# MoME: Mixture of Multimodal Experts for Cancer Survival Prediction

Conghao Xiong[1(✉)[0000−0003−0134−9397]], Hao Chen[2,3[0000−0002−8400−3780]], Hao Zheng[4(✉)], Dong Wei[4], Yefeng Zheng[4[0000−0003−2195−2847]], Joseph J. Y. Sung[5[0000−0003−3125−5199]], and Irwin King[1[0000−0001−8106−6447]]

[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China
[2] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
[3] Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
[4] Jarvis Research Center, Tencent YouTu Lab, Shenzhen, China
[5] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore
chxiong21@cse.cuhk.edu.hk, howzheng@tencent.com

**Abstract.** Survival prediction requires integrating Whole Slide Images (WSIs) and genomics, a task complicated by significant heterogeneity and complex inter- and intra-modal interactions between modalities. Previous methods used co-attention, fusing features only once after separate encoding, which is insufficient to model such a complex task due to modality heterogeneity. To this end, we propose a Biased Progressive Encoding (BPE) paradigm, performing encoding and fusion simultaneously. This paradigm uses one modality as a reference when encoding the other, fostering deep fusion of the modalities through multiple iterations, progressively reducing the cross-modal disparities and facilitating complementary interactions. Besides, survival prediction involves biomarkers from WSIs, genomics, and their integrative analysis. Key biomarkers may exist in different modalities under individual variations, necessitating the model flexibility. Hence, we further propose a Mixture of Multimodal Experts layer to dynamically select tailored experts in each stage of the BPE paradigm. Experts incorporate reference information from another modality to varying degrees, enabling a balanced or biased focus on different modalities during the encoding process. The experimental results demonstrate the superior performance of our method on various datasets, including TCGA-BLCA, TCGA-UCEC and TCGA-LUAD. Codes are available at https://github.com/BearCleverProud/MoME.

**Keywords:** Multimodal Learning · Survival Prediction · Computational Pathology.

## 1 Introduction

Survival analysis via **W**hole **S**lide **I**mages (WSIs) and genomic data is crucial in cancer prognosis as it assesses the risk of death and provides important references

for the treatment plans. The key to this task is how to effectively utilize information from both modalities, for instance, to detect image-omic biomarkers as well as to explore interactions between tumor microenvironment in histopathology images and co-expression of genomic data. In recent years, the focus of research also has shifted from single-modal prediction [1,9,12,14,24] to the more complicated survival analysis utilizing multimodal information [2,5,21,22,26,30,31].

One of the key challenges in this task is the significant heterogeneity between histopathology images and genomic data [13], stemming from their inherent disparities and distinct pre-processing methods. Additionally, the inter- and intra-modal interactions are highly complex, as both modalities possess abundant information, but only a small fraction of them can be mutually correlated and utilized for survival prediction. Previous approaches have attempted to tackle this challenge by using cross-modality attention (co-attention) [25] based methods [2,26,31]. However, feature fusion is conducted only once throughout the entire process. These approaches might be considered shallow given the complexity of the task and the significant differences between the two modalities.

To tackle these issues, we propose a **B**iased **P**rogressive **E**ncoding (BPE) paradigm. Unlike previous methods that encode modalities separately before fusion, our approach simultaneously encodes and fuses features. In this approach, one modality is encoded while utilizing the other modality as a reference, which can assist in extracting more relevant information. Furthermore, the encoding of features from the two modalities is performed alternately, thereby progressively reducing the differences between their feature spaces. This allows for a deeper fusion process and facilitates the discovery of interactions between modalities.

In addition to modality heterogeneity, inter-individual variations can cause key features for survival analysis to appear in different modalities for each patient. This presents a new challenge in designing the model structure, as it requires selective focus on a specific modality or the interactions between the two modalities. To achieve this, we propose a **M**ixture **o**f **M**ultimodal **E**xperts (MoME) layer, employing our BPE paradigm. The MoME layer consists of multiple specialized experts capable of modeling complex inter- and intra-modal correlations. In addition, these experts incorporate reference information from another modality to varying degrees, enabling a focus on different modalities during the encoding process. Moreover, we enable flexible selection of experts in each layer, as the function of reference information may differ. In the shallow layers of the network, our MoME layer could use the reference information as a filter to eliminate task-irrelevant features and enhance relevant ones within each modality. Conversely, in the deeper layers of the network, it could be used as a guidance to seeks cross-modal combination representations as biomarkers.

