# Imbalance-Aware Distributional Alignment of Heterogeneous Modalities for HER2 Status Prediction

Ximeng Huang[1*], Wenjie Tang[2*], Yeli Wu[1], Jinhui Cai[3], Qingyu Liu[3], Weiwen Wu[1], Yuan Guo[2(✉)], and Jianjia Zhang[1(✉)]

[1] School of Biomedical Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China {huangxm77,wuyli29}@mail2.sysu.edu.cn; {zhangjj225,wuweiw7}@mail.sysu.edu.cn

[2] Department of Radiology, Guangzhou First People's Hospital, South China University of Technology, Guangzhou 510180, China {eywenjietang,eyguoyuan}@scut.edu.cn

[3] Department of Radiology, The Seventh Affiliated Hospital, Sun Yat-sen University, Shenzhen 518107, China liuqingyu@sysush.com; caijinhui123456@163.com

**Abstract.** Fusing whole-slide images (WSIs) and magnetic resonance imaging (MRI) provides a comprehensive perspective from both microscopic and macroscopic levels, and could potentially improve the accuracy of human epidermal growth factor receptor 2 (HER2) status prediction. However, directly applying existing fusion methods to such disparate modalities faces two major limitations. First, they often directly interact across these highly heterogeneous modalities, failing to adequately consider the huge domain gap between WSI and MRI. WSI and MRI differ fundamentally in imaging perspectives, information density, and dimensionality, making their correspondence inherently ambiguous. Second, they fail to effectively resolve modality imbalance caused by differences in information density through mere feature-level strategies. WSI contains substantially richer information and usually dominates the fusion process, causing models to inadequately learn from the non-dominant MRI modality. Consequently, both issues limit the model's ability to leverage multimodal information, sometimes even yielding performance inferior to unimodal baselines. To address these issues, we propose a novel Imbalance-Aware Distributional Alignment (IADA) framework for fusing WSI and MRI via distributional representations and dynamic gradient modulation. Specifically, a heterogeneous modality alignment strategy with distributional representations is introduced to mitigate the domain gap by indirectly characterizing and promoting semantic consistency across modalities. Furthermore, a dynamic gradient modulation module is incorporated to mitigate modality imbalance by adaptively suppressing the gradient contribution of the dominant modality, enabling both modalities to be fully leveraged in the fusion process. Extensive experiments on data from three medical centers demonstrate the superior generalization and effectiveness of the proposed method.

[*] These authors contributed equally to this work.

**Keywords:** Multimodal learning · Breast cancer · HER2 Status Prediction.

## 1 Introduction

As one of the most important biomarkers for breast cancer subtyping and treatment decision-making, human epidermal growth factor receptor 2 (HER2) status is closely associated with the selection of targeted therapies and patient prognosis [16,18]. Fusing whole-slide images (WSIs) and magnetic resonance imaging (MRI) can effectively exploit the complementary strengths of microscopic and macroscopic information, thereby improving HER2 status prediction. Pathological assessment captures cellular morphology, tissue architecture, and the tumor microenvironment at the microscopic level, serving as the gold standard for diagnosis and molecular subtyping, while MRI provides complete structural and functional information at the macroscopic level, offering a broader characterization of the overall tumor.

Many multimodal approaches have been proposed and they usually perform fusion via attention mechanisms, contrastive learning, or modality weighting strategies. For example, co-attention has been used to guide attention toward critical WSI regions [14,4], enhance token-level matching via Optimal Transport [24] or orthogonality-guided scoring [27], or establish denser interactions through cross-attention [28,12]. Contrastive learning-based methods typically align deterministic feature vectors from different modalities into a unified representation space [20] while weighting-based methods generally fuse modalities based on their contributions [7,21].

Nevertheless, directly applying existing fusion methods to disparate modalities of WSI and MRI faces two major limitations. First, existing methods often directly model the interactions between modalities, failing to adequately consider the huge domain gap between WSI and MRI. The imaging perspectives, information density, and dimensionality of WSI and MRI are fundamentally different, and their explicit correspondences are inherently ambiguous and challenging to model. Second, existing fusion methods may encounter the modality imbalance issue when applied to WSI and MRI. In comparison with MRI, WSI contains substantially richer information and thus may dominate the fusion process, causing the model to rely excessively on WSI while neglecting the learning of MRI, resulting in insufficient multimodal fusion and sometimes even inferior multimodal performance compared with unimodal baselines.

