# DualGGR-MoME: A Mixture of Multimodal Experts with Dual-Branch Gating and Global-Guided Routing for Cancer Survival Prediction

Kaiwen Tan[1,2], Jie Feng[1,2], Siqi Zhang[1,2⋆], Yongbing Zhang[1,2], Yuxin Huang[1,2], and Zhengtao Yu[1,2]

[1] Faculty of Information Engineering and Automation, Kunming University of Science and Technology
[2] Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology
zhangsiqi.mo@foxmail.com

**Abstract.** Mixture-of-Multimodal-Experts (MoME) models have shown strong performance in cancer survival prediction, as different experts can either integrate modalities in a balanced manner or focus on specific features during encoding. However, real-world multimodal cancer data inevitably contain substantial noise, arising from issues such as read errors and missing calls in gene sequencing, as well as uneven preparation and staining of pathological slides. Existing MoME models compute expert weights directly from inputs, which exposes weight computation to input noise, thereby degrading predictive performance and stability. To address this problem, we propose DualGGR-MoME, a Mixture of Multimodal Experts with Dual-Branch Gating and Global-Guided Routing for Cancer Survival Prediction, which incorporates Dual-Branch Gating (DBG) and Global-Guided Routing (GGR) to achieve a noise-robust expert routing. Unlike conventional MoME methods, we decouple the computation of expert weights into two stages: gating and routing, where gating is responsible for generating noise-robust routing signals and routing is responsible for computing the final noise-robust expert weights. Extensive experiments demonstrate that, compared with MoME, DualGGR-MoME improves the C-index by 3.98%, 5.08%, and 2.57% on the TCGA-BLCA, TCGA-LUAD, and TCGA-UCEC datasets, respectively. Noise-robustness analyses further validate the effectiveness of DualGGR-MoME.

**Keywords:** Cancer Survival Prediction · Mixture of Experts · Noise-Robust Routing.

## 1 Introduction

Whole slide images (WSIs) characterize macro-level tumor microenvironment phenotypes, whereas genomics data reveal microscopic molecular drivers [1].

⋆ Corresponding author

Therefore, integrating WSIs and genomics for survival prediction has become a major research focus. Existing methods fall into three categories: **Early fusion**, directly concatenating multimodal representations [2, 3]; **Late fusion**, aggregating predictions from independent models [4] and **Intermediate fusion**, using attention or graph networks for feature alignment [5, 6, 7, 8]. However, these methods mostly rely on fixed strategies lacking sample-specific adaptability. Recent Mixture of Experts (MoME) methods assign adaptive fusion weights to address this limitation. [9, 10, 11]. Yet real-world multimodal cancer data inevitably contain substantial noise. For instance, genomic data often contain sequencing errors and missing values, while pathological images suffer from variations in tissue preparation and staining. Existing MoME methods compute expert weights directly from inputs, which exposes weight computation to input noise. Consequently, noise distorts expert weights computation, causing expert weight misallocation and degraded stability.

To address these limitations, we propose DualGGR-MoME, which integrates Dual-Branch Gating (DBG) and Global-Guided Routing (GGR) for noise-robust routing. Instead of computing expert weights from the raw input, we condition them on the experts' outputs together with a global sample representation. This decouples the unified router of existing MoME methods [9, 10] into two stages: gating, which generates routing signals from expert outputs, and routing, which computes the final weights. DBG and GGR are the concrete instantiations of these two stages. Our main contributions are:

1. **Noise-Robust MoME Framework.** We propose a new MoME framework that decouples routing into two stages: gating for routing signal generation and routing for expert weights computation.
2. **Dual-Branch Gating and Global-Guided Routing.** We introduce Dual-Branch Gating (DBG) and Global-Guided Routing (GGR) to implement the noise-robust MoME framework. DBG generates multi-dimensional routing signals from expert outputs, avoiding direct exposure to input noise. GGR integrates these signals with global sample representations through adaptive fusion and regularization to compute robust expert weights.
3. **Noise-Robust Validation.** We designed comprehensive perturbation experiments (Gaussian, dropout, and outlier noise) on three TCGA datasets: BLCA, UCEC, and LUAD. Results demonstrate that our method significantly enhances model robustness.