The contributions of this paper can be summarized as follows:

1. We propose a biased progressive encoding paradigm which integrates information from one modality into the feature encoding of the other modality as a reference for more effective feature extraction and interaction modeling.

2. We design a mixture of multimodal experts layer which enables the network to selectively focus on the information from a specific modality and utilizes the reference information in different forms across encoding stages.
3. We extensively evaluate our method on three TCGA datasets: BLCA, UCEC, and LUAD. The results demonstrate the superior performance of our method.

## 2 Methodology

### 2.1 Problem Formulation

The WSIs are usually formulated under the **A**ttention-**B**ased **M**ultiple **I**nstance **L**earning (AB-MIL) framework [10,23]. This involves dividing the WSIs into bags of patches and extracting features from them using a pre-trained neural network. An MIL aggregator is then used to process the features, generate bag features, and make predictions. The WSI patch features are represented as $\boldsymbol{P} \in \mathbb{R}^{n_p \times d}$, where $n_p$ is the number of patches in the WSI bag and $d$ is the dimension for both WSI and genomic latent features. The genomic data consists of various $1 \times 1$ values, including RNA sequencing, copy number variation, DNA methylation, *etc*. Following previous works [2], the genomic data are categorized into the following genomic sequences: 1) Tumor Suppression, 2) Oncogenesis, 3) Protein Kinases, 4) Cellular Differentiation, 5) Transcription, and 6) Cytokines and Growth. These sequences are stacked and fed into a fully connected layer to obtain genomic features, which are denoted as $\boldsymbol{G} \in \mathbb{R}^{n_g \times d}$, where $n_g = 6$ is the number of genomic sequences. In survival analysis, given the input pair ($\boldsymbol{P}$,$\boldsymbol{G}$), rather than predicting the exact time of death for patients, we initially estimate the hazard function $h(t) = h(T = t | T \geq t, (\boldsymbol{P}, \boldsymbol{G})) \in [0, 1]$, which is the probability of death for a patient right after the time point $t$. Subsequently, an ordinal value is obtained via integrating the negated hazard functions: $s(t|(\boldsymbol{P}, \boldsymbol{G})) = \prod_{u=1}^{t}(1 - h(u))$.

### 2.2 Biased Progressive Encoding Paradigm

The overview of our BPE paradigm is shown in Fig. 1(a). For simplicity, we refer to the modality being encoded as $\boldsymbol{F}_1^{(i)}$, and the reference modality as $\boldsymbol{F}_2^{(i)}$, where $\boldsymbol{F}_1$ could be either $\boldsymbol{P}$ or $\boldsymbol{G}$, and $i$ is the times of encoding. Our BPE paradigm follows a progressive learning strategy, where $\boldsymbol{F}_1^{(i)}$ is encoded to discover the complex interactions with $\boldsymbol{F}_2^{(i)}$ being a reference. This process is then reversed to encode the other modality $\boldsymbol{F}_2^{(i)}$ with the encoded first modality $\boldsymbol{F}_1^{(i+1)}$. The deep feature extraction and progressive learning strategy in our BPE enable deep fusion to reduce the inter-modal discrepancy. The complete progressive encoding process, which involves encoding both modalities, is given as,

$$\boldsymbol{F}_1^{(i+1)} = \mathrm{BPE}_{2i}(\boldsymbol{F}_1^{(i)}, \boldsymbol{F}_2^{(i)}), \qquad \boldsymbol{F}_2^{(i+1)} = \mathrm{BPE}_{2i+1}(\boldsymbol{F}_2^{(i)}, \boldsymbol{F}_1^{(i+1)}), \tag{1}$$

where $\mathrm{BPE}.(\cdot,\cdot)$ is our MoME, detailed in the next section. We encode both features twice for all datasets, facilitating a deep fusion of the two modalities. Following the encoding, both features are fed into an attention block along with a classification token, which is then used for the final survival prediction.

