Focus on Focus: Focus-oriented Representation Learning and Multi-view Cross-modal Alignment for Glioma Grading

Li Pan[1*], Yupei Zhang[1*], Qiushi Yang[2], Tan Li[3], Xiaohan Xing[4], Maximus C. F. Yeung[1], Zhen Chen[5†]

[1]*The University of Hong Kong*   [2]*City University of Hong Kong*   [3]*The Hang Seng University of Hong Kong*
[4]*Stanford University*   [5]*Centre for Artificial Intelligence and Robotics (CAIR), HKISI-CAS*

Abstract—Recently, multimodal deep learning, which integrates histopathology slides and molecular biomarkers, has achieved a promising performance in glioma grading. Despite great progress, due to the intra-modality complexity and inter-modality heterogeneity, existing studies suffer from inadequate histopathology representation learning and inefficient molecular-pathology knowledge alignment. These two issues hinder existing methods to precisely interpret diagnostic molecular-pathology features, thereby limiting their grading performance. Moreover, the real-world applicability of existing multimodal approaches is significantly restricted as molecular biomarkers are not always available during clinical deployment. To address these problems, we introduce a novel Focus on Focus (FoF) framework with paired pathology-genomic training and applicable pathology-only inference, enhancing molecular-pathology representation effectively. Specifically, we propose a Focus-oriented Representation Learning (FRL) module to encourage the model to identify regions positively or negatively related to glioma grading and guide it to focus on the diagnostic areas with a consistency constraint. To effectively link the molecular biomarkers to morphological features, we propose a Multi-view Cross-modal Alignment (MCA) module that projects histopathology representations into molecular subspaces, aligning morphological features with corresponding molecular biomarker status by supervised contrastive learning. Experiments on the TCGA GBM-LGG dataset demonstrate that our FoF framework significantly improves the glioma grading. Remarkably, our FoF achieves superior performance using only histopathology slides compared to existing multimodal methods. The source code is available at https://github.com/peterlipan/FoF.

Index Terms—Glioma grading, Multimodal learning, Missing modality, Representation learning

I. INTRODUCTION

As the most common type of brain tumors, gliomas are classified into Grade II to IV by the World Health Organization, correlating with varied prognoses and intervention approaches [1]. The gold standard for grading gliomas is the observation of representative histopathology features in biopsies [2]. However, histopathology slides present a complex milieu of cells, necrosis, and microenvironments, which complicates the localization of tumor foci, necessitating the expertise of senior pathologists [3]. The recent advances in computer-assisted cancer grading reveal promising performance in identifying glioma grades from histopathology slides [4], [5].

Recent studies [6], [7] in cancer grade classification benefit from multimodal approaches that combine histopathology features with molecular biomarkers from tissue biopsies, offering a comprehensive and accurate tumor analysis. Existing multimodal methods [6], [8] heavily rely on paired pathology-genomic data for inference when deployed. However, obtaining molecular biomarkers requires staining and/or sequencing, which is not routinely available in clinical practice due to high costs and technical challenges, thereby limiting their practical applications [9]. To address this limitation, cutting-edge multimodal techniques [10], [11] propose achieving unimodal inference by distilling knowledge from a pathology-genomic teacher, expanding the applicability of multimodal grading in clinical settings. Nevertheless, these methods are inherently constrained by the performance of multimodal teachers, highlighting an emergent need for a novel approach that can effectively learn from multimodal data while enabling unimodal inference.

The effectiveness of existing multimodal methods is limited by the difficulties in accurately identifying diagnostic molecular-pathology features, which can be attributed to two major issues. The first issue is the *insufficient histopathology representation learning*, attributed to a lack of focus on the most representative regions associated with gliomas. Existing methods [4] attempt to correlate the complex images with diagnostic grades by uniformly processing the whole histopathology slides. Nonetheless, without an explicit regularization, these approaches [5], [12] are prone to overfitting the minor textual details rather than high-level diagnostic patterns. Existing studies [6], [13] attempt to tackle this issue by augmenting the histopathology inputs with handcrafted morphological features but lack the generalization ability. To this end, there is an urgent need for a novel method to boost histopathology representation learning by directing the model focus toward areas most relevant to glioma grading.

Another challenge lies in the *inefficient molecular-pathology knowledge alignment*. Contemporary multimodal methods tend to fuse the pathology-genomic information in the feature space through simple combinations. However, due to the notable heterogeneity between the two modalities, these methods struggle to align the morphological features of tumor niches with molecular biomarkers, such as Intradialytic hypotension

---

L. Pan and Y. Zhang contribute equally to this work.

(IDH) mutation status and 1p/19q codeletion presence, which are crucial for precise glioma grading [1], [14]. Thus, there is a high demand for novel strategies that can effectively correlate genomic biomarkers with histopathology abnormalities to jointly identify the molecular-pathology features of gliomas toward precision grading.

To tackle these challenges, we propose the Focus on Focus (FoF) framework that enhances the identification of diagnostic molecular-pathology features by boosting the histopathology representations with Focus-oriented Representation Learning (FRL), and effectively integrating the molecular biomarkers with Multi-view Cross-modal Alignment (MCA). Specifically, to improve the representations of histopathology slides, we propose FRL to identify the regions positively and negatively correlated to cancer grading and encourage the model to focus on diagnostic regions with a consistency constraint. To correlate the genomic biomarkers with histopathology properties, we propose MCA to cluster the multi-view visual representations that share compatible molecular contexts while distancing those in discrepancies through supervised contrastive learning. By precisely locating the tumor niches and linking them with molecular biomarkers, FoF enriches the visual representations of histopathology slides and facilitates accurate glioma grading only using images. Experimental results prove that our FoF framework outperforms state-of-the-art histopathology and multimodal methods on the TCGA GBM-LGG dataset.

