Weakly Supervised ViT-Conditioned Text Explanations in Glioma Histology

***Abstract*—Predicting the promoter methylation of O^6-methylguanine–DNA–methyltransferase (MGMT) from hematoxylin–eosin (H&E)–stained glioma biopsies using deep learning (DL) remains challenging due to limited interpretability. In this work, we propose a weakly supervised framework that jointly performs classification and generates textual explanations grounded in visual evidence. We employ a lightweight convolutional neural network (CNN) to predict MGMT status from histology images. Building on these predictions, we introduce a Vision Transformer (ViT)–conditioned language decoder guided by SmoothGrad-CAM saliency maps to produce class-aware textual explanations. Patch-level predictions and saliency maps guide the generation of concise rationales aligned with visually relevant tissue regions. We further assess explanation faithfulness using perturbation-based insertion and deletion analyses. On a glioblastoma (GBM) cohort, the proposed model achieves a patient-level F1-score of 0.94, while the generated explanations remain consistent with both saliency maps and model predictions, demonstrating that class-conditioned textual rationales can provide a faithful and interpretable interface for histology-based MGMT prediction.**

***Index Terms*—Explainable AI, MGMT Methylation, Glioblastoma, Weakly Supervised Learning, Attention-Based Deep Learning, Vision Transformers, Textual Rationales.**

I. Introduction

Diffuse gliomas are the most common primary tumors of the central nervous system, with glioblastoma (GBM) representing the most aggressive subtype despite advances in surgery, radiotherapy, and chemotherapy [1]. Among clinically relevant molecular biomarkers, the methylation status of the O^6-methylguanine–DNA methyltransferase (MGMT) promoter plays a central role in treatment planning and prognosis. MGMT-methylated tumors typically exhibit increased sensitivity to alkylating agents and improved survival outcomes [2]. Reliable determination of MGMT status is therefore essential for guiding therapeutic decisions.

Recent advances in computational pathology have shown that deep learning (DL) models can predict molecular biomarkers directly from H&E-stained histology images [3]–[6]. Weakly supervised approaches are particularly effective for whole-slide analysis, enabling learning from slide-level labels while localizing discriminative regions [7]. However, existing explainability methods often remain disconnected from the prediction process, producing visual or textual outputs that are not tightly coupled to model decisions.

More recently, transformer-based architectures have achieved strong performance in computer vision tasks, with Vision Transformers (ViTs) providing powerful global feature representations [13]. Vision–language models have also been explored for medical image captioning and automated report generation, enabling natural language descriptions to accompany visual predictions [14]–[16]. However, most existing approaches treat classification, visual explanation, and language generation as independent components, which can lead to explanations that are weakly connected to the underlying prediction process.

To address these limitations, we propose a weakly supervised framework that jointly predicts MGMT promoter methylation from glioma histology and generates aligned textual explanations. The approach integrates a lightweight DL classifier with a ViT-based captioning module guided by SmoothGrad-CAM saliency maps, enabling class-aware explanations grounded in visually discriminative regions without requiring region-level annotations.

The classifier first produces MGMT predictions, while the corresponding saliency maps identify the most informative tissue regions. These visual cues, combined with the predicted class, are then used to condition a transformer-based decoder that generates concise and interpretable textual rationales. This tight integration between prediction, visual attribution, and language generation ensures that the resulting explanations remain faithful to the model's decision process.

The main contributions of this work are threefold: (i) a unified weakly supervised framework for joint classification and explanation generation; (ii) a lightweight MobileNetV2-based architecture enhanced with attention mechanisms and stabilized NdLinear projections for MGMT prediction; and (iii) a comprehensive evaluation of explanation faithfulness using insertion–deletion analyses and caption–saliency alignment metrics.

The rest of this paper is organized as follows: Section II details the data and preprocessing, Section III presents the methodology, Section IV reports the results, Section V discusses the findings, and Section VI concludes the paper.