To address these limitations, we propose a novel Imbalance-Aware Distributional Alignment (IADA) framework for fusing WSI and MRI via distributional representations and dynamic gradient modulation. For the first limitation, we introduce a heterogeneous modality alignment strategy via distributional representations to bridge the domain gap between WSI and MRI. By modeling features as distributions rather than directly modeling the feature interactions, this strategy characterizes and promotes semantic consistency through the overlap between cross-modal latent distributions. Unlike existing methods, distribu-

tional alignment does not seek explicit local correspondences across modalities, and therefore can better align WSI and MRI with inherently ambiguous correspondences. This effectively mitigates the semantic domain gap and ultimately improves the prediction performance and generalization. In terms of the second limitation, a dynamic gradient modulation module is incorporated into the proposed method to mitigate modality imbalance by adaptively suppressing the gradient contribution of the dominant modality [19]. Specifically, the dynamic gradient modulation module uses unimodal prediction scores to quantify the contributions of individual modalities and dynamically balance the gradient updates of modality-specific parameters. By doing this, it effectively alleviates the modality imbalance issue and enables more effective exploration of the complementary information within WSI and MRI. The key contributions of this paper can be summarized as follows:

1. Considering the huge domain gap between WSI and MRI, we propose a novel distributional cross-modality alignment strategy that models the features of both modalities as distributional representations and promotes their semantic consistency.
2. To effectively mitigate the modality imbalance issue, we incorporate a dynamic gradient modulation module to dynamically adjust the contributions of WSI and MRI during model optimization, enabling more effective exploitation of fused information.
3. We conduct extensive experiments on a multi-center cohort collected from three independent medical centers, demonstrating the superior generalization and effectiveness of the proposed method.

## 2 Method

### 2.1 Overall Architecture

The proposed Imbalance-Aware Distributional Alignment (IADA) framework is illustrated in Fig. 1. Given paired WSI and MRI data, modality-specific features are first extracted through dedicated encoders to obtain compact representations. To reduce the domain gap between heterogeneous modalities, each representation is projected into a Gaussian distribution parameterized by its mean vector and standard deviation vector, forming distributional embeddings within a shared latent space. A symmetrized KL divergence loss is employed to align the cross-modal distributions, thereby enhancing semantic consistency across modalities. The mean vector of each modality is further refined via a feature gating module to produce discriminative representations, which are subsequently concatenated and fed into a classifier for final prediction. To alleviate modality imbalance, the unimodal prediction scores of each modality are fed into a Dynamic Gradient Modulation (DGM) module, which adaptively rescales modality-specific gradients based on their batch-level contribution ratio to promote balanced optimization. The entire framework is optimized end-to-end under

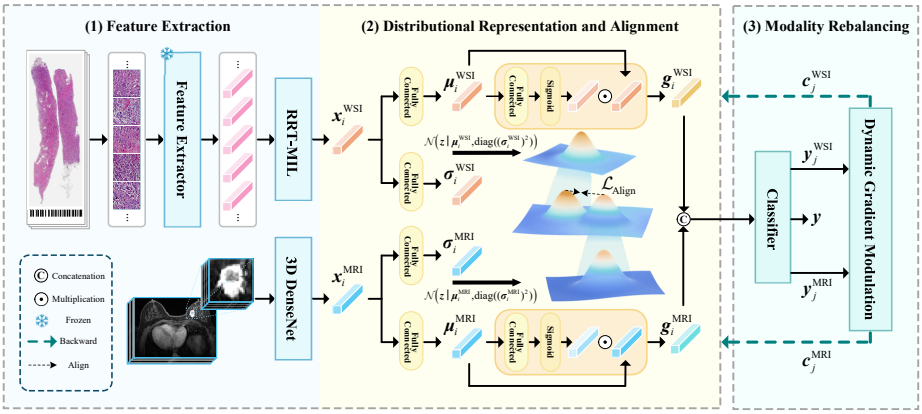


**Fig. 1.** Overview of the proposed framework. (1) **Feature Extraction** encodes WSI and MRI via RRT-MIL and 3D DenseNet, respectively. (2) **Distributional Representation and Alignment** maps modality-specific features into Gaussian distributions in a shared latent space, with a symmetrized KL divergence loss promoting cross-modal alignment. (3) **Modality Rebalancing** employs a DGM module to adaptively scale gradient updates and mitigate modality imbalance during training.