## 2 Methodology

As illustrated in Fig. 1, DualGGR-MoME consists of four components: an input representation module, a DualGGR-MoME Module, an Attention Fusion module, and a Survival Prediction module.

### 2.1 Input Representation Module

Inputs include WSI representation $\mathbf{P} \in \mathbb{R}^{N_p \times d}$ and genomics representation $\mathbf{G} \in \mathbb{R}^{N_g \times d}$, where $d$ is the feature dimension, $N_p$ and $N_g$ denote the number

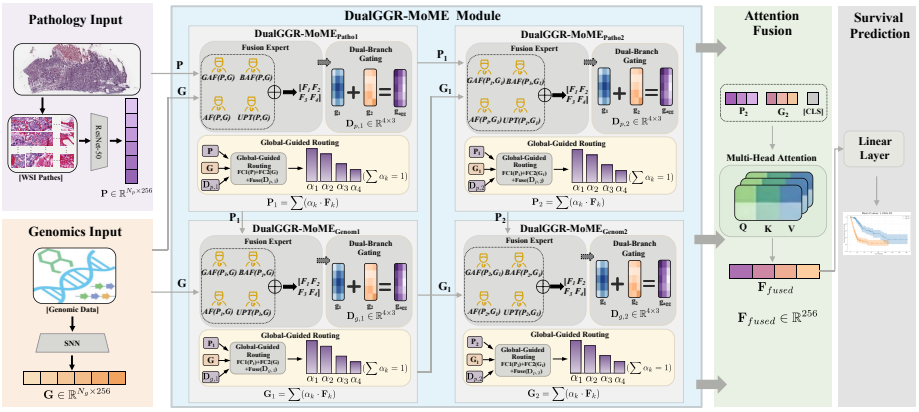


**Fig. 1.** Overview of DualGGR-MoME.

of patches and biological pathways [5], respectively. Specifically, each WSI is segmented into $256 \times 256$ patches and encoded via a pretrained ResNet-50 to obtain $\mathbf{P}$ (via CLAM [12]). The raw multi-omics data are organized by biological pathway annotations and projected through SNN to obtain $\mathbf{G}$.

### 2.2 DualGGR-MoME Module

MoME couples signal extraction and weight computation in a single input-conditioned router. Consequently, input noise enters the router directly and distorts the resulting weights. We instead separate the gating stage, which drives routing, from the routing stage, which computes the weights, so that no stage takes the raw input directly as its routing argument. Gating derives routing signals from expert outputs, where raw input perturbations are no longer directly exposed; routing then computes the final weights from these signals together with a global sample representation. The DualGGR-MoME module enables noise-robust expert fusion by processing expert outputs to generate routing signals and integrating them with global sample representations. As illustrated in Fig. 2, this module comprises three core components: **(i)** four parallel multimodal fusion experts with diverse strategies (Fig. 2a); **(ii)** DBG that generates routing signals from each expert's outputs (Fig. 2b) and **(iii)** GGR that computes robust expert weights by fusing routing signals with global sample representations (Fig. 2c).

**Multimodal Fusion Experts.** Following previous work [9], we design four experts operating on a target modality $\mathbf{x}_1$ and reference $\mathbf{x}_2$. The *Global Attention Fusion (GAF)* concatenates $[\mathbf{x}_1; \mathbf{x}_2]$ for unified self-attention, enabling deep interactions. The *Bottleneck Attention Fusion (BAF)* employs $n_b$ tokens $\mathbf{B}$ to first compress $\mathbf{x}_2$ via attention and then interact with $\mathbf{x}_1$, facilitating selective extraction without over-reliance. The *Additive Fusion (AF)* efficiently broadcasts auxiliary context by adding the global average of SNN-processed $\mathbf{x}_2$ to $\mathbf{x}_1$. The