II. Data and Preprocessing

We consider a private GBM cohort comprising $N_{\mathrm{GBM}} = 299$ patients (121 MGMT-methylated (mMGMT) and 178 MGMT-unmethylated (uMGMT)), with patient-level MGMT promoter methylation labels provided by neuropathological assessment [17]. Whole-slide H&E images are subdivided into non-overlapping patches of size 512×512 pixels using a sliding-window strategy. Each patch inherits the MGMT label of its parent slide.

To suppress background regions and reduce noise from non-informative tissue, we compute a hematoxylin-based cellu-

larity score for each patch. For a given patch I, the score C is defined as: $C = \frac{1}{|I|}\sum_{x\in I}\mathbb{1}(H(x) > \tau)$, where $H(x)$ denotes the hematoxylin intensity at pixel x, $\tau = 0.05$ is the predefined threshold, and $\mathbb{1}$ is the indicator function. Patches with negligible tissue (i.e., $C \approx 0$) are discarded.

For each patient, patches are ranked according to their cellularity, and only the top $k = 5$ are retained. Patient-level predictions are then obtained by averaging the probabilities across the selected patches. This mean aggregation provides a simple yet robust and parameter-free strategy that captures distributed evidence while avoiding the sensitivity of max pooling to outliers and the additional complexity of attention-based aggregation.

The choice of $k = 5$ reflects a trade-off between capturing intra-tumoral heterogeneity and limiting redundancy: smaller values ($k < 5$) increased prediction variance, whereas larger values ($k > 5$) yielded marginal gains with higher overlap between patches.

Selected patches undergo Macenko stain normalization to reduce inter-slide variability, followed by data augmentation including 90° rotations, horizontal and vertical flips, and mild color jittering to improve robustness while preserving relevant morphology.

At the patient level, the dataset is split into 209, 30, and 60 patients for training, validation, and testing, respectively. To prevent information leakage, all patches from a given patient are assigned to a single subset. Five-fold cross-validation is performed on the 239-patient training–validation cohort with class balance preserved across folds. The final model is selected based on cross-validation performance, retrained on the full training–validation set, and evaluated on the held-out test set.

III. Methods

A. Base Classifier

The proposed classifier comprises five main components: (i) a MobileNetV2 backbone for hierarchical feature extraction, (ii) a 1×1 convolutional projection layer, (iii) a shared attention mechanism that jointly captures channel-wise, spatial, and local contextual relevance, (iv) a stack of multidimensional NdLinear layers for expressive yet stable feature projection, and (v) a final linear classification head that produces patient-level MGMT predictions.

1) MobileNet Backbone + 1×1 Convolution Layer: We use a fully trainable MobileNetV2 backbone [4] followed by a 1×1 projection layer to adapt feature dimensionality while preserving spatial structure.

2) Shared Attention Mechanism: We introduce a shared attention mechanism that jointly models channel importance, spatial saliency, and local neighborhood interactions. Specifically, channel attention captures global dependencies by applying global average pooling followed by a 1×1 convolution to reweight feature maps according to their semantic relevance. In parallel, spatial attention learns pixel-level importance through stacked 3×3 and 1×1 convolutions, enabling the network to focus on salient regions. Complementing these components, a neighbor attention module incorporates local context via depthwise filtering with center-suppressed kernels, enhancing edge-aware interactions and promoting spatial continuity.

To further guide attention toward meaningful structures, we introduce a hierarchical correlation signal by sequentially applying spatial max pooling and global average pooling. This signal acts as a soft voting mechanism, highlighting discriminative super-pixels while suppressing background noise, offering a more robust alternative to conventional global pooling.

The outputs of the three attention branches are fused through a lightweight 1×1 convolution and applied to the input feature maps via element-wise modulation. This attention-guided reweighting allows the network to dynamically emphasize salient spatial regions and semantically relevant channels.

3) NdLinear Layers: To increase representational expressiveness without sacrificing parameter efficiency, we employ a sequence of NdLinear layers [18] following the attention module. To ensure stability, weights are initialized using scaled Kaiming initialization, reducing gradient instability during training. This scaling constrains the effective Lipschitz constant of each NdLinear block, resulting in smoother training dynamics and improved stability without sacrificing final performance.