a unified objective function:

$$\mathcal{L} = (1-\alpha)\,\mathcal{L}_{\text{CLS}} + \alpha\,\mathcal{L}_{\text{Dist}}, \tag{1}$$

where $\mathcal{L}_{\text{CLS}}$ denotes the classification loss, $\mathcal{L}_{\text{Dist}} = \mathcal{L}_{\text{Align}} + \mathcal{L}_{\text{Reg}}$ denotes the distributional loss with $\mathcal{L}_{\text{Align}}$ being the symmetrized KL divergence that promotes distributional overlap between the two modalities and $\mathcal{L}_{\text{Reg}}$ being the KL regularization that anchors the learned distributions to a standard Gaussian prior, and $\alpha$ is a hyperparameter balancing the two objectives.

### 2.2 Feature Extraction

We extract features from WSI and MRI. For WSI, we first segment the tissue region and crop a series of non-overlapping $512 \times 512$ patches at a magnification of $20\times$. Instance-level features are extracted from patches with pre-trained ResNet50 [8] and stacked as a bag-level embedding $\boldsymbol{P}_i^{\text{WSI}} \in \mathbb{R}^{N \times d_p}$, where $N$ is the number of patches and $d_p$ is the patch feature dimension. The bag-level embedding is then fed into the RRT [23] encoder to re-embed instance features by modeling spatial dependencies, and aggregated into a slide-level feature representation $\boldsymbol{x}_i^{\text{WSI}} \in \mathbb{R}^{1 \times d}$. For MRI, we crop tumor-centered 3D volumes of size $64 \times 64 \times 64$ from the region of interest (ROI) delineated by expert radiologists and feed them into a 3D DenseNet-121 [10] to extract the MRI feature representation $\boldsymbol{x}_i^{\text{MRI}} \in \mathbb{R}^{1 \times d}$, where $d$ is the shared feature dimension for both modalities.

### 2.3 Distributional Representation and Alignment

To bridge the domain gap between heterogeneous modalities, we model each modality's features as distributional representations and promote cross-modal semantic consistency through distributional overlap. Modeling features as distributions can better capture their implicit alignment rather than directly modeling their explicit correspondence, which naturally accommodates the inherently ambiguous correspondence between WSI and MRI.

We denote the modality index as $m \in \{\text{WSI}, \text{MRI}\}$. For the $i$-th sample, the feature $\boldsymbol{x}_i^m \in \mathbb{R}^{1\times d}$ extracted from modality $m$ is mapped to a mean vector $\boldsymbol{\mu}_i^m \in \mathbb{R}^{1\times d_z}$ and a standard deviation vector $\boldsymbol{\sigma}_i^m \in \mathbb{R}^{1\times d_z}$ as:

$$\boldsymbol{\mu}_i^m = f_\mu^m(\boldsymbol{x}_i^m), \quad \boldsymbol{\sigma}_i^m = f_\sigma^m(\boldsymbol{x}_i^m), \tag{2}$$

where $f_\mu^m$ and $f_\sigma^m$ are modality-specific fully connected mappings with independent learnable parameters. Rather than producing a single deterministic embedding, we represent each sample as a Gaussian distribution in the latent space:

$$q_i^m(\boldsymbol{z}) = \mathcal{N}\big(\boldsymbol{z} \mid \boldsymbol{\mu}_i^m, \operatorname{diag}((\boldsymbol{\sigma}_i^m)^2)\big), \tag{3}$$

where $q_i^m(\boldsymbol{z})$ denotes the latent distribution and $\boldsymbol{z}$ is the latent random variable. To promote semantic consistency between WSI and MRI in the latent space, the distributional loss $\mathcal{L}_{\text{Dist}}$ is defined as:

$$\begin{aligned}\mathcal{L}_{\text{Dist}} = \frac{1}{2}\left(\text{KL}(q_i^{\text{MRI}} \parallel q_i^{\text{WSI}}) + \text{KL}(q_i^{\text{WSI}} \parallel q_i^{\text{MRI}})\right) \\ + \frac{\lambda}{2}\sum_m \text{KL}(q_i^m \parallel \mathcal{N}(\mathbf{0}, \mathbf{I})),\end{aligned} \tag{4}$$

where the first term $\mathcal{L}_{\text{Align}}$ is a symmetrized KL divergence that encourages the latent distributions of the two modalities to overlap, and the second term $\mathcal{L}_{\text{Reg}}$ anchors the learned distributions to a standard Gaussian prior to prevent over-concentration, where $\mathbf{I}$ denotes the identity matrix and $\lambda$ controls the regularization strength. Together, they bridge the domain gap between heterogeneous modalities by promoting cross-modal semantic consistency while maintaining a well-structured latent space.