The contributions of this work are summarized as follows:

- To achieve accurate grading of gliomas, we propose a novel FoF framework that directs the focus of models to diagnostic regions. To the best of our knowledge, this is the first work to adjust the model focus on areas of both histopathology and molecular-pathology significance.
- We devise the FRL module to identify regions that positively or negatively contribute to grading, encouraging the model to focus on important diagnostic niches through a consistency constraint on histopathology features.
- In response to the latest WHO classification of gliomas, we propose the MCA module to align the histopathology features directly with molecular biomarkers and balance the impact of each biomarker through a novel multi-view contrastive learning.
- Extensive experiments on the TCGA GBM-LGG dataset prove the effectiveness, and our FoF outperforms multimodal grading state-of-the-arts using only histopathology slides, demonstrating great clinical significance.

II. RELATED WORK

A. Glioma Grading

Histopathology grading Glioma is a significant type of malignant tumor that occurs in the glial cells of the brain or spinal cord and can be classified into Grades II to IV, each correlating with different prognoses [2]. Histologic properties observed under a microscope from biopsy samples are the gold standard for tumor assessments in clinical practice, *e.g.*, the presence of

necrosis and/or microvascular proliferation leads to the diagnosis of high-grade glioma (HGG) [15]. To reduce the workload of pathologists and promote consistent diagnosis, researchers have utilized traditional machine learning and deep learning models to classify morphological features [10]. For example, Ker *et al.* employed a pre-trained convolutional neural network to classify histopathology slides into normal, low-grade glioma (LGG), or HGG categories [16]. Rathore *et al.* extracted conventional features (*e.g.*, intensity) and textural features (*e.g.*, gray-level co-occurrence matrix) from specimens, and used a support vector machine for classification [17]. However, handcrafted features only capture detailed textural information, and deep learning models are prone to overfitting it due to the high complexity of histopathology slides. As a result, these approaches are inefficient at capturing the representative patterns of gliomas, lacking both generalization ability and interpretability.

Multimodal grading According to the latest WHO classification of tumors of the central nerve system [1], integrating molecular biomarkers with histopathology slides provides a more comprehensive and precise analysis of gliomas. Inspired by this medical insight, researchers attempt to project the histopathology and molecular data into a uniform latent space, fusing the features of the two modalities through various combinations, such as concatenation [7], [18], Kronecker Product [6], [19], and cross attention [8]. Yet, molecular biomarkers, such as the status of IDH mutation and 1p/19q codeletion, contain precise information that directly correlates with the grading of gliomas [14]. As a result of this heterogeneity between the two modalities, existing methods that rely on uniform projection and late fusion are insufficient at aligning cross-modal information through simple combinations and overlook the correlations among different molecular biomarkers. In contrast, our FoF framework encourages the model to adaptively focus on representative regions and directly align these regions with biomarkers, resulting in more robust representation learning and improved interpretability.

B. Multimodal Glioma Grading with Missing Modality

Although molecular biomarkers are essential for the clinical assessment of cancers, their acquisition necessitates immuno-histochemistry (IHC) staining and DNA/RNA sequencing. These processes are time-consuming and costly, thus resulting in limited accessibility, particularly in underrepresented areas [20]. Fusion-based multimodal methods [6] necessitate the presence of paired histopathology slides and molecular biomarkers, further restricting their practical application in real-world clinical settings. To alleviate this challenge, DDM-net [21] proposed to reconstruct the features of unavailable modality from the available one with the transformation function learned from paired data. Most recent studies have implemented distillation-based methods to enhance the image-based model by distilling knowledge from a pathology-genomic teacher [10]. For instance, Xing *et al.* [19] devised a novel distillation framework that effectively transfers knowledge from a multimodal teacher to a uni-modal student, achieving
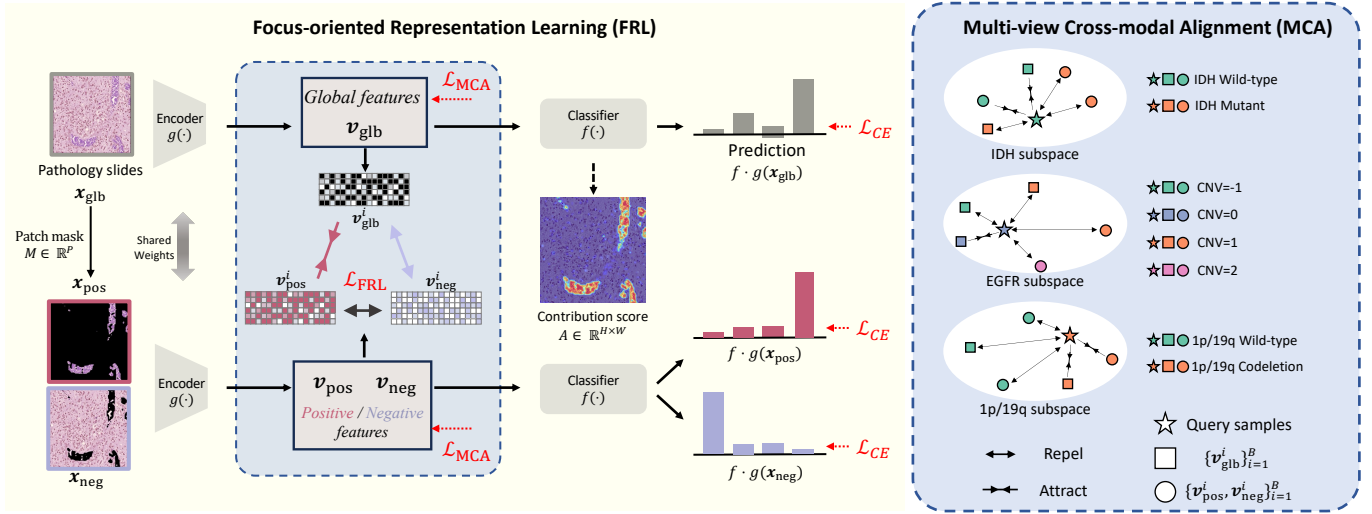
Fig. 1. The FoF framework. FRL identifies the positive x_{pos} and negative regions x_{neg}, encouraging the model to focus on the most important regions with a consistency constrict. The MCA module correlates the histopathology representations $\{v_{\text{glb}}, v_{\text{pos}}, v_{\text{neg}}\}$ with molecular biomarkers.