4) Final Classification Head: The output of the NdLinear layers is flattened and passed to a fully connected classification head, producing class logits. A sigmoid activation converts the logits into a probability of MGMT promoter methylation, used for both patient-level prediction and explanation conditioning.

5) Training Setup: The model is trained using AdamW with warm-up and cosine annealing, binary cross-entropy loss, and gradient clipping.

B. ViT-Conditioned Captioner

The explanation module g_ϕ generates a concise textual rationale $s = (w_1, \ldots, w_T)$ for a given histology patch x, explicitly conditioned on a binary MGMT class label $c \in \{0, 1\}$. Caption generation follows an auto-regressive language modeling formulation, in which the conditional probability of a sequence factorizes over time steps as $p_\phi(s \mid x, c) = \prod_{t=1}^{T} p_\phi(w_t \mid w_{<t}, x, c)$. This formulation enables the model to jointly account for visual features, MGMT class prediction, and linguistic context when producing textual explanations. The overall architecture of the proposed framework is illustrated in Fig. 1:

1) Image Encoding and Class Conditioning: Global visual representations are extracted using ViT-B/16 encoder. Given an input patch x, the ViT outputs a sequence of token embeddings, including a dedicated `[CLS]` token summarizing the overall image content. We retain this token as a compact image-level descriptor: $h_{\text{img}} = \text{ViT}(x)_{\texttt{[CLS]}} \in \mathbb{R}^{d_{\text{ViT}}}, \quad d_{\text{ViT}} = 768$. To align the visual representation with the language decoder embedding space, h_{img} is linearly projected as $v = W_{\text{img}} h_{\text{img}} + b_{\text{img}}$, where $W_{\text{img}} \in \mathbb{R}^{d \times d_{\text{ViT}}}$ and $d = 512$ denotes the decoder model dimension.

Class conditioning is introduced via a learnable embedding of the MGMT label. Specifically, the discrete class c is mapped

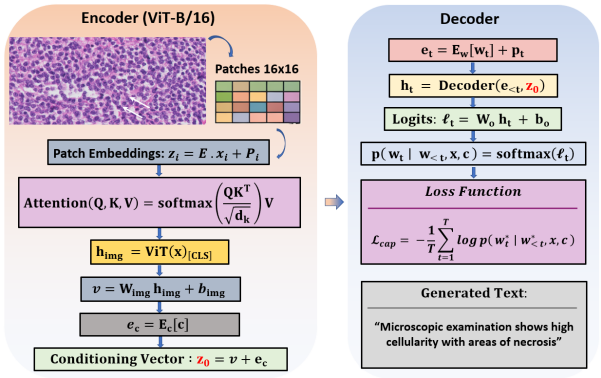

Fig. 1: Architecture of the proposed ViT-based encoder–decoder framework for conditional text generation.

to $e_c = E_c[c] \in \mathbb{R}^d$, where $E_c \in \mathbb{R}^{2\times d}$ is a trainable class embedding matrix. The projected image representation and the class embedding are combined through simple addition to form the conditioning token: $z_0 = v + e_c$, which serves as memory input to the transformer decoder. During, training, the ground-truth MGMT label is used for conditioning, whereas at inference time captions are conditioned on the patient-level classifier prediction $\hat{y}_{\text{pat}}$, ensuring that the generated explanations remain aligned with the model's decision.

2) Transformer Decoder: Textual rationales are generated using a transformer decoder composed of $L = 3$ layers, with model dimension $d = 512$, 8 attention heads, and a feedforward dimension of 2048. Let $E_w \in \mathbb{R}^{V\times d}$ denote the word embedding matrix and p_t the positional embedding at decoding step t. The decoder input embedding is given by $e_t = E_w[w_t] + p_t$. Using causal self-attention masking, the decoder produces hidden states $h_t \in \mathbb{R}^d$, which are mapped to token-level logits $\ell_t = W_o h_t + b_o$. Token probabilities are obtained via a softmax operation, $p_\phi(w_t \mid w_{<t}, x, c) = \text{softmax}(\ell_t)_{w_t}$. The captioning module is trained using a sequence-level cross-entropy loss, $\mathcal{L}_{\text{cap}} = -\frac{1}{T}\sum_{t=1}^{T} \log p_\phi(w_t^* \mid w_{<t}^*, x, c)$, where (w_t^*) denotes the weakly supervised pseudo-caption associated with patch x, as described in the next Section III-C.