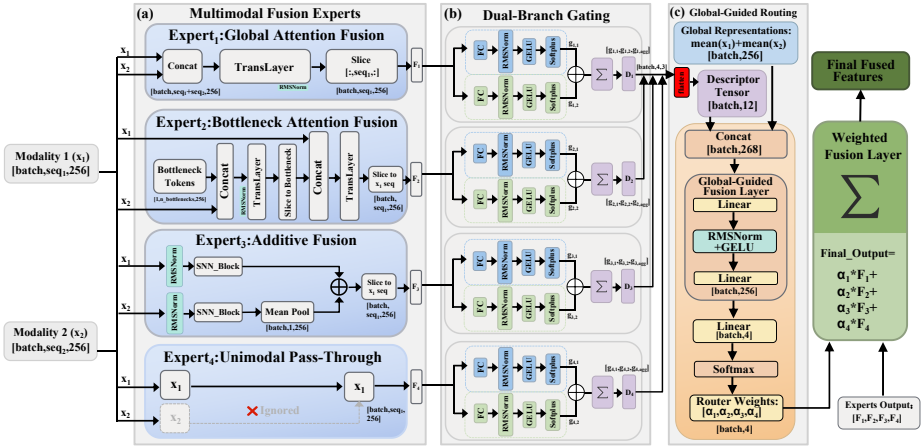


**Fig. 2.** Architecture of DualGGR-MoME. (a) Four parallel multimodal fusion experts. (b) DBG generates routing signals from expert outputs. (c) GGR integrates routing signals with global sample representations to compute robust weights for expert fusion.

*Unimodal Pass-Through (UPT)* serves as a conservative strategy, preserving $\mathbf{x}_1$ without incorporating the reference modality to avoid introducing noise interference.

**Dual-Branch Gating.** To generate multi-dimensional routing signals from expert outputs, we employ dual-branch gating. Let $\mathbf{F}_k \in \mathbb{R}^{\text{seq}_1 \times d}$ denote the output matrix of the $k$-th expert, where $\mathbf{F}_k^{(i)} \in \mathbb{R}^d$ is the feature vector of the $i$-th token (patch for WSI, pathway for genomics), and $\text{seq}_1$ is the target modality's sequence length ($\text{seq}_1 = N_p$ for WSI, $\text{seq}_1 = N_g$ for genomics). DBG generates two routing signals from the globally pooled representation $\bar{\mathbf{F}}_k = \frac{1}{\text{seq}_1} \sum_{i=1}^{\text{seq}_1} \mathbf{F}_k^{(i)} \in \mathbb{R}^d$:

$$g_{k,1} = f_1(\bar{\mathbf{F}}_k), \quad g_{k,2} = f_2(\bar{\mathbf{F}}_k) \tag{1}$$

where $f_1, f_2$ are 2-layer MLPs with RMSNorm, GELU, and Softplus activations, producing $g_{k,1}$ and $g_{k,2}$ from separate branches respectively. The two signals are summed to form $g_{k,agg} = g_{k,1} + g_{k,2}$, yielding the routing signal vector $\mathbf{D}_k = [g_{k,1}, g_{k,2}, g_{k,agg}] \in \mathbb{R}^3$ for robust expert fusion.

**Global-Guided Routing**. This routing computes expert weights by integrating routing signals with global sample representations. To capture holistic sample characteristics, modal representations $\mathbf{P}$ and $\mathbf{G}$ are processed by independent MLPs, $\text{fc}_P$ and $\text{fc}_G$. Representations are then compressed via global mean pooling and summed to yield a global context vector $\mathbf{h}$:

$$\mathbf{h} = \text{Mean}\big(\text{fc}_P(\mathbf{P})\big) + \text{Mean}\big(\text{fc}_G(\mathbf{G})\big) \in \mathbb{R}^d \tag{2}$$

When DBG is enabled during training, the routing network also receives the routing signal tensor $\mathbf{D} \in \mathbb{R}^{4\times 3}$ from all experts. The global context vector $\mathbf{h}$

is concatenated with the flattened signal tensor $\mathbf{D}_{\text{flat}} \in \mathbb{R}^{12}$ and processed by a signal fusion layer:

$$\mathbf{h}' = f_{\text{fusion}}\big([\mathbf{h}; \mathbf{D}_{\text{flat}}]\big) \in \mathbb{R}^d \tag{3}$$

where $[\cdot;\cdot]$ denotes concatenation, and $f_{\text{fusion}}$ is a fusion layer that integrates global representations with expert-generated routing signals. The fused representation $\mathbf{h}'$ is then used to compute expert weights via a classifier for weighted summation:

$$\boldsymbol{\alpha} = \text{Softmax}(f_{cls}(\mathbf{h}')) \in \mathbb{R}^4, \quad \mathbf{P}' = \sum\nolimits_{k=1}^{4} \alpha_k \cdot \mathbf{F}_k \tag{4}$$

where $f_{\text{cls}}$ is a single-layer linear projection, and $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3, \alpha_4]$. This design enables the routing to combine global sample representations with expert routing signals to compute noise-robust expert weights.