performance comparable to multimodal grading using only histopathology slides. Despite efforts to maximize the efficiency of transferring knowledge from multiple modalities to a single modality, these approaches [10], [19], [21] are naturally limited by the patterns learned from pathology-genomic pairs, leading to the aforementioned problem of insufficient cross-modal alignment. Unlike these methods, our FoF enhances the representation learning on histopathology slides directly with the guidance of individual molecular biomarkers, efficiently correlating the representative morphological patterns with molecular properties.

III. METHODOLOGY

A. Overview

As illustrated in Fig. 1, our FoF framework improves the glioma grading by highlighting the diagnostic molecular-pathology features. To enhance the representation learning on histopathology slides, we introduce FRL to identify the positive x_{pos} and negative regions x_{neg} from the input images x_{glb} with the pixel-wise contribution score $\mathbf{A} \in \mathbb{R}^{H \times W}$, encouraging the model $f \cdot g$ to focus on the most important areas. In MCA, we project the multi-view features $\{v_{\text{glb}}^i, v_{\text{pos}}^i, v_{\text{neg}}^i\}_{i=1}^{B}$ into individual molecular subspaces with distinct projectors $h^n(\cdot)$ and employ biomarkers as labels. During inference, the model predicts the glioma grades $f \cdot g(x_{\text{glb}})$ with only histopathology slides.

B. Focus-oriented Representation Learning

Existing representation learning on histopathology slides is inadequate, which can be attributed to the lack of focus on the diagnostic morphological patterns [13]. To boost histopathology representation learning, we propose the Focus-oriented Representation Learning module to encourage the focus of the model on areas most relevant to gliomas and enhance these representations with a consistency constraint. Specifically,

FRL quantifies the contribution score of each pixel towards accurate classification $\mathbf{A} \in \mathbb{R}^{H \times W}$ and identifies the areas that are positively and negatively related to grading using a threshold. This module separately feeds the positive x_{pos} and negative x_{neg} regions into the ViT encoder g and enhances the visual representations with a consistency constraint on the positive v_{pos} and global features v_{glb}.

Technically, locating the focus of models has been formulated as identifying the input pixels that contribute most significantly to the output [22]. To identify the diagnostic regions within a specific slide x, FRL initially processes the whole image through the model to obtain the predictions $f \cdot g(x)$. It then aggregates the gradient of the prediction for the ground-truth class s across each feature map layer v^k, thereby estimating the contribution score of each pixel towards accurate grading as follows:

$$
\begin{aligned}
\alpha_k &= \sum_i \sum_j \frac{\partial s}{\partial v_{ij}^k}, \\
\mathbf{A} &= \Upsilon(\text{ReLU}(\sum_k \alpha_k v^k), [H, W]),
\end{aligned}
\tag{1}
$$

where i and j represent the location indices within the feature map v, α_k is the weight associated with the k-th feature map v^k, and $\Upsilon(\cdot, [H, W])$ refers to the reshape and resample function that upscales the feature map to match the input image size. Consequently, the matrix $\mathbf{A}$ indicates the contribution of each pixel toward accurately classifying the input image to the target class. Following this step, we divide the pixel-wise contribution score into $P = \lfloor \frac{H}{p} \rfloor \times \lfloor \frac{W}{p} \rfloor$ patches, where p is the patch size. The FRL calculates the average patch-wise contribution score, filtering it by a threshold to generate the mask of patches $M \in \mathbb{R}^P$. We apply the patch-wise mask on the patchified image to select only the positive and negative

patches as follows:

$$M_{i,j} = \frac{1}{p^2} \sum_{m=0}^{p-1} \sum_{n=0}^{p-1} \mathbf{A}_{pi+m,pj+n} > \theta, \qquad (2)$$

$$\boldsymbol{x}_{\mathrm{pos}} = \boldsymbol{x}_{\mathrm{glb}} \odot M, \; \boldsymbol{x}_{\mathrm{neg}} = \boldsymbol{x}_{\mathrm{glb}} \odot (1 - M),$$

where θ is a pre-defined threshold for contribution scores set at 0.5 in this study, $\odot$ denotes the element-wise multiplication operator and $x_{\mathrm{glb}} \in \mathbb{R}^{P \times C}$ corresponds to the patches of the input image. Consequently, $\boldsymbol{x}_{\mathrm{pos}}$ and $\boldsymbol{x}_{\mathrm{neg}}$ denote the positive and negative patches, respectively. To promote a consistent grading on global and positive regions as well as distinguish the negative ones, FRL applies the Cross-Entropy loss on all the regions $\{\boldsymbol{x}_{\mathrm{glb}}, \boldsymbol{x}_{\mathrm{pos}}, \boldsymbol{x}_{\mathrm{neg}}\}$ as follows:

$$\sigma(\boldsymbol{x}, y) = \mathcal{L}_{\mathrm{CE}}(f \cdot g(\boldsymbol{x}), y)),$$