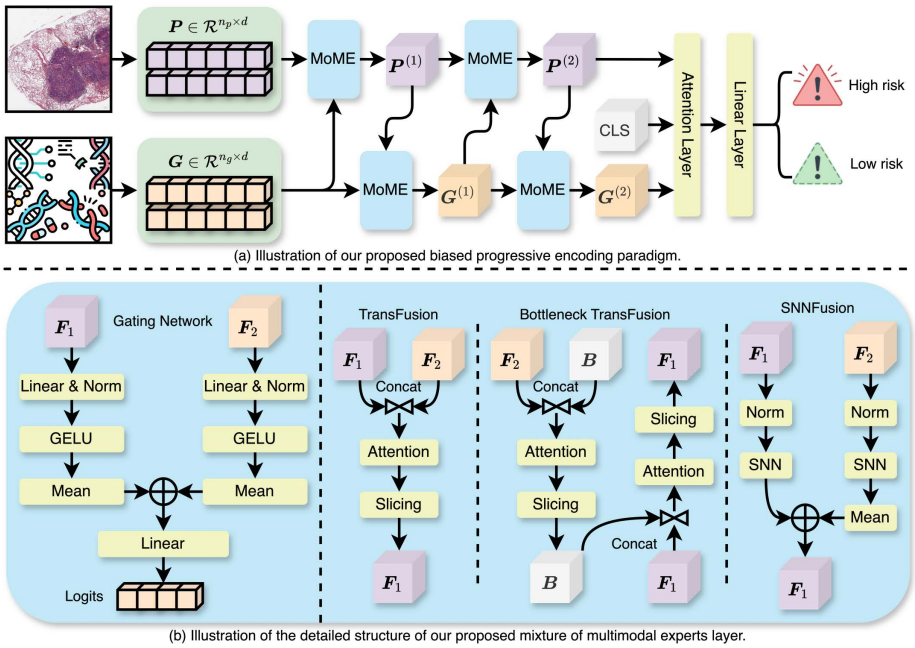


**Fig. 1.** Illustrations of (a) our biased progressive encoding paradigm, and (b) the structure of our mixture of multimodal experts. The left section of (b) represents our gating network, and the right section of (b) depicts our proposed three expert components designed for different degrees of integration of the reference modality.

### 2.3 Mixture of Multimodal Experts

The structure of our MoME layer is depicted in Fig. 1(b). Our MoME layer is derived from the traditional **M**ixture **o**f **E**xperts (MoE) [6,16,19], consisting of a set of parallel feed forward networks (experts) and a gating network that controls the selection of experts. Distinct from the classic MoE that operate at the token level, routing tokens within a sequence to various experts, our MoME innovates by functioning at the sample and layer levels, *i.e.*, different samples within the same layer or identical samples across different layers can be routed to distinct experts. This innovation is pivotal for handling the informative yet sparse characteristics of WSIs and genomics, where isolated features may be nondescript, underscoring the necessity to process them collectively.

Our MoME comprises two components: 1) a gating network and 2) a set of specially designed experts for multimodal survival analysis. Initially, the features are passed to the gating network, which determines the most suitable expert to utilize. The selected expert then performs fusion and encoding for $\boldsymbol{F}_1^{(i)}$.

**Gating Network.** The gating network is designed to be lightweight yet informative to select experts. It consists of linear layers, **G**aussian **E**rror **L**inear **U**nits

(GELUs) [8], and **R**ooted **M**ean **S**quare Layer **Norm**alization (RMSNorm) layers [28]. These modules map features into a new space, and the mapped features are averaged to obtain the multimodal one. The gating network is given as,

$$\begin{aligned}\text{logits} = \boldsymbol{W} \cdot (\text{mean}(\text{GELU}(\text{RMSNorm}(\boldsymbol{W}_1\boldsymbol{F}_1)))) \\ + \boldsymbol{W} \cdot (\text{mean}(\text{GELU}(\text{RMSNorm}(\boldsymbol{W}_2\boldsymbol{F}_2)))),\end{aligned} \tag{2}$$

where $\boldsymbol{W}_1, \boldsymbol{W}_2$ and $\boldsymbol{W}$ are learnable matrices. The obtained logits are utilized to select the appropriate expert. Unlike traditional MoE models that employ a weighted sum of multiple experts [16], our approach enforces the selection of only one expert within the module [27]. By adopting this strategy, the gating network can make careful expert choices while also reducing computational costs.

**Multimodal Expert Pool.** We design the following four experts based on the two principles: 1) inclusion of experts specializing in WSI, genomics, and interactions between them, and 2) capability of simultaneous fusion and encoding.