**Bidirectional Encoding.** Following MoME [9], we employ bidirectional alternating encoding to progressively deepen multimodal interactions. Given initial representations $\mathbf{P}$ and $\mathbf{G}$, the encoding proceeds in two iterations:

**First iteration:**

$$\begin{aligned} \mathbf{P}_1, \mathbf{D}_{p,1} &= \text{DualGGR-MoME}_{\text{Patho1}}(\mathbf{P}, \mathbf{G}), \\ \mathbf{G}_1, \mathbf{D}_{g,1} &= \text{DualGGR-MoME}_{\text{Genom1}}(\mathbf{G}_0, \mathbf{P}_1). \end{aligned} \tag{5}$$

**Second iteration:**

$$\begin{aligned} \mathbf{P}_2, \mathbf{D}_{p,2} &= \text{DualGGR-MoME}_{\text{Patho2}}(\mathbf{P}_1, \mathbf{G}_1), \\ \mathbf{G}_2, \mathbf{D}_{g,2} &= \text{DualGGR-MoME}_{\text{Genom2}}(\mathbf{G}_1, \mathbf{P}_2). \end{aligned} \tag{6}$$

where each DualGGR-MoME module takes the first argument as target modality and the second as reference modality, producing enhanced representations and routing signal tensors ($\mathbf{D}_{p,i}$ for pathology, $\mathbf{D}_{g,i}$ for genomics, where $i$ denotes the iteration) for computing expert weights.

### 2.3 Attention Fusion module

After bidirectional encoding through the DualGGR-MoME Module, the resulting $\mathbf{P}_2$ and $\mathbf{G}_2$ are fused via multi-head attention to generate $F_{fused} \in \mathbb{R}^{256}$ for survival prediction.

### 2.4 Survival Prediction module

A linear layer takes $\mathbf{F}_{fused}$ as input and outputs hazard probabilities $\mathbf{h} = [h_1, h_2, ..., h_K] \in \mathbb{R}^K$ for $K$ discretized time intervals, where $h_k$ represents the conditional probability of an event in the $k$-th interval.

### 2.5 Loss Functions

To enable robust multi-expert fusion, we optimize the model using a composite objective that balances survival risk modeling, routing signal regularization, and expert diversity. The total loss function is defined as:

$$L_{total} = L_{surv} + \lambda_s L_{signal} + \lambda_e L_{entropy} \tag{7}$$

where $\lambda_s$ and $\lambda_e$ serve as balancing coefficients.

**Survival Risk Loss ($L_{surv}$).** We employ a discrete-time survival analysis framework with Negative Log-Likelihood (NLL) loss [13, 14]. The survival probability at time bin $k$ is computed as $S_k = \prod_{j=1}^{k}(1 - h_j)$, representing the cumulative probability of surviving beyond interval $k$. For patient $i$ with observed time bin $y_i$ and censorship status $c_i \in \{0, 1\}$, the NLL loss is formulated as:

$$L_{surv} = -\frac{1}{N}\sum_{i=1}^{N}[(1 - c_i)(\log S_i(y_i) + \log h_i(y_i)) + c_i \log S_i(y_i + 1)] \tag{8}$$

where $N$ is the number of patients, $S_i(y_i)$ denotes the survival probability at the observed time, and $h_i(y_i)$ is the hazard at that time.

**Routing Signal Loss ($L_{signal}$).** To create learnable discriminative patterns for routing, we introduce a regularization term across all modules in $M = 4$ DualGGR-MoME modules (Patho1, Patho2, Genom1, Genom2):

$$L_{signal} = \frac{1}{4}\sum_{m \in M}\sum_{k=1}^{4} \alpha_k^{(m)} \cdot g_{k,agg}^{(m)} \tag{9}$$

where $g_{k,agg}^{(m)} = g_{k,1}^{(m)} + g_{k,2}^{(m)}$ denotes the aggregated routing signal for expert $k$ in module $m$. By minimizing this weighted sum, the formulation promotes expert differentiation. Since $\partial L_{signal}/\partial g_k \propto \alpha_k$, experts with higher routing weights face stricter constraints, while lower-weight experts experience milder regularization. This asymmetric pressure breaks signal uniformity, providing discriminative information for robust weight computation.