$$\mathcal{L}_{\mathrm{CLS}} = \frac{1}{B} \sum_{i=1}^{B} \sigma(\boldsymbol{x}_{\mathrm{glb}}^i, y^i) + \sigma(\boldsymbol{x}_{\mathrm{pos}}^i, y^i) + \sigma(\boldsymbol{x}_{\mathrm{neg}}^i, c), \qquad (3)$$

where B indicates the number of samples in a mini-batch, $\mathcal{L}_{\mathrm{CE}}$ denotes the Cross-Entropy loss, y^i is the label of the i-th sample, and c is a constant label for negative areas indicating background. To further boost the representations of histopathology slides, especially for the positive regions, we propose a consistency constraint that promotes a consistent mapping of positive and overall regions as well as a distinguishable mapping of negative regions. This approach pulls the positive and global features closer together as well as pushing the negative features far apart by the InfoNCE loss [23] defined on the features $\boldsymbol{v} = g(\boldsymbol{x})$ as follows:

$$\phi(\boldsymbol{v}_i, \boldsymbol{v}_j) = e^{\frac{\cos(\boldsymbol{v}_i, \boldsymbol{v}_j)}{\tau}},$$

$$\psi(\boldsymbol{v}) = \frac{\phi(\boldsymbol{v}_{\mathrm{glb}}, \boldsymbol{v}_{\mathrm{pos}})}{\phi(\boldsymbol{v}_{\mathrm{glb}}, \boldsymbol{v}_{\mathrm{pos}}) + \phi(\boldsymbol{v}_{\mathrm{glb}}, \boldsymbol{v}_{\mathrm{neg}}) + \phi(\boldsymbol{v}_{\mathrm{pos}}, \boldsymbol{v}_{\mathrm{neg}})}, \quad (4)$$

$$L_{\mathrm{FRL}} = -\frac{1}{B} \sum_{i=1}^{B} \log \psi(\boldsymbol{v}_i),$$

where $\boldsymbol{v}_{\mathrm{glb}}^i, \boldsymbol{v}_{\mathrm{pos}}^i, \boldsymbol{v}_{\mathrm{neg}}^i$ denote the features extracted from global, positive and negative regions, respectively. The $\cos(\cdot, \cdot)$ represents the cosine similarity function and $\tau \in \mathbb{R}^+$ is a scalar temperature parameter. By locating the areas that contributed the most significantly to accurate grading and regularizing the representations of different regions with the consistency constraint, FRL encourages the model to focus on diagnostic areas, promoting representation learning on the histopathology slides, thus improving the performance of glioma grading.

C. Multi-view Cross-modal Alignment

Current multimodal glioma grading methods encode and fuse the histopathology slides with molecular biomarkers [6]. Notably, unlike histopathology representations, molecular biomarkers usually with integer values, e.g., IDH mutation status, provide direct insights related to glioma grading without necessitating neural network processing [1]. This heterogeneity between two modalities presents a challenge for existing fusion-based methods, which struggle to align molecular

indicators with histologic features in the high-dimensional feature space [14]. To tackle this issue, we propose a Multi-view Cross-modal Alignment module that employs genomic biomarkers as unique labels and aligns the histological feature representations within each molecular subspace by supervised contrastive learning.

Specifically, MCA leverages the biomarkers with discrete values, including IDH mutation status, 1p/19q codeletion presence, and the Copy Number Variation (CNV), as labels. To accommodate the co-occurrence of biomarkers, this module projects multi-view histopathology representations $\boldsymbol{v} \in \{\boldsymbol{v}_{\mathrm{glb}}, \boldsymbol{v}_{\mathrm{pos}}, \boldsymbol{v}_{\mathrm{neg}}\}$ into individual molecular subspaces, employing distinct projection heads $h^n(\cdot)$ for each. Within these subspaces, it aims to bring features that share the same molecular biomarker values closer together while distancing those that differ, thereby enhancing the concordance of histopathology features regarding each biomarker. The loss of Multi-view Cross-modal Alignment is formulated as follows:

$$\omega(\boldsymbol{v}, i, j) = \frac{\phi(\boldsymbol{v}_i, \boldsymbol{v}_j)}{\sum_{k=1}^{B} \mathbb{1}_{i \neq k} \cdot \phi(\boldsymbol{v}_i, \boldsymbol{v}_k)},$$

$$\mathcal{L}_{\mathrm{MCA}}^n = -\frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{B} \mathbb{1}_{i \neq j} \cdot \mathbb{1}_{y_i^n = y_j^n} \cdot \log \omega(h^n(\boldsymbol{v}), i, j),$$

$$\mathcal{L}_{\mathrm{MCA}} = \sum_{n=1}^{N} \mathcal{L}_{\mathrm{MCA}}^n,$$

(5)

where N denotes the number of gene biomarkers, n indicates the index of the biomarker, $\mathbb{1}$ is an indicator function that returns 1 if the condition is satisfied and 0 otherwise, and y^n represents the status of the n-th biomarker. We set the gene labels of positive regions to the same as global regions and set the negative gene labels to the normal status (*i.e.*, wildtype for IDH and 1p/19q, CNV equals 0 for other biomarkers). Following recent medical studies [1], [24], we include the IDH mutation status, 1p/19q codeletion presence, and CNV of PTEN, EGFR, CARD11, and FGFR2 as molecular biomarkers in this study. By enhancing the histological feature representations with FRL and aligning molecular biomarkers to them, FoF promotes a reliable and accurate grading of gliomas, reaching even better performance with sole images than existing multimodal counterparts.