*TransFusion.* This multimodal expert is based on self-attention. This expert maximizes the utilization of the reference modality by enabling full information exchange between the two modalities with self-attention. Given the input pair $(\boldsymbol{F}_1, \boldsymbol{F}_2)$, our proposed **T**rans**F**usion (TF) expert can be expressed as,

$$\text{TF}(\boldsymbol{F}_1, \boldsymbol{F}_2) = \text{SA}([\boldsymbol{F}_1, \boldsymbol{F}_2])[:n_1, :], \tag{3}$$

where $\text{SA}(\cdot)$ is the **S**elf-**A**ttention [20], $[\boldsymbol{F}_1, \boldsymbol{F}_2] \in R^{(n_1+n_2)\times d}$ is the concatenation of $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$, $[:n_1, :]$ indicates to select the first $n_1$ rows of the matrix, and $n_1, n_2$ denote the number of features of the two modalities $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$.

*Bottleneck TransFusion.* Both modalities contain a vast amount of information, but only a small portion of them is useful for survival analysis. Therefore, it is necessary to distill the information and focus only on the portions that are pertinent to the outcome. To tackle this, we propose **B**ottleneck **T**rans**F**usion (BTF) expert that avoids direct interactions between the two modalities with the bottleneck features acting as independent features communicating between the features of the two modalities [17]. The reference modality is lower utilized compared to TransFusion as it does not allow complete mutual communication of the two modalities. Let $\boldsymbol{B} \in \mathbb{R}^{n_b \times d}$ denote the bottleneck features, where $n_b$ is the number of bottleneck features, the BTF expert is given as,

$$\text{BTF}(\boldsymbol{F}_1, \boldsymbol{F}_2) = \text{SA}_1(\boldsymbol{F}_1, \text{SA}_2(\boldsymbol{B}, \boldsymbol{F}_2)[:n_b, :])[:n_1, :]. \tag{4}$$

*SNNFusion.* The **S**NN**F**usion (SF) expert is designed for fusion that is genomic dominant, as it has promising results when applied solely to genomic data. The utilization of reference modality is even lower when adopting this expert. In SNNFusion, there are two SNNs [12] and our SF expert is given as,

$$\text{SF}(\boldsymbol{F}_1, \boldsymbol{F}_2) = \text{SNN}_1(\text{RMSNorm}(\boldsymbol{F}_1)) + \text{mean}(\text{SNN}_2(\text{RMSNorm}(\boldsymbol{F}_2))), \tag{5}$$

where $\text{SNN}(\cdot)$ consists of a linear layer, an **E**xponential **L**inear **U**nit (ELU) activation layer [4] and an alpha dropout layer [12].

*DropF2Fusion.* This expert drops $\boldsymbol{F}_2$ during fusion, serving as a skip layer, as it completely stops utilizing $\boldsymbol{F}_2$. It is particularly useful when using one modality is accurate enough. Mathematically, the **D**ropF2**F**usion (DF) expert is given as,

$$\mathrm{DF}(\boldsymbol{F}_1, \boldsymbol{F}_2) = \boldsymbol{F}_1. \tag{6}$$

### 2.4 Why Self-attention over Co-attention

Mathematically, the **C**o-**A**ttention (CA) and SA can be given as,

$$\mathrm{CA}(\boldsymbol{F}_1, \boldsymbol{F}_2) = \mathrm{Softmax}(\frac{(\boldsymbol{F}_1\boldsymbol{Q})(\boldsymbol{F}_2\boldsymbol{K})^T}{\sqrt{d}})(\boldsymbol{F}_2\boldsymbol{V}). \tag{7}$$

$$\begin{aligned} \mathrm{SA}([\boldsymbol{F}_1, \boldsymbol{F}_2]) &= \mathrm{Softmax}(\frac{\begin{bmatrix}\boldsymbol{F}_1\boldsymbol{Q}\\ \boldsymbol{F}_2\boldsymbol{Q}\end{bmatrix}\begin{bmatrix}(\boldsymbol{F}_1\boldsymbol{K})^T & (\boldsymbol{F}_2\boldsymbol{K})^T\end{bmatrix}}{\sqrt{d}})\begin{bmatrix}\boldsymbol{F}_1\boldsymbol{V}\\ \boldsymbol{F}_2\boldsymbol{V}\end{bmatrix}, \\ &= \mathrm{Softmax}(\frac{\begin{bmatrix}(\boldsymbol{F}_1\boldsymbol{Q})(\boldsymbol{F}_1\boldsymbol{K})^T & \{(\boldsymbol{F}_1\boldsymbol{Q})(\boldsymbol{F}_2\boldsymbol{K})^T\}\\ (\boldsymbol{F}_2\boldsymbol{Q})(\boldsymbol{F}_1\boldsymbol{K})^T & (\boldsymbol{F}_2\boldsymbol{Q})(\boldsymbol{F}_2\boldsymbol{K})^T\end{bmatrix}}{\sqrt{d}})\begin{bmatrix}\boldsymbol{F}_1\boldsymbol{V}\\ \{\boldsymbol{F}_2\boldsymbol{V}\}\end{bmatrix}. \end{aligned} \tag{8}$$