**Expert Diversity Loss ($L_{entropy}$).** To prevent over-reliance on a single expert, we apply a weight entropy regularization:

$$L_{entropy} = \frac{1}{4}\sum_{m=1}^{4}(\log 4 - H(\alpha^{(m)})) \tag{10}$$

where $H(\alpha^{(m)}) = -\sum_{k=1}^{4} \alpha_k^{(m)} \log \alpha_k^{(m)}$ is the entropy of routing weights in module $m$. This regularization encourages diverse expert utilization across samples, preventing any single expert from dominating the predictions.

# 3 Experiments and Results

## 3.1 Datasets and Implementation Details

We utilize multimodal datasets from TCGA [13], comprising whole slide images (WSIs), multi-omics data and clinical labels, for three cancer cohorts: Bladder Urothelial Carcinoma (BLCA, $n = 373$), Uterine Corpus Endometrial Carcinoma (UCEC, $n = 480$), and Lung Adenocarcinoma (LUAD, $n = 453$). The model is trained end-to-end using the Adam optimizer [14] with lr=$2 \times 10^{-4}$ and weight decay=$1 \times 10^{-5}$ for 20 epochs [15]. To prevent memory overflow, we employ a mini-batch technique [15] with a size of 8.

## 3.2 Comparative Experiments

We systematically benchmark DualGGR-MoME against 11 state-of-the-art methods, including unimodal methods (ABMIL [16], TransMIL [17], SNN [18], SNNTrans [18]) and multimodal methods (PIBD [19], CMTA [20], MMP [21], TMSE [22], ProSurv [23],M2Surv [24], MoME [9]). As illustrated in Table 1, DualGGR-MoME consistently achieves superior performance across all cohorts. with improvements of 3.98%, 5.08%, and 2.57% over the second-best methods, respectively.

**Table 1.** Performance comparison on TCGA datasets. The C-index is reported using 5-fold cross-validation. **P.**: Pathomics (WSI); **G.**: Genomics (Omics).

| **Methods** | **Modality** | | **BLCA** | **LUAD** | **UCEC** |
|---|---|---|---|---|---|
| | P. | G. | ($n = 373$) | ($n = 453$) | ($n = 480$) |
| ABMIL [16] | ✓ | | $0.624 \pm 0.032$ | $0.602 \pm 0.043$ | $0.615 \pm 0.014$ |
| TransMIL [17] | ✓ | | $0.609 \pm 0.034$ | $0.626 \pm 0.029$ | $0.602 \pm 0.030$ |
| SNN [18] | | ✓ | $0.587 \pm 0.032$ | $0.617 \pm 0.033$ | $0.675 \pm 0.039$ |
| SNNTrans [18] | | ✓ | $0.583 \pm 0.061$ | $0.635 \pm 0.049$ | $0.626 \pm 0.032$ |
| PIBD [19] | ✓ | ✓ | $0.667 \pm 0.061$ | $0.712 \pm 0.063$ | $0.673 \pm 0.056$ |
| CMTA [20] | ✓ | ✓ | $0.681 \pm 0.042$ | $0.686 \pm 0.036$ | $0.698 \pm 0.041$ |
| MMP [21] | ✓ | ✓ | $0.635 \pm 0.051$ | $0.642 \pm 0.037$ | $0.666 \pm 0.045$ |
| TMSE [22] | ✓ | ✓ | $0.673 \pm 0.053$ | $0.641 \pm 0.032$ | $0.669 \pm 0.059$ |
| ProSurv [23] | ✓ | ✓ | $0.655 \pm 0.037$ | $0.693 \pm 0.055$ | $0.673 \pm 0.034$ |
| M2Surv [24] | ✓ | ✓ | $0.668 \pm 0.038$ | $0.671 \pm 0.043$ | $0.692 \pm 0.042$ |
| MoME [9] | ✓ | ✓ | $0.679 \pm 0.041$ | $0.689 \pm 0.040$ | $0.701 \pm 0.038$ |
| **DualGGR-MoME(Ours)** | ✓ | ✓ | $\mathbf{0.706 \pm 0.030}$ | $\mathbf{0.724 \pm 0.052}$ | $\mathbf{0.719 \pm 0.033}$ |