To further enhance the discriminability of the latent representation, $\boldsymbol{\mu}_i^m$ is fed into a fully connected layer followed by a sigmoid activation to generate a gate vector, which is then multiplied element-wise with $\boldsymbol{\mu}_i^m$ to obtain the gated representation $\boldsymbol{g}_i^m \in \mathbb{R}^{1\times d_z}$. The gated representations $\boldsymbol{g}_i^{\text{WSI}}$ and $\boldsymbol{g}_i^{\text{MRI}}$ are subsequently concatenated and fed into a shared classifier to produce the fused prediction score $y$.

### 2.4 Dynamic Gradient Modulation

To mitigate the modality imbalance caused by the significant disparities in information density between WSI and MRI, we incorporate a dynamic gradient modulation module that quantifies the contribution of each modality via unimodal

prediction scores and dynamically balances their respective gradient updates. Specifically, in addition to the fused prediction score $y$, the classifier is also applied separately to $\boldsymbol{g}_i^{\mathrm{WSI}}$ and $\boldsymbol{g}_i^{\mathrm{MRI}}$ to produce modality-specific prediction scores $y_i^{\mathrm{WSI}}$ and $y_i^{\mathrm{MRI}}$. The contribution discrepancy between WSI and MRI can then be quantified as the ratio:

$$u_j^{\mathrm{WSI}} = \frac{\sum_{i \in B_j} y_i^{\mathrm{WSI}}}{\sum_{i \in B_j} y_i^{\mathrm{MRI}}}, \tag{5}$$

where $B_j$ denotes the $j$-th training batch, $u_j^{\mathrm{WSI}} > 1$ indicates that WSI contributes more than MRI to the current prediction, reflecting a state of modality imbalance. Symmetrically, $u_j^{\mathrm{MRI}}$ is defined as the reciprocal of $u_j^{\mathrm{WSI}}$. Based on the contribution ratio, a gradient modulation coefficient is defined as:

$$c_j^m = \begin{cases} 1 - \tanh(\beta \cdot u_j^m), & u_j^m > 1 \\ 1, & \text{otherwise}, \end{cases} \tag{6}$$

where $\beta$ is a hyperparameter controlling the degree of modulation. The resulting $c_j^m \in (0, 1]$ is integrated into the parameter update rule. The modality-specific parameters $\boldsymbol{\theta}^m$ are updated as:

$$\boldsymbol{\theta}^m \leftarrow \boldsymbol{\theta}^m - \eta \cdot c_j^m \cdot \nabla_{\boldsymbol{\theta}^m} \mathcal{L}, \tag{7}$$

where $\eta$ is the learning rate and $\nabla_{\boldsymbol{\theta}^m} \mathcal{L}$ denotes the gradient with respect to $\boldsymbol{\theta}^m$. When one modality dominates the prediction, its gradient update is reduced to limit its influence during optimization, while keeping the non-dominant modality updated at full scale. Since $\eta$ is small and $c_j^m$ is constrained within $(0, 1]$, DGM introduces only mild modulation at each iteration, resulting in gradual rather than abrupt optimization changes. This dynamic and stable adjustment promotes a more balanced use of complementary WSI and MRI information.

## 3 Experiments

### 3.1 Dataset

To evaluate the robustness and generalizability of our proposed method, we collected data from 866 breast cancer patients across three independent medical centers between 2021 and 2024. For each patient, the multimodal dataset comprises a WSI derived from core needle biopsy, the third enhancement phase of DCE-MRI, and the corresponding clinically confirmed HER2 status. Cohort A consists of 364 patients, including 93 HER2-positive and 271 HER2-negative cases; Cohort B consists of 385 patients, including 96 HER2-positive and 289 HER2-negative cases; and Cohort C consists of 117 patients, including 33 HER2-positive and 84 HER2-negative cases. For internal validation, Cohorts A and B were pooled and evaluated using five-fold cross-validation. For external evaluation, Cohorts A and B were used for training, and Cohort C served as an independent external test set. The model performance was evaluated using the area under the curve (AUC), accuracy (ACC), F1-score (F1), sensitivity (SEN), and specificity (SPE) on both the internal validation and external test.