The matrix multiplication result of the red brace portions in Eq. (8) matches $\mathrm{CA}(\boldsymbol{F}_1, \boldsymbol{F}_2)$ in Eq. (7). $\mathrm{CA}(\boldsymbol{F}_2, \boldsymbol{F}_1)$, $\mathrm{SA}(\boldsymbol{F}_1)$ and $\mathrm{SA}(\boldsymbol{F}_2)$ are also embedded in Eq. (8). Hence, we can conclude that CA is a sub-optimal substitute for SA. Therefore, we design the experts based on self-attention instead of co-attention.

## 3 Experiments and Results

### 3.1 Datasets

**T**he **C**ancer **G**enome **A**tlas (TCGA[6]) project provides extensive information on patients under study, including WSIs, genomic data, and ground truth survival time. The datasets used in our experiments include 373 samples of **B**ladder **U**rothelial **CA**rcinoma (BLCA), 480 samples of **U**terine **C**orpus **E**ndometrial **C**arcinoma (UCEC), and 453 samples of **LU**ng **AD**enocarcinoma (LUAD).

### 3.2 Implementation Details

**Training Settings.** We select a wide range of baseline methods, including those focused on genomic data, WSIs, and both modalities. The methods for comparison are: SNN [12], SNNTrans [12], AttnMIL [10], CLAM [15], TransMIL [18], DTFD-MIL [29], Porpoise [3], MCAT [2], MOTCAT [26] and CMTA [31]. We choose the **C**oncordance index (C-index), a commonly employed metric in survival analysis, as our evaluation metric. To evaluate the performance of these methods, we conduct a five-fold cross-validation. Each model is training for 20 epochs, and the best validation performance obtained among these epochs is considered as the final performance for the respective fold. The means and standard deviations of the C-index for each method on different datasets are reported.

[6] https://www.cancer.gov/tcga

**Table 1.** C-index Results for different methods on three TCGA datasets. The best results are underlined in red, while the second best are italicized in blue. "Geno." denotes the utilization of genomic profiles, and "Patho." signifies the use of WSIs.

| | Modality | | Dataset | | | |
|---|---|---|---|---|---|---|
| | Geno. | Patho. | BLCA | UCEC | LUAD | Overall |
| SNN [12] | ✓ | | 0.618±0.022 | 0.679±0.040 | 0.611±0.047 | 0.636 |
| SNNTrans [12] | ✓ | | 0.659±0.032 | 0.656±0.038 | 0.638±0.022 | 0.651 |
| AttnMIL [10] | | ✓ | 0.599±0.048 | 0.658±0.036 | 0.620±0.061 | 0.626 |
| CLAM-SB [15] | | ✓ | 0.559±0.034 | 0.644±0.061 | 0.594±0.063 | 0.599 |
| CLAM-MB [15] | | ✓ | 0.565±0.027 | 0.609±0.082 | 0.582±0.072 | 0.585 |
| TransMIL [18] | | ✓ | 0.575±0.034 | 0.655±0.046 | 0.642±0.046 | 0.624 |
| DTFD-MIL [29] | | ✓ | 0.546±0.021 | 0.656±0.045 | 0.585±0.066 | 0.596 |
| MCAT [2] | ✓ | ✓ | 0.672±0.032 | 0.649±0.043 | 0.659±0.027 | 0.660 |
| Porpoise [3] | ✓ | ✓ | 0.636±0.024 | *0.692*±0.048 | 0.647±0.031 | 0.658 |
| MOTCAT [26] | ✓ | ✓ | *0.682*±0.023 | 0.671±0.053 | 0.673±0.040 | 0.675 |
| CMTA [31] | ✓ | ✓ | 0.672±0.038 | 0.691±0.066 | *0.676*±0.037 | *0.680* |
| MoME (Ours) | ✓ | ✓ | 0.686±0.041 | 0.706±0.038 | 0.691±0.040 | 0.694 |

**Hyper-parameters.** Adam [11] optimizer is used in our experiment. The learning rate and weight decay are set to $2\times10^{-4}$ and $1\times10^{-5}$, respectively [26]. The WSIs are split into patches sized of 256 × 256 pixels at 20× magnification and ResNet-50 [7] pre-trained on ImageNet is used to extract features from them. The number of bottleneck features in our experiments $n_b$ is 2. We utilize the micro-batch technique [26] and the size of the micro-batch is 4,096.