## 3.3 Dual-Branch Gating Analysis

To validate the effectiveness of dual-branch gating, we conducted an ablation study comparing different configurations (Table 2): baseline without branches,

single-branch variants, and dual-branch. DualGGR-MoME with dual-branch gating significantly outperforms the baseline, achieving gains of 3.82% on BLCA, 5.69% on LUAD and 2.42% on UCEC. It also consistently surpasses single-branch variants, which show smaller improvements. This demonstrates that dual-branch design enables more effective routing decisions compared to conventional routing methods or single-branch alternatives.

**Table 2.** Dual-branch gating analysis. "Branch A" and "Branch B" represent two independent heads that generate routing signals.

| Model | Routing Signal Setup | | BLCA | LUAD | UCEC |
|---|---|---|---|---|---|
| | Branch A | Branch B | | | |
| DualGGR -MoME | ✗ | ✗ | $0.680 \pm 0.039$ | $0.685 \pm 0.032$ | $0.702 \pm 0.045$ |
| | ✓ | ✗ | $0.688 \pm 0.036$ | $0.693 \pm 0.041$ | $0.711 \pm 0.040$ |
| | ✗ | ✓ | $0.693 \pm 0.030$ | $0.711 \pm 0.036$ | $0.709 \pm 0.036$ |
| | ✓ | ✓ | **0.706 ± 0.028** | **0.724 ± 0.052** | **0.719 ± 0.033** |

### 3.4 Noise Robustness Analysis

As illustrated in Fig. 3, we evaluate the noise robustness of DualGGR-MoME against MoME on three TCGA datasets (BLCA, LUAD, and UCEC) across WSI and genomics modalities. Three types of noise are applied independently to each modality: Gaussian noise (levels: 0.1, 0.2, 0.3, 0.5), Dropout noise (levels: 0.1, 0.2, 0.3, 0.5), and Outlier noise (levels: 0.05, 0.1, 0.15, 0.2). Performance is measured by the mean absolute change in risk score $r = -\sum_{k=1}^{K} S_k$ relative to clean inputs. A smaller change indicates greater robustness to noise. DualGGR-MoME consistently maintains stable risk predictions across all noise types and levels, whereas MoME exhibits sharp risk score increases as noise intensifies. This robustness stems from our dual-stage design: DBG extracts routing signals from expert outputs, while GGR integrates them with global representations to compute noise-resilient fusion weights.

## 4 Conclusion

This paper proposes DualGGR-MoME, a Mixture-of-Multimodal-Experts framework with DBG and GGR for survival prediction. Addressing the noise vulnerability of existing routing methods, we decouple routing into two stages: DBG generates routing signals from expert outputs to reduce noise exposure, then GGR integrates these signals with global representations to compute robust weights. DualGGR-MoME employs four fusion experts with bidirectional alternating encoding for deep multimodal interaction. Experiments on three TCGA datasets demonstrate superior performance, with ablation studies confirming that dual-branch gating significantly improves accuracy and robustness. Noise robustness analysis validates the framework's stability and risk stratification capability.

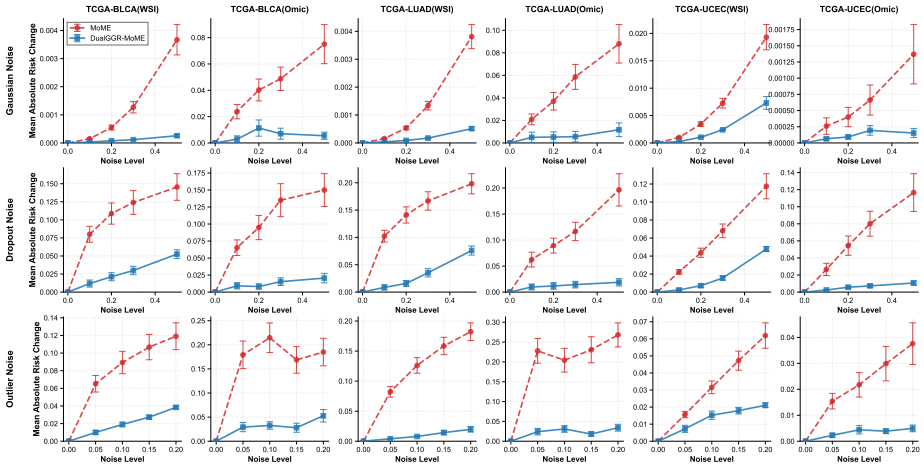


**Fig. 3.** Noise robustness analysis against feature perturbations. Mean absolute risk change (y-axis) is plotted against varying noise intensities (x-axis).