D. Training and Inference

The overall optimization objective of our FoF framework is summarized as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CLS}} + \lambda_1 \mathcal{L}_{\mathrm{FRL}} + \lambda_2 \mathcal{L}_{\mathrm{MCA}}, \qquad (6)$$

where λ_1 and λ_2 indicate the coefficients to control the trade-off of FRL and MCA, respectively. During training, FoF estimates the contribution score $\mathbf{A} \in \mathbb{R}^{H \times W}$ in the first back-propagation and calculates the losses of FRL and MCA modules in the second back-propagation. At the inference phase, the model $f \cdot g$ directly outputs the predictions of the input images with no additional computational cost.

TABLE I
COMPARISON WITH **HISTOPATHOLOGY GRADING** ON THE TCGA GBM-LGG DATASET.

Method	AUC (%)	AP (%)	Accuracy (%)	Kappa (%)
Baseline	90.37 ± 0.82	83.05 ± 2.39	74.36 ± 2.14	61.16 ± 3.84
PathCNN [19]	90.46 ± 1.01	82.84 ± 2.07	74.07 ± 1.81	60.59 ± 3.05
PathGCN [6]	90.25 ± 1.53	82.69 ± 2.11	74.20 ± 1.49	61.27 ± 2.92
SwinV2-Tiny [25]	91.56 ± 1.74	83.27 ± 1.26	74.34 ± 1.56	62.35 ± 2.46
KL div [26]	91.67 ± 0.43	85.01 ± 0.95	74.58 ± 1.15	61.39 ± 2.17
PKT [27]	91.52 ± 0.52	84.67 ± 1.03	74.77 ± 1.05	61.71 ± 1.89
Feats KL [28]	91.70 ± 0.43	85.06 ± 0.74	74.82 ± 1.66	61.76 ± 2.97
SP [29]	91.89 ± 0.45	85.38 ± 0.54	75.60 ± 1.12	62.96 ± 1.96
RKD [30]	91.67 ± 0.35	85.03 ± 0.65	74.97 ± 1.06	62.00 ± 2.01
CRD [31]	91.86 ± 0.68	85.33 ± 1.34	75.71 ± 1.27	63.16 ± 2.04
SCKD [32]	91.31 ± 0.34	84.44 ± 0.76	73.91 ± 1.28	60.42 ± 2.32
HKD [9]	91.73 ± 0.32	85.18 ± 0.59	75.21 ± 0.85	62.37 ± 1.49
CLAT [10]	92.42 ± 0.58	86.34 ± 1.23	76.47 ± 0.65	64.34 ± 0.85
FoF *w/o* FRL	93.20 ± 0.53	88.44 ± 1.48	76.89 ± 1.60	64.31 ± 2.13
FoF *w/o* MCA	93.64 ± 0.42	88.26 ± 0.93	76.55 ± 2.32	63.85 ± 2.55
FoF	**94.17 ± 0.68**	**89.98 ± 1.35**	**79.98 ± 2.17**	**69.05 ± 2.24**

TABLE II
COMPARISON WITH **MULTIMODAL GRADING** ON THE TCGA GBM-LGG DATASET.

Method	AUC (%)	AP (%)	Accuracy (%)	Kappa (%)
SCNN [7]	91.18 ± 0.85	84.40 ± 1.83	74.40 ± 2.09	61.02 ± 1.34
Pathomic [6]	92.30 ± 0.79	85.80 ± 1.55	75.74 ± 1.39	63.15 ± 2.31
Gpdbn [18]	92.01 ± 0.56	85.42 ± 1.12	75.98 ± 1.42	63.02 ± 1.56
Hfbsurv [33]	91.61 ± 0.75	84.79 ± 1.39	74.67 ± 2.21	61.84 ± 1.65
DDM-net [21]	92.53 ± 0.91	85.22 ± 1.01	76.92 ± 1.76	64.07 ± 1.95
FOAA [8]	93.15 ± 0.76	86.70 ± 1.14	77.90 ± 2.41	65.36 ± 2.38
DGKD [19]	93.43 ± 0.40	87.75 ± 0.86	78.08 ± 0.73	66.73 ± 1.31
FoF	**94.17 ± 0.68**	**89.98 ± 1.35**	**79.98 ± 2.17**	**69.05 ± 2.24**

IV. EXPERIMENT

A. Dataset and Implementation Details

Dataset. We follow [19] to evaluate our FoF framework on the combination of TCGA-GBM and TCGA-LGG datasets [34], which comprise paired histopathology slides and genomic profiles. There are a total of 736 patients with standard grading labels, including 182 grade II, 205 grade III, and 350 grade IV. Following [6], [7], we utilize $1,325$ Region of Interest (ROI) images of dimensions $1,024 \times 1,024$ from pathology modality. The ROI images are augmented through random cropping, color jittering, flipping, and distortion. We perform 5-fold cross-validation and report the average test performance with the standard deviation.

Implementation. We implement the proposed framework with the PyTorch library [35] and employ ViT-Tiny [36], [37] as the encoder for images. All the experiments are done on four NVIDIA GeForce GTX 1080 Ti GPUs with a batch size of 4. For network optimization, we employ AdamW optimizer, configuring it with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.01. The initial learning rate is set as 1×10^{-4} and follows a cosine decay schedule down to 0. The framework is trained for 70 epochs. Both the loss weights λ_1 and λ_2 are set as 0.2. The temperature parameter τ is set as 0.07.