**Table 1.** The classification performance (%) of different methods on internal validation and external test.

| Modality | Model | Internal Validation | | | | | External Test | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUC | ACC | F1 | SEN | SPE | AUC | ACC | F1 | SEN | SPE |
| WSI | TransMIL [22] | 66.9 | 76.4 | 52.1 | 51.4 | 84.7 | 73.2 | 70.4 | 59.5 | 80.6 | 66.7 |
| | CLAM [17] | 76.7 | 78.4 | 60.0 | 64.9 | 82.9 | 78.1 | 67.8 | 60.2 | 90.3 | 59.5 |
| | AMIL [11] | 76.0 | 72.3 | 59.4 | 81.1 | 69.3 | 79.7 | 73.0 | 57.8 | 87.1 | 67.9 |
| | RRT-MIL [23] | 78.4 | 72.9 | 58.3 | 75.7 | 72.1 | 80.4 | 70.4 | 60.5 | 83.9 | 65.5 |
| MRI | ConvNext [15] | 63.6 | 58.1 | 47.5 | 52.3 | 75.7 | 59.8 | 65.8 | 47.4 | 56.3 | 69.4 |
| | nnMamba [6] | 65.5 | 52.7 | 43.9 | 75.0 | 45.5 | 61.3 | 56.0 | 41.3 | 62.1 | 54.0 |
| | SENet [9] | 64.0 | 64.9 | 43.5 | 54.1 | 68.5 | 64.1 | 64.3 | 48.1 | 61.3 | 65.5 |
| | ResNet50 [8] | 65.2 | 61.5 | 45.7 | 64.9 | 60.4 | 62.7 | 67.5 | 47.2 | 53.1 | 72.9 |
| | DenseNet [10] | 67.0 | 75.0 | 46.4 | 43.2 | 85.6 | 67.7 | 67.0 | 52.5 | 67.7 | 66.7 |
| Fusion | MCAT [4] | 74.4 | 78.4 | 59.0 | 62.2 | 83.8 | 78.2 | 73.0 | 60.8 | 77.4 | 71.4 |
| | MoPE [13] | 77.9 | 70.9 | 59.0 | 83.8 | 66.7 | 78.7 | 70.4 | 58.5 | 77.4 | 67.9 |
| | Pathomic Fusion [3] | 78.1 | 75.0 | 60.2 | 75.7 | 75.7 | 78.5 | 67.0 | 57.8 | 83.9 | 60.7 |
| | CMTA [28] | 78.3 | 70.3 | 58.5 | 83.8 | 65.8 | 79.4 | 70.4 | 59.5 | 80.7 | 66.7 |
| | Weight | 78.8 | 75.0 | 61.9 | 81.1 | 73.0 | 79.4 | 72.2 | 59.0 | 74.2 | 71.4 |
| | Concat [2] | 78.5 | 76.4 | 61.5 | 75.7 | 76.6 | 79.9 | 73.0 | 60.0 | 74.2 | 72.6 |
| | GMU [1] | 78.0 | 75.0 | 61.1 | 78.4 | 73.9 | 80.0 | 70.0 | 60.0 | 83.8 | 64.3 |
| | EAU [5] | 78.4 | 75.7 | 59.1 | 70.3 | 77.5 | 80.2 | 74.0 | 60.3 | 75.7 | 74.8 |
| | Multimodal Dynamics [7] | 78.2 | 74.3 | 60.4 | 78.4 | 73.0 | 80.3 | 71.3 | 61.2 | 83.9 | 66.7 |
| | OGM [19] | 79.1 | 81.1 | 64.1 | 67.6 | 85.6 | 80.9 | 73.0 | 61.7 | 80.6 | 70.2 |
| | AMSS [25] | 79.5 | 76.4 | 63.2 | 81.1 | 74.8 | 83.6 | 69.6 | 62.4 | 93.5 | 60.7 |
| | LFM [26] | 79.5 | 81.1 | 65.0 | 70.3 | 84.7 | 83.3 | 75.7 | 64.1 | 80.6 | 73.8 |
| | **IADA (proposed)** | **81.7** | **83.8** | **68.4** | 70.3 | 88.3 | **86.1** | **77.4** | **64.9** | 82.8 | 75.6 |

