A. Discussion

A.1. Differences with attention mechanism

Different motivation. The proposed method is a general framework to handle multimodal dynamics for more trustworthy multimodal fusion. Both gating strategies and attention mechanism can be used under the proposed framework. *Efficient implementation*. We employ a more efficient way to obtain feature informativeness while the traditional attention mechanism is more complex, e.g., Transformer, leading to larger model size.

A.2. Other ways to obtain multimodal dynamics

It is interesting to propose a more principled framework that can handle multimodal dynamics. Furthermore, other uncertainty-based methods are also choices to model multimodal dynamics. For example, each feature of the input data can be modeled by the uncertainty estimation algorithm to model the feature informativeness, and the feature informativeness can be dynamically aggregated within the modality to obtain the modal informativeness.

A.3. Differences of MCP and TCP

Compared with MCP, we can obtain a more reliable estimation of modality confidence through an approximation of TCP during the test stage to guide multimodal fusion. By taking the largest softmax probability, MCP leads to high confidence values even for erroneous predictions while TCP could be more likely close to a low value, reflecting the fact that the model made an error.

B. Implementation Details

The implementation details of our method are detailed here: *Feature encoder:* a one-layer perceptron with a sigmoid activation function. *Depth of the f^m classifiers:* 2-3 linear layers for all datasets. *Adaptation for each dataset:* The difference between different datasets lies in the choice of hyperparameters, e.g., learning rate (selected from 1e-4, 5e-5, and 1e-5) and network layers (selected from 2 and 3).

Implementation details of compared methods. We reimplemented GMU, CF, and TMC methods. For all the reimplemented methods, we tune their hyperparameters for the best performance on all datasets. For other comparison methods, we directly use the experimental results in MOGONET [62].

C. Comparison with transformer based method

We conducted comparison experiments with transformer-based early and late fusion methods. The experimental results are shown in Tab. 4. It is observed that the proposed method achieves the best performance. A possible reason is that Transformer is more difficult to train on small-scale datasets due to the large model size and overfitting. We show the transformer-based model size in Tab. 5, which is much larger than other methods leading to higher computational cost and training difficulty.

Table 4. Comparison with Transformer-based methods in terms of accuracy.

Method	BRCA	KIPAN	LGG	ROSMAP
Transformer Late	79.9±1.4	99.3±0.4	80.2±1.5	77.4±2.7
Transformer Early	81.4±1.0	99.2±0.4	81.6±1.2	78.8±2.6
Ours	**87.7±0.3**	**99.9±0.2**	**83.3±1.0**	**84.2±1.3**

D. Model size and architecture.

For fair comparison, we used the same architecture (2-4 linear layers) for all re-implemented methods. The numbers of model parameters are shown in Tab. 5, where all models share similar model sizes and it is noteworthy that on both LGG and ROSMAP data, we use fewer parameters to obtain better performance.

Table 5. Number of parameters for different methods.

Method	BRCA	KIPAN	LGG	ROSMAP
MOGONET	1.74M	2.76M	2.81M	0.32M
CF	5.52M	5.95M	6.06M	0.72M
GMU	4.01M	5.96M	6.06M	0.73M
Transformer Late	81.22M	295.32M	299.01M	4.35M
Transformer Early	225.66M	711.48M	744.83M	12.99M
Ours	4.79M	14.16M	0.97M	0.31M