Evaluation Metrics. Following the evaluation protocol in [10], [19], we adopt four commonly-used metrics to comprehensively evaluate the performance of glioma grading, including the Average Precision (AP), Accuracy, Area Under
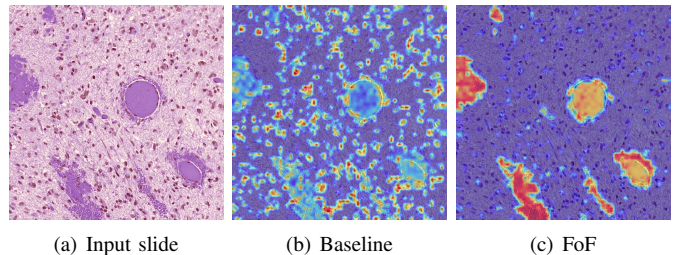

(a) Input slide   (b) Baseline   (c) FoF

Fig. 2. Visualization of (a) the input pathology slide, (b) the CAM produced by baseline method ViT-Tiny, and (c) the CAM generated by the proposed FoF framework. As illustrated, FoF focuses on microvascular proliferation, which leads to the diagnosis of Glioblastoma (Grade IV).

the Curve (AUC), and Kappa score. Higher scores of these metrics represent a more accurate classification of the gliomas.

B. Comparison with Histopathology Grading Methods

We compare our FoF with state-of-the-art methods that require only histopathology slides as input during inference, including popular image classification models (ViT-Tiny as Baseline, PathCNN [19], PathGCN [6], and SwinV2 [25]). For a more fair comparison, we implement various latest relevant comparisons based on knowledge distillation, which also utilizes multimodal data for training while requiring image-only data for inference. As shown in Table I, FoF outperforms other histopathology grading methods in all metrics on the TCGA GBM-LGG dataset, achieving an Accuracy of 79.98%,
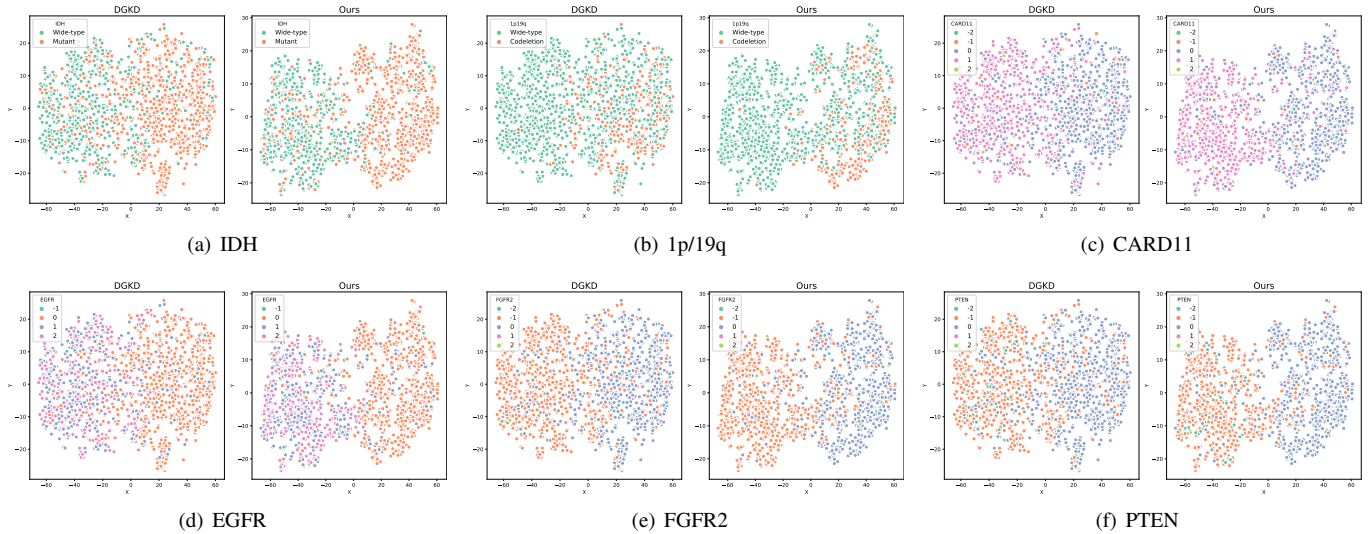
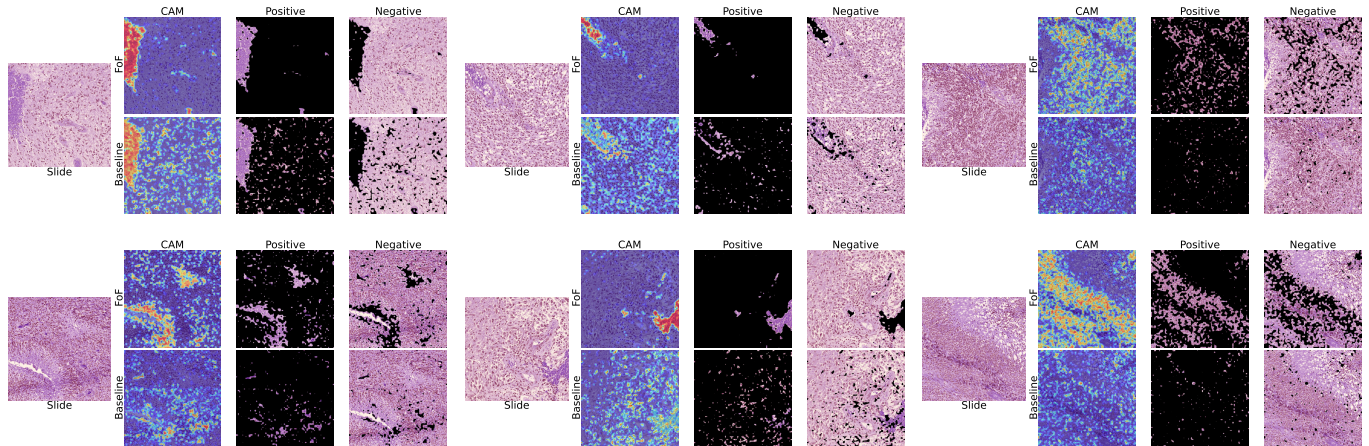Fig. 3. Comparisons with the baseline on the molecular-pathology knowledge alignment.