### 3.2 Implementation Details

All experiments were conducted on an NVIDIA RTX A6000 GPU (48 GB). We employed the Adam optimizer with an initial learning rate $\eta$ of $1 \times 10^{-4}$. The batch size was set to 2 due to the high memory cost of WSIs. The classification branch was optimized using focal loss. The hyperparameters $\alpha$, $\beta$, and $\lambda$ were empirically set to 0.2, 0.1, and $2 \times 10^{-2}$, respectively. The maximum number of training epochs was set to 100. The WSI encoder, MRI encoder, and fusion module contain 2.828M, 11.507M, and 1.054M parameters, respectively, for a total of 15.389M parameters. The implementation is available at https://github.com/embers819/IADA.

### 3.3 Experimental Results

**Comparison with State-of-the-Art Methods** Table 1 presents a comprehensive comparison between the proposed method and state-of-the-art single-modality and multimodal fusion approaches. The proposed IADA achieves AUCs of 81.7% and 86.1% on the internal validation and external test. IADA outperforms the second-best AUC by 2.2% and 2.5% on the internal validation and external test, respectively. Meanwhile, IADA significantly outperforms the

**Table 2.** Ablation study (%) on the contribution of each model component.

| Baseline | Alignment | | DGM | Internal Validation | | | | | External Test | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | point | distribution | | AUC | ACC | F1 | SEN | SPE | AUC | ACC | F1 | SEN | SPE |
| ✓ | | | | 78.2 | 81.8 | 63.0 | 62.2 | 88.3 | 79.9 | 66.1 | 58.9 | 90.3 | 57.1 |
| ✓ | ✓ | | | 79.4 | 78.4 | 62.8 | 73.0 | 80.2 | 79.6 | 72.2 | 63.6 | 90.3 | 65.5 |
| ✓ | | ✓ | | 80.5 | 75.0 | 61.1 | 78.4 | 73.9 | 83.0 | 74.8 | **65.9** | 90.3 | 69.0 |
| ✓ | | ✓ | ✓ | **81.7** | **83.8** | **68.4** | 70.3 | 88.3 | **86.1** | **77.4** | 64.9 | 82.8 | 75.6 |

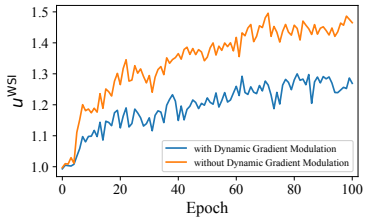


**Fig. 2.** Dynamics of the modality contribution ratio between WSI and MRI.

strongest unimodal WSI and MRI baselines. Specifically, it improves the AUC over the best WSI method RRT-MIL and the best MRI method DenseNet by 3.3% ($p < 0.001$, fold-level paired $t$-test) and 14.7% ($p < 0.001$, fold-level paired $t$-test) on internal validation, and by 5.7% ($p = 0.031$, DeLong test) and 18.4% ($p < 0.001$, DeLong test) on the external test, respectively, highlighting the advantage of WSI-MRI fusion. Notably, AUC, ACC, F1, and SPE all reach the highest values in both settings.

**Ablation Study** Table 2 presents the ablation study validating the contribution of distributional alignment and DGM. Compared with point-based alignment, distributional alignment improves AUC by 1.1% and 3.4% on the internal validation and external test, respectively. These improvements suggest that modeling features as distributions can better capture implicit cross-modal alignment than directly enforcing explicit correspondence. Incorporating DGM further improves AUC by 1.2% and 3.1% on the internal validation and external test, respectively, highlighting its role in mitigating modality imbalance. Fig. 2 further shows the trend of the modality contribution ratio $u^{\mathrm{WSI}}$ during training with and without DGM. Without DGM, $u^{\mathrm{WSI}}$ increases more rapidly and reaches a higher magnitude. With DGM, this growth is notably moderated, indicating that DGM alleviates the dominance of the WSI modality and facilitates a more balanced optimization process. Moreover, the contribution-ratio curve with DGM exhibits a smooth trend, with small fluctuations comparable to the setting without DGM, demonstrating the stable behavior of the modulation strategy.