### 3.3 Comparison Results

We conduct a comprehensive comparison of our method with both unimodal methods and other state-of-the-art multimodal methods. The results are presented in Table 1. Our method consistently outperforms all other methods across all datasets, particularly on the UCEC and LUAD datasets, where it exhibits a significant performance advantage over previous approaches. Our method achieves improvements of 0.4%, 1.4%, and 1.5% on the C-index of the three datasets, respectively, compared to previous methods, as well as a 1.4% improvement in overall performance. These results suggest that our MoME is applicable to general survival analysis settings. Notably, despite their simple structures, methods based on genomic data outperform those based on WSIs, highlighting the importance of genomic data in survival analysis. Furthermore, multimodal methods consistently outperform unimodal methods, further demonstrating the efficacy of multimodal approaches and the necessity of incorporating multiple modalities.

### 3.4 Ablation Studies

**Choices of Experts.** We conduct experiments to assess the effectiveness of each expert by deactivating them individually. The experimental results are presented

**Table 2.** Ablation study results for using different experts in our MoME (left) and sensitivity analysis in the number of bottleneck features (right). "T." represents TransFusion, "B." represents Bottleneck TransFusion, "S." represents SNNFusion, "D." represents DropF2Fusion, and "#B." denotes the number of bottleneck features.

| | Experts | | | | Dataset | | #B. | Dataset | |
|---|---|---|---|---|---|---|---|---|---|
| | T. | B. | S. | D. | UCEC | LUAD | | UCEC | LUAD |
| MoME | ✓ | ✓ | ✓ | ✓ | *0.706*±0.038 | 0.691±0.040 | 1 | 0.690±0.029 | 0.669±0.029 |
| MoME | | ✓ | ✓ | ✓ | 0.693±0.041 | 0.673±0.043 | 2 | *0.706*±0.038 | 0.691±0.040 |
| MoME | ✓ | | ✓ | ✓ | 0.703±0.052 | 0.663±0.064 | 4 | 0.699±0.053 | *0.673*±0.040 |
| MoME | ✓ | ✓ | | ✓ | 0.738±0.060 | 0.669±0.045 | 8 | 0.717±0.072 | 0.658±0.035 |
| MoME | ✓ | ✓ | ✓ | | 0.684±0.053 | *0.685*±0.047 | 16 | 0.704±0.044 | 0.662±0.034 |
| TF | ✓ | | | | 0.690±0.059 | 0.655±0.061 | / | / | / |

in the left part of Table 2. We observe that our MoME with all experts achieves the best performance on LUAD and the second-best performance on UCEC. The MoME without the SNNFusion expert achieves the best performance on UCEC. These results indicate that the most crucial components for UCEC and LUAD differ, suggesting that a MoME focusing on a specific modality could be beneficial. This further supports the necessity of our MoME, which dynamically selects different experts for different samples. Additionally, the performance of our MoME compared to TransFusion reaffirms its superiority.

**Sensitivity Analysis on BTF.** We conduct a sensitivity analysis on UCEC and LUAD by varying the number of bottleneck features used in BTF from 1 to 16. The results are presented in the right part of Table 2. We observe that our MoME achieves the overall best performance when the number of bottleneck features is 2, which strikes a balance between these two datasets. Specifically, our MoME achieves the best performance on UCEC when the number of bottleneck features is 8, however, its performance on LUAD is the lowest.

## 4 Conclusion

In this paper, we introduced a BPE paradigm and a MoME layer for cancer survival analysis. The BPE paradigm enables deep fusion by performing feature encoding and fusion simultaneously, leveraging one modality as a reference to encode the other. With this, our BPE can addresses the severe heterogeneity between WSI and genomic features. Additionally, our MoME layer dynamically selects the most appropriate expert to model the intricate inter- and intra-modal correlations, addressing the challenges posed by variations in key features. Through extensive experiments, we demonstrated that our method outperforms other multimodal learning approaches in survival prediction, and the results suggest that our method could be applied to a general survival analysis setting.