C. Weakly Supervised Pseudo-Captions

We adopt a weakly supervised strategy to construct pseudo-captions encoding coarse morphological and spatial information. For each MGMT class c, we define a class-specific set $\mathcal{S}_c$ of domain-informed phrases describing common histological patterns. Rather than relying on rigid templates, pseudo-captions are generated by randomly sampling a subset of 2–4 phrases $s_k \in \mathcal{S}_c$ and composing them into lightweight sentence structures. This stochastic composition introduces variability in both content and ordering, encouraging the model to learn associations between visual features and semantic descriptors rather than memorizing fixed sentence patterns.

To introduce spatial grounding, we leverage SmoothGrad-CAM saliency maps computed from the base classifier. For each patch x, the saliency map M is partitioned into four quadrants R_1, R_2, R_3, R_4, and the mean activation values $(q_1, \ldots, q_4)$ are computed using $q_i = \frac{1}{|R_i|}\sum_{(u,v)\in R_i} M(u,v)$.

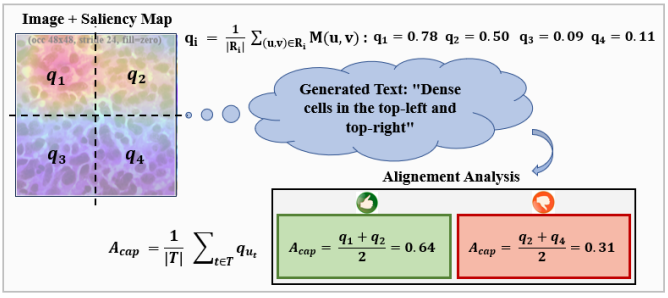

Fig. 2: Example of text–saliency alignment.

Quadrants exhibiting high activation are associated with spatial descriptors (e.g. "in the top-left quadrant"), which are optionally appended to morphological phrases. This mechanism encourages the decoder to associate visual features, class conditioning, and saliency-derived spatial cues within a unified textual explanation.

For captions that explicitly reference spatial locations, we define a caption-saliency alignment score: $A_{\text{cap}} = \frac{1}{|T|}\sum_{t\in T} q_{u_t}$, where T denotes the set of spatially tagged tokens in the caption and q_{u_t} is the SmoothGrad-CAM activation of the quadrant referenced by token t. This score provides a compact quantitative measure of how closely textual location cues align with visually salient regions. An example of text-saliency is presented in Fig. 2:

IV. Results

A. Evaluation Metrics

In this study, we report standard classification metrics, which are widely used in classification task, including: Accuracy (Acc), Precision (Prec), Recall (Rec), and F1-score (F1). All quantitative results are reported at the patient level.

B. Classification Performance and Model Comparison

To benchmark the proposed MobileNetV2 + Attention + NdLinear classifier (MBV2+Att+NdLinear), we compare against widely used convolutional backbones in computational pathology, including ResNet-50 and EfficientNet-B0 (EffNet-B0), all initialized with ImageNet pretraining. For each model, the original classification head is replaced with a single-unit sigmoid output to predict MGMT promoter methylation probability.

To ensure a controlled and fair comparison, all backbones are trained using the same optimization strategy, learning rate schedule, data augmentation pipeline, batch size, and number of epochs. In addition, all models are augmented with the same attention and NdLinear modules.

Table I reports patient-level performance obtained using five-fold cross-validation on the training–validation cohort. For each model, reported training and validation metrics correspond to the mean values computed across the five cross-validation folds.