## 4 Conclusion

In sum, this paper proposes a novel multimodal fusion framework, Imbalance-Aware Distributional Alignment (IADA), that jointly leverages WSI and MRI for accurate HER2 status prediction. The introduced distributional alignment strategy and dynamic gradient modulation module effectively address the domain gap between heterogeneous modalities and the modality imbalance problem, respectively. Extensive experiments on a multi-center cohort of 866 patients demonstrate the superior performance and generalization of the proposed method over state-of-the-art approaches.

**Acknowledgments.** This work was supported in part by Shenzhen Medical Research Fund under Grant B2402030; in part by the National Natural Science Foundation of China under Grant 62471501, Grant 92570107, Grant 62471502; in part by Shenzhen Science and Technology Program under Grant JCYJ20240813151026034; in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2026B1515020019.

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

## References

1. Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A.: Gated multimodal units for information fusion. arXiv preprint arXiv:1702.01992 (2017)
2. Boehm, K.M., Aherne, E.A., Ellenson, L., Nikolovski, I., Alghamdi, M., Vázquez-García, I., Zamarin, D., Long Roche, K., Liu, Y., Patel, D., et al.: Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. Nature cancer **3**(6), 723–733 (2022)
3. Chen, R.J., Lu, M.Y., Wang, J., Williamson, D.F., Rodig, S.J., Lindeman, N.I., Mahmood, F.: 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 (2020)
4. 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: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4025 (2021)
5. Gao, Z., Jiang, X., Xu, X., Shen, F., Li, Y., Shen, H.T.: Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26876–26885 (2024)
6. Gong, H., Kang, L., Wang, Y., Wang, Y., Wan, X., Wu, X., Li, H.: nnmamba: 3d biomedical image segmentation, classification and landmark detection with state space model. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2025)
7. Han, Z., Yang, F., Huang, J., Zhang, C., Yao, J.: Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20707–20717 (2022)

8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
10. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
11. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
12. Jaume, G., Vaidya, A., Chen, R.J., Williamson, D.F., Liang, P.P., Mahmood, F.: Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11579–11590 (2024)
13. Jiang, R., Liu, L., Chen, C.: Mope: Mixture of prompt experts for parameter-efficient and scalable multimodal fusion. arXiv preprint arXiv:2403.10568 (2024)
14. Li, Z., Jiang, Y., Lu, M., Li, R., Xia, Y.: Survival prediction via hierarchical multimodal co-attention transformer: A computational histology-radiology solution. IEEE Transactions on Medical Imaging **42**(9), 2678–2689 (2023)
15. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
16. Loibl, S., Gianni, L.: Her2-positive breast cancer. The Lancet **389**(10087), 2415–2429 (2017)
17. 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)
18. Oh, D.Y., Bang, Y.J.: Her2-targeted therapies—a role beyond breast cancer. Nature reviews Clinical oncology **17**(1), 33–48 (2020)
19. Peng, X., Wei, Y., Deng, A., Wang, D., Hu, D.: Balanced multimodal learning via on-the-fly gradient modulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8238–8247 (2022)
20. Qu, L., Huang, D., Zhang, S., Wang, X.: Multi-modal data binding for survival analysis modeling with incomplete data and annotations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 501–510. Springer (2024)
21. Qu, M., Yang, G., Di, D., Su, T., Gao, Y., Song, Y., Fan, L.: Multimodal cancer survival analysis via hypergraph learning with cross-modality rebalance. arXiv preprint arXiv:2505.11997 (2025)
22. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
23. Tang, W., Zhou, F., Huang, S., Zhu, X., Zhang, Y., Liu, B.: Feature re-embedding: Towards foundation model-level performance in computational pathology. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11343–11352 (2024)
24. Xu, Y., Chen, H.: 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. pp. 21241–21251 (2023)

25. Yang, Y., Pan, H., Jiang, Q.Y., Xu, Y., Tang, J.: Learning to rebalance multimodal optimization by adaptively masking subnetworks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
26. Yang, Y., Wan, F., Jiang, Q.Y., Xu, Y.: Facilitating multimodal classification via dynamically learning modality gap. Advances in Neural Information Processing Systems **37**, 62108–62122 (2024)
27. Zhang, S., Du, S., Sun, C., Li, B., Shao, L., Zhang, L., Wang, K., Liu, Z., Tian, J.: M2fusion: multi-time multimodal fusion for prediction of pathological complete response in breast cancer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 458–468. Springer (2024)
28. Zhou, F., Chen, H.: Cross-modal translation and alignment for survival analysis. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 21485–21494 (2023)