**Acknowledgments.** This work is supported by the National Natural Science Foundation of China (62302201), Major Science and Technology Special Projects of Yunnan Province (No. 202502AS080003, 202402AG050007), and Yunnan Xingdian Talent Support Plan Project.

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

## References

1. Mukesh Verma and Sudhir Srivastava. Epigenetics in cancer: implications for early detection and prevention. *The lancet oncology*, 3(12):755–763, 2002.
2. Richard J. Chen, Ming Y. Lu, Jingwen Wang, Drew F. K. Williamson, Scott J. Rodig, Neal I. Lindeman, and Faisal Mahmood. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4):757–770, 2022.
3. Ruiqing Li, Xingqi Wu, Ao Li, and Minghui Wang. Hfbsurv: hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction. *Bioinformatics*, 38(9):2587–2594, 2022.
4. Dominic Flack, Aakash Tripathi, Asim Waqas, Ghulam Rasool, and Dimah Dera. Robust multimodal fusion for survival prediction in cancer patients. *Cancer Informatics*, 24:11769351251376192, 2025.
5. Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4025, 2021.

6. Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer cell*, 40(8):865–878, 2022.
7. Kexin Ding, Mu Zhou, Dimitris N Metaxas, and Shaoting Zhang. Pathology-and-genomics multimodal transformer for survival outcome prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 622–631. Springer, 2023.
8. Guillaume Jaume, Anurag Vaidya, Richard Chen, Drew Williamson, Paul Liang, and Faisal Mahmood. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
9. Conghao Xiong, Hao Chen, Hao Zheng, Dong Wei, Yefeng Zheng, Joseph JY Sung, and Irwin King. Mome: Mixture of multimodal experts for cancer survival prediction. *arXiv preprint arXiv:2406.09696*, 2024.
10. Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. *arXiv preprint arXiv:2404.10237*, 2024.
11. Yufeng Jiang and Yiqing Shen. M4oE: A Foundation Model for Medical Multimodal Image Segmentation with Mixture of Experts . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15012. Springer Nature Switzerland, October 2024.
12. Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.
13. John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.
14. Diederik Kinga, Jimmy Ba Adam, et al. A method for stochastic optimization. In *International conference on learning representations (ICLR)*, volume 5. California;, 2015.
15. Yingxue Xu and Hao Chen. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21241–21251, October 2023.
16. Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*, 2018.
17. Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021.

18. Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *Advances in neural information processing systems*, 30, 2017.
19. Yilan Zhang, Yingxue Xu, Jianqi Chen, Fengying Xie, and Hao Chen. Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. In *The Twelfth International Conference on Learning Representations*, 2024.
20. Fengtao Zhou and Hao Chen. Cross-modal translation and alignment for survival analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21485–21494, 2023.
21. Andrew H Song, Richard J Chen, Guillaume Jaume, Anurag Jayant Vaidya, Alexander Baras, and Faisal Mahmood. Multimodal prototyping for cancer survival prediction. In *Forty-first International Conference on Machine Learning*, 2024.
22. Ruofan Zhang, Mengjie Fang, Shengyuan Liu, Zipei Wang, Jie Tian, and Di Dong. Tmse: Tri-modal survival estimation with context-aware tissue prototype and attention-entropy interaction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 640–650. Springer, 2025.
23. Fengchun Liu, Linghan Cai, Zhikang Wang, Zhiyuan Fan, Jin-gang Yu, Hao Chen, and Yongbing Zhang. Prototype-guided cross-modal knowledge enhancement for adaptive survival prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 522–532. Springer, 2025.
24. Mingcheng Qu, Guang Yang, Donglin Di, Yue Gao, Yang Song, and Lei Fan. M3surv: Fusing multi-slide and multi-omics for memory-augmented robust survival prediction. *Medical Image Analysis*, page 103846, 2025.