**Acknowledgments.** This research was done with Jarvis Research Center, Tencent YouTu Lab and partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (CUHK 14222922, RGC GRF 2151185).

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine **25**(8), 1301–1309 (2019) 2
2. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: IEEE/CVF International Conference on Computer Vision. pp. 4015–4025 (2021) 2, 3, 6, 7
3. Chen, R.J., Lu, M.Y., Williamson, D.F., Chen, T.Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., Mahmood, F.: Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell **40**(8), 865–878.e6 (2022) 6, 7
4. Clevert, D., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). In: International Conference on Learning Representations (2016) 5
5. Ding, K., Zhou, M., Metaxas, D.N., Zhang, S.: Pathology-and-genomics multimodal transformer for survival outcome prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 622–631 (2023) 2
6. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research **23**, 120:1–120:39 (2022) 4
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016) 7
8. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv:1606.08415 (2016) 5
9. Ianni, J.D., Soans, R.E., Sankarapandian, S., Chamarthi, R.V., Ayyagari, D., Olsen, T.G., Bonham, M.J., Stavish, C.C., Motaparthi, K., Cockerell, C.J., et al.: Tailored for real-world: a whole slide image classification system validated on uncurated multi-site data emulating the prospective pathology workload. Scientific Reports **10**(1), 3217 (2020) 2
10. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning. pp. 2127–2136 (2018) 3, 6, 7
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) International Conference on Learning Representations (2015) 7
12. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-normalizing neural networks. In: Advances in Neural Information Processing Systems. vol. 30 (2017) 2, 5, 6, 7

13. Li, R., Wu, X., Li, A., Wang, M.: HFBSurv: hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction. Bioinformatics **38**(9), 2587–2594 (Feb 2022) 2
14. Litjens, G., Sánchez, C.I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-Van De Kaa, C., Bult, P., Van Ginneken, B., Van Der Laak, J.: Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Scientific Reports **6**(1), 26286 (2016) 2
15. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021) 6, 7
16. Masoudnia, S., Ebrahimpour, R.: Mixture of experts: a literature survey. Artif. Intell. Rev. **42**(2), 275–293 (2014) 4, 5
17. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. In: Advances in Neural Information Processing Systems. pp. 14200–14213 (2021) 5
18. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., Zhang, Y.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In: Advanced in Neural Information Processing Systems (2021) 6, 7
19. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q.V., Hinton, G.E., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: International Conference on Learning Representations (2017) 4
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017) 5
21. Wang, Z., Ma, J., Gao, Q., Bain, C., Imoto, S., Liò, P., Cai, H., Chen, H., Song, J.: Dual-stream multi-dependency graph neural network enables precise cancer survival analysis. Medical Image Analysis p. 103252 (2024) 2
22. Xing, X., Chen, Z., Zhu, M., Hou, Y., Gao, Z., Yuan, Y.: Discrepancy and gradient-guided multi-modal knowledge distillation for pathological glioma grading. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 636–646 (2022) 2
23. Xiong, C., Chen, H., Sung, J.J.Y., King, I.: Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. In: International Joint Conference on Artificial Intelligence. pp. 1587–1595 (2023) 3
24. Xiong, C., Lin, Y., Chen, H., Sung, J.J.Y., King, I.: Knowledge transfer via multi-head feature adaptation for whole slide image classification. arXiv preprint arXiv:2303.05780 (2023) 2
25. Xu, P., Zhu, X., Clifton, D.A.: Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 12113–12132 (2023) 2
26. Xu, Y., Chen, H.: Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In: IEEE/CVF International Conference on Computer Vision. pp. 21241–21251 (October 2023) 2, 6, 7
27. Xue, Z., Marculescu, R.: Dynamic Multimodal Fusion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 2575–2584 (Jun 2023) 5
28. Zhang, B., Sennrich, R.: Root mean square layer normalization. In: Advances in Neural Information Processing Systems. pp. 12360–12371 (2019) 5

29. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: DTFD-MIL: double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18780–18790. IEEE (2022) 6, 7
30. Zhang, Y., Xu, Y., Chen, J., Xie, F., Chen, H.: Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. In: International Conference on Learning Representations (2024) 2
31. Zhou, F., Chen, H.: Cross-modal translation and alignment for survival analysis. In: IEEE/CVF International Conference on Computer Vision. pp. 21485–21494 (2023) 2, 6, 7