seq is defined as:

$$X^{L}_{\text{WSI}\leftrightarrow\text{RNA}} = \frac{1}{X^{L^*}_{WSI}} \sum_{a} (z^{WSI}_{a^*} + \sum_{b} (\gamma^{a,b}_{\text{WSI}\leftrightarrow\text{RNA}} \cdot z^{RNA}_{b^*}) \tag{1}$$

where $X^{L^*}_{WSI}$ and $X^{L^*}_{RNA}$ are the sets of K key local features of RNA and WSI (i.e., $\{z^{WSI}_{j^*} | j = 1, \ldots, K\}$ and $\{z^{RNA}_{j^*} | j = 1, \ldots, K\}$).

$\gamma^{a,b}_{\text{WSI}\leftrightarrow\text{RNA}}$ is the normalized correlation coefficient computed as $\gamma^{a,b}_{\text{WSI}\leftrightarrow\text{RNA}} = \frac{\exp((z^{WSI}_{a^*})^\top z^{RNA}_{b^*})}{\sum_b \exp((z^{WSI}_{a^*})^\top z^{RNA}_{b^*})}$.

This process is similarly applied for interactions between other modalities at the local level. After computing the pairwise interactions, the final aggregated representations at the local level are obtained using the same methodology as for the global level using Eqs. 5-8 in the main paper.

Dataset	BRCA	NSCLC	RCC
Two Modalities			
Total Patients	IDC: 786 ILC: 197	LUAD: 472 LUSC: 476	KIRC: 510 KIRP: 272 KICH: 66
Total Slides	IDC: 838 ILC: 210	LUAD: 534 LUSC: 510	KIRC: 516 KIRP: 296 KICH: 66
Three Modalities			
Total Patients	IDC: 737 ILC: 190	LUAD: 438 LUSC: 440	KIRC: 498 KIRP: 261 KICH: 65
Total Slides	IDC: 788 ILC: 203	LUAD: 499 LUSC: 474	KIRC: 501 KIRP: 283 KICH: 65

Table 1. Data Statistics for BRCA, NSCLC, and RCC Datasets

B. Feature Selection via Mixture of Experts (MoE) and Top-K Activation

For an input WSI, the local representation is defined as $X^{L}_{WSI} = \{x_j \mid j = 1, \ldots, C\}$. We define a set of expert networks $\{EXP_k(\cdot)\}_{k=1}^{K}$, where each expert EXP_k independently processes the local representations.Each expert consists of two connected layers of size 1024 with ReLU activation. A gating network $Gate(\cdot)$ determines the contribution of each expert by assigning selection weights:

$$\alpha_k = Gate_k(x_j), \quad \sum_{k=1}^{K} \alpha_k = 1, \quad \alpha_k \geq 0. \tag{2}$$

where α_k represents the gating score for expert k, computed using a softmax function over expert logits. To enforce sparsity and encourage selective activation, we apply a Top-K selection mechanism, retaining only the highest-scoring experts per local representation:

$$\mathcal{S}_j = \text{Top-K}(\alpha_1, \alpha_2, \ldots, \alpha_K), \quad |\mathcal{S}_j| = K' \tag{3}$$

where $\mathcal{S}_j$ denotes the selected subset of experts, and K' is the number of active experts, which is set to 5. The final expert representation is computed as a weighted sum over the selected experts:

$$z_j = \sum_{k \in \mathcal{S}_j} \alpha_k EXP_k(x_j) \tag{4}$$

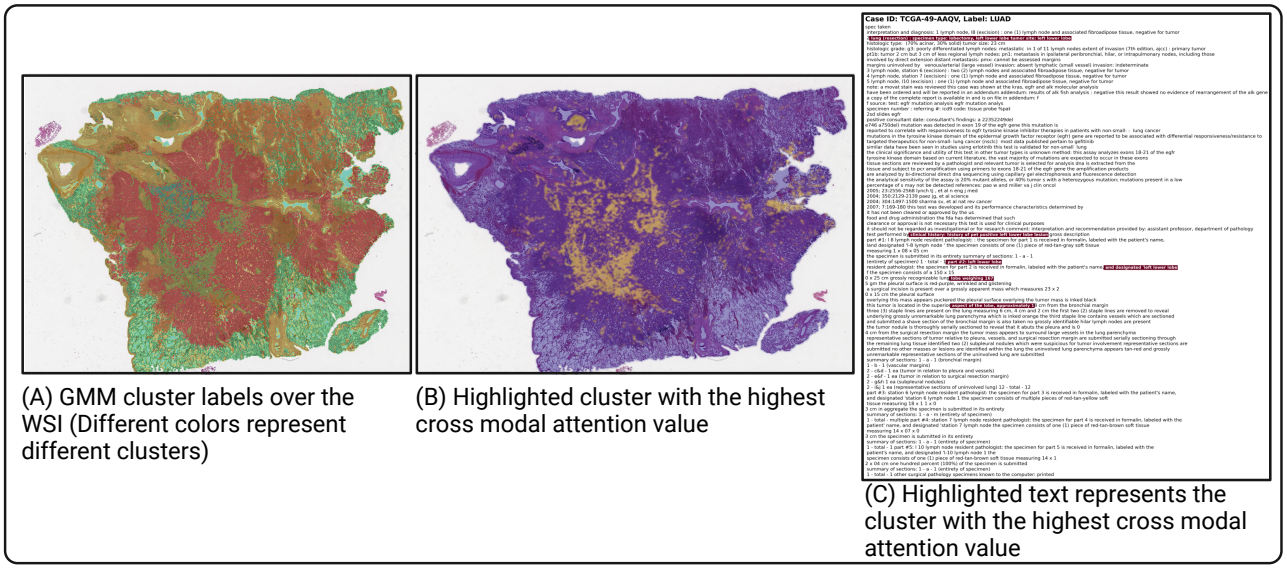
To further refine local representations, we apply an activation gating mechanism $Act(\cdot)$ that assigns selection scores ($score_j = Act(z_j)$). We retain only the k most discriminative local features:

$$\mathcal{A} = \text{Top-k}(score_1, score_2, \ldots, score_C) \tag{5}$$

where $\mathcal{A}$ denotes the k selected active local representations out of C (i.e. number of local representations per WSI). We set k to 3. The final feature set is obtained via:

$$X^{L^*}_{WSI} = Z \odot \mathbf{1}_{\mathcal{A}} = \{z^{WSI}_{1^*}, z^{WSI}_{2^*}, \ldots, z^{WSI}_{k^*}\} \tag{6}$$

where $\odot$ represents element-wise multiplication of the set of MoE outputs for local representations, denoted as $Z = \{z_j \mid j = 1, \ldots, C\}$, with an activation mask $\mathbf{1}_{\mathcal{A}}$. This Mixture of Experts with Top-K activation adaptively selects the most informative features while suppressing redundancy, enhancing model efficiency and robustness in multimodal learning.



(A) GMM cluster labels over the WSI (Different colors represent different clusters)

(B) Highlighted cluster with the highest cross modal attention value

(C) Highlighted text represents the cluster with the highest cross modal attention value

(A) GMM cluster labels over the WSI (Different colors represent different clusters)

(B) Highlighted cluster with the highest cross modal attention value

(C) Highlighted text represents the cluster with the highest cross modal attention value

Figure 1. Illustration showing how cross-modal attention can be used for interpretation. Highlighted text and images show local clusters with highest cross-modal attention values.