Fig. 4. Comparison with the baseline on model focus. As illustrated, our FoF framework concentrates more precisely on the key structures of Glioblastoma, such as microvascular proliferation and pseudopalisadeing necrosis.

an AUC of 94.17%, an Average Precision of 89.98%, and a Kappa of 69.05%. FoF significantly outperforms the baseline ViT model, showing 5.62%, 3.80%, 6.93%, 7.89% increases in Accuracy, AUC, Average Precision, Kappa, respectively. Notably, compared with the most recent study CLAT [10], FoF achieves a 3.51% increase in Accuracy, 1.75% increase in AUC, 3.64% increase in Average Precision, and 4.71% increase in Kappa. These comparisons confirm the effectiveness of our FoF framework.

We have also compared the Class Activation Map generated by the baseline model and our FoF framework. As illustrated in Fig. 2, our FoF framework focuses more precisely on regions of microvascular proliferation, a key histopathology feature of Glioblastoma (Grade IV).

Ablation study. We perform the ablation study to validate the effectiveness of our proposed FRL and MCA modules on the TCGA GBM-LGG dataset. As shown in Table I, both FRL and MCA modules remarkably improve the performance over the

baselines. In particular, we apply two ablative baselines of the proposed FoF framework by disabling the FRL (denoted as FoF w/o FRL) and the MCA (denoted as FoF w/o MCA) individually. In detail, as shown in Table I, for FRL, the Accuracy increases by 3.09%, AUC increases by 0.97%, Average Precision increases by 1.54%, and Kappa increases by 4.74%. For MCA, the Accuracy increases by 3.43%, AUC increases by 0.53%, Average Precision increases by 1.72%, and Kappa increases by 5.20%. The ablation study verifies the effectiveness of our FRL and MCA modules.

C. Comparison with Multimodal Grading Methods

We compare our FoF with multimodal grading methods that require paired histopathology slides and genomic data for glioma grading. These methods implement various feature fusion methods, including concatenation (SCNN [7], Gpdbn [18]), Kronecker Product (Pathomic [6], DGKD [19]), hierarchical fusion (Hfbsurv [33]), and attention (DDM-net

[21], FOAA [8]). As shown in Table II, our FoF with sole histopathology slides outperforms all the multimodal grading methods on all metrics, significantly illustrating the effectiveness of the proposed FoF framework. In detail, our FoF outperforms the most recent multimodal grading method DGKD [19] with improvements of 1.9% in Accuracy, 0.74% in AUC, 2.23% in Average Precision, and 2.32% in Kappa. Compared to FOAA [8], which employs more advanced attention-based feature fusion, FoF achieved 2.08%, 1.02%, 3.28%, 3.69% increases in Accuracy, AUC, Average Precision, Kappa, respectively.

D. Qualitative Analysis

We have illustrated the distributions of morphological features regarding the different values of molecular biomarkers in Fig. 3. Compared to DGKD [19], our FoF presents a clearer clustering of histopathology features with respect to the molecular biomarkers, illustrating a more effective alignment between the two modalities.

Furthermore, we present more comparisons of the model focus on input histopathology slides in Fig. 4. All the images presented have been reviewed by senior pathologists. It is worth noting that our FoF concentrates on both microvascular proliferation and pseudopalisading necrosis, critical indicators for diagnosing HGG, whereas the baseline model focuses on other, less relevant details.

V. CONCLUSION

In this work, we propose the FoF framework that utilizes pathology-genomic knowledge toward accurate glioma grading with histopathology slides. To enhance the representation learning on histopathology slides, we propose the FRL module that encourages the model to focus on diagnostic regions and the MCA scheme which efficiently aligns the molecular biomarkers with visual representations. Experimental results indicate that FoF improves the glioma grading significantly. Particularly, with the sole histopathology slides, FoF achieves superior performance with existing multimodal approaches, which is of great clinical significance.

REFERENCES

[1] D. N. Louis, A. Perry, P. Wesseling, D. J. Brat, I. A. Cree, D. Figarella-Branger, C. Hawkins, H. Ng, S. M. Pfister, G. Reifenberger *et al.*, "The 2021 who classification of tumors of the central nervous system: a summary," *Neuro-oncology*, vol. 23, no. 8, pp. 1231–1251, 2021.

[2] D. N. Louis, A. Perry, G. Reifenberger, A. Von Deimling, D. Figarella-Branger, W. K. Cavenee, H. Ohgaki, O. D. Wiestler, P. Kleihues, and D. W. Ellison, "The 2016 world health organization classification of tumors of the central nervous system: a summary," *Acta neuropathologica*, vol. 131, pp. 803–820, 2016.

[3] J. Shanes, J. Ghali, M. Billingham, V. Ferrans, J. Fenoglio, W. Edwards, C. Tsai, J. Saffitz, J. Isner, and S. Furner, "Interobserver variability in the pathologic interpretation of endomyocardial biopsy results." *Circulation*, vol. 75, no. 2, pp. 401–405, 1987.

[4] D. Anand, S. Gadiya, and A. Sethi, "Histographs: graphs in histopathology," in *Medical Imaging 2020: Digital Pathology*, vol. 11320. SPIE, 2020, pp. 150–155.