C. Comparative Analysis of Explanation Methods

To justify the choice of SmoothGrad-CAM as the primary explanation method, we compare it qualitatively with two

TABLE I: Patient-level performance (5-fold CV) on the GBM cohort.

Model	Phase	Loss	Acc.	Prec.	Rec.	F1
EffNet-B0 †	Train	0.256	0.894	0.866	0.872	0.859
	Val	0.266	0.870	0.796	0.920	0.839
ResNet-50†	Train	0.187	0.930	0.923	0.924	0.917
	Val	0.820	0.802	0.764	0.846	0.763
Custom	Train	0.101	0.955	0.950	0.941	0.942
	Val	0.081	0.979	0.950	0.941	0.945
	Test	-	0.950	0.960	0.923	0.941

† Baselines use the same attention and NdLinear modules for fair comparison.

widely used model-agnostic approaches, namely LIME [11] and SHAP [12]. As illustrated in Fig. 3, although LIME and SHAP provide useful insights into feature importance at the super-pixel level, their explanations tend to be spatially fragmented and highly sensitive to the underlying image segmentation. This behavior limits their effectiveness for fine-grained localization in histopathological images. In contrast, SmoothGrad-CAM leverages gradient information from deep convolutional layers to produce spatially coherent attribution maps that better reflect the hierarchical feature extraction process of the classifier. This results in more consistent and anatomically meaningful localization of discriminative regions.

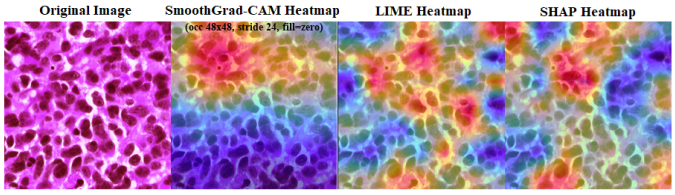

Fig. 3: Qualitative comparison of explanation methods on a representative histopathology patch.

While the proposed classifier incorporates internal attention mechanisms, these are primarily designed to modulate intermediate feature representations and do not directly correspond to class-discriminative evidence at the pixel level. Consequently, attention maps alone cannot be reliably interpreted as explanations of model decisions.

Given the need for spatially coherent and class-specific localization—particularly to support region-aware textual explanations in the ViT-conditioned captioning module—we adopt SmoothGrad-CAM as a principled and effective method for visual attribution.

D. ViT-Conditioned Textual Explanations

Building on the spatial consistency of SmoothGrad-CAM explanations, we evaluate the ViT-conditioned captioning module. As illustrated in Fig. 4, the generated explanations remain aligned with both the predicted MGMT status and the spatial structure of the saliency maps. This demonstrates that the ViT captioner does not act as an independent language generator, but instead produces rationales that are tightly coupled to the classifier's decision process.

E. Faithfulness of Saliency Maps

To assess whether the SmoothGrad-CAM saliency maps faithfully reflect the classifier's decision process, we adopt the insertion and deletion evaluation protocol commonly used in explainable image classification. These analyses quantify how the predicted confidence of the model changes when pixels ranked as salient are progressively revealed or removed.

For a given patch x and its predicted class, let $p_{\text{orig}} = s(x)$ denote the classifier confidence. The SmoothGrad-CAM map $\tilde{\text{CAM}}(x)$ is first normalized and used to rank pixels in decreasing order of saliency. In the *insertion* setting, we start from a heavily blurred version of the image, x_0, and progressively insert pixels from x according to this ranking, yielding a sequence of images $\{x_{\text{ins}}^{(p)}\}_{p\in[0,1]}$, where p denotes the fraction of inserted pixels. In the *deletion* setting, we start from the original image and progressively mask the most salient pixels, producing a corresponding sequence $\{x_{\text{del}}^{(p)}\}_{p\in[0,1]}$.

At each step, we evaluate the classifier confidence $p^{(p)} = s(x^{(p)})$ for the ground-truth MGMT class. A faithful saliency map is expected to yield a rