6 Supplementary Material

6.1 Experimental Details

For selecting appropriate values for these hyperparameters, we can employ empirical metrics as detailed in the experiment section. By taking a subset of 500 image-text pairs each from the Conceptual Captions and MS-CXR datasets—ensuring no overlap with the test sets used in subsequent experiments—a grid search was conducted to determine the most effective combination of hyperparameter values (Tab. 2).

The hyperparameter σ dictates the magnitude of noise infused into the intermediate representations of the model. A diminutive σ results in added noise values tending towards zero, thereby exerting a negligible effect on these representations. The hyperparameter β serves to balance the significance of the compression component within the information bottleneck architecture. As β increases, the flow of information through the bottleneck is more restricted. The positioning of the bottleneck layer, denoted by ℓ, also influences the resulting attributions. Placing the bottleneck at an earlier layer could hinder the model's ability to develop informative features, while positioning it too late in the network may diminish the bottleneck's intended influence [23].

Θ_A encompasses all trainable weights and biases inherent to the cross-attention mechanism, including the Query Weights and Biases (W_q, b_q), the Key Weights and Biases (W_k, b_k), the Value Weights and Biases (W_v, b_v), the Output Transformation Weights and Biases (W_o, b_o), and the Scaling Factor (d_k). After the attention scores are computed and applied to the value vectors, this final set of parameters is used to transform the aggregated output into the final attention-modulated embedding $\tilde{E}_{m'}$.

The cross-attention process unfolds as follows, underpinned by Θ_A. First, the Query-Key Matching, where for each element in the source modality, query vectors (Q) are matched against key vectors (K) from the target modality. This matching is quantified by computing the dot product between every query and key pair, scaled by $\frac{1}{\sqrt{d_k}}$ to control the variance of the dot products. Then, the Attention Weight Calculation, where Softmax is applied across these scaled dot products for each query, yielding attention weights that signify the relevance of each key to the query. Value Aggregation follows, in which attention weights are then used to weigh the corresponding value vectors (V), and the results are summed up to produce the attention output for each query. Finally, the attention output is transformed by W_o and b_o to produce $\tilde{E}_{m'}$, the attention-modulated embedding of the target modality. Θ_A thus encapsulates the essence of cross-modal attention within the CA-M2IB framework, facilitating an adaptive interplay between modalities to enhance mutual information and interpretability.

Table 2: Hyperparameter Selection

Hyper-parameter	MSCXR Image	MSCXR Text	CC Image	CC Text	Range
β	1	1	1	1	$\{0.01, 0.1, 1\}$
σ^2	1	1	0.1	0.1	$\{0.01, 0.1, 1\}$
ℓ	9	9	9	7	$\{7, 8, 9\}$
γ	0.1	0.1	1	1	$\{0.01, 0.1, 1\}$
Q/K/V dim.	64	64	128	128	$\{64, 88, 128\}$
Nb. of heads	8	8	8	8	$\{4, 6, 8\}$
Dropout rate	0.3	0.3	0.3	0.3	$\{0.1, 0.2, 0.3\}$
LayerNorm	True	True	True	True	{True, False}