[5] J. Wang, R. J. Chen, M. Y. Lu, A. Baras, and F. Mahmood, "Weakly supervised prostate tma classification via graph convolutional networks," in *ISBI*. IEEE, 2020, pp. 239–243.

[6] R. J. Chen, M. Y. Lu, J. Wang, D. F. Williamson, S. J. Rodig, N. I. Lindeman, and F. Mahmood, "Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," *IEEE Transactions on Medical Imaging*, 2020.

[7] P. Mobadersany, S. Yousefi, M. Amgad, D. A. Gutman, J. S. Barnholtz-Sloan, J. E. Velázquez Vega, D. J. Brat, and L. A. Cooper, "Predicting cancer outcomes from histology and genomics using convolutional networks," *Proceedings of the National Academy of Sciences*, vol. 115, no. 13, pp. E2970–E2979, 2018.

[8] O. Alwazzan, I. Patras, and G. Slabaugh, "Foaa: Flattened outer arithmetic attention for multimodal tumor classification," *arXiv preprint arXiv:2403.06339*, 2024.

[9] S. Zhou, Y. Wang, D. Chen, J. Chen, X. Wang, C. Wang, and J. Bu, "Distilling holistic knowledge with graph neural networks," in *ICCV*, 2021, pp. 10 387–10 396.

[10] X. Xing, M. Zhu, Z. Chen, and Y. Yuan, "Comprehensive learning and adaptive teaching: Distilling multi-modal knowledge for pathological glioma grading," *Medical Image Analysis*, vol. 91, p. 102990, 2024.

[11] S. Zhang, S. Zheng, Z. Shui, H. Li, and L. Yang, "Multi-modal learning with missing modality in predicting axillary lymph node metastasis," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2023, pp. 2395–2400.

[12] X. Wang, D. Wang, Z. Yao, B. Xin, B. Wang, C. Lan, Y. Qin, S. Xu, D. He, and Y. Liu, "Machine learning models for multiparametric glioma grading with quantitative result interpretations," *Frontiers in neuroscience*, vol. 12, p. 1046, 2019.

[13] Y. Zhou, S. Graham, N. Alemi Koohbanani, M. Shaban, P.-A. Heng, and N. Rajpoot, "Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images," in *ICCV workshops*, 2019, pp. 0–0.

[14] X. Wang, S. Price, and C. Li, "Multi-task learning of histology and molecular markers for classifying diffuse glioma," *arXiv preprint arXiv:2303.14845*, 2023.

[15] A. Perry and P. Wesseling, "Histologic classification of gliomas," *Handbook of clinical neurology*, vol. 134, pp. 71–95, 2016.

[16] J. Ker, Y. Bai, H. Y. Lee, J. Rao, and L. Wang, "Automated brain histology classification using machine learning," *Journal of Clinical Neuroscience*, vol. 66, pp. 239–245, 2019.

[17] S. Rathore, T. Niazi, M. A. Iftikhar, and A. Chaddad, "Glioma grading via analysis of digital pathology images using machine learning," *Cancers*, vol. 12, no. 3, p. 578, 2020.

[18] Z. Wang, R. Li, M. Wang, and A. Li, "Gpdbn: deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction," *Bioinformatics*, vol. 37, no. 18, pp. 2963–2970, 2021.

[19] X. Xing, Z. Chen, M. Zhu, Y. Hou, Z. Gao, and Y. Yuan, "Discrepancy and gradient-guided multi-modal knowledge distillation for pathological glioma grading," in *MICCAI*. Springer, 2022, pp. 636–646.

[20] X. Xing, Z. Chen, Y. Hou, and Y. Yuan, "Gradient modulated contrastive distillation of low-rank multi-modal knowledge for disease diagnosis," *Medical Image Analysis*, vol. 88, p. 102874, 2023.

[21] L. Qiu, L. Zhao, W. Zhao, and J. Zhao, "Dual-space disentangled-multimodal network (ddm-net) for glioma diagnosis and prognosis with incomplete pathology and genomic data," *Physics in Medicine & Biology*, vol. 69, no. 8, p. 085028, 2024.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626.

[23] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[24] P. Śledzińska, M. G. Bebyn, J. Furtak, J. Kowalewski, and M. A. Lewandowska, "Prognostic and predictive biomarkers in gliomas," *International journal of molecular sciences*, vol. 22, no. 19, p. 10373, 2021.

[25] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *CVPR*, 2022, pp. 12 009–12 019.

[26] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[27] N. Passalis and A. Tefas, "Learning deep representations with probabilistic knowledge transfer," in *ECCV*, 2018, pp. 268–284.

[28] M. Hu, M. Maillard, Y. Zhang, T. Ciceri, G. La Barbera, I. Bloch, and P. Gori, "Knowledge distillation from multi-modal to mono-modal segmentation networks," in *MICCAI*. Springer, 2020, pp. 772–781.

[29] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *ICCV*, 2019, pp. 1365–1374.

[30] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *CVPR*, 2019, pp. 3967–3976.

[31] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," *arXiv preprint arXiv:1910.10699*, 2019.

[32] Y. Zhu and Y. Wang, "Student customized knowledge distillation: Bridging the gap between student and teacher," in *ICCV*, 2021, pp. 5057–5066.

[33] R. Li, X. Wu, A. Li, and M. Wang, "Hfbsurv: hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction," *Bioinformatics*, vol. 38, no. 9, pp. 2587–2594, 2022.

[34] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "Review the cancer genome atlas (tcga): an immeasurable source of knowledge," *Contemporary Oncology/Współczesna Onkologia*, vol. 2015, no. 1, pp. 68–77, 2015.

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[37] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your vit? data, augmentation, and regularization in vision transformers," *arXiv preprint arXiv:2106.10270*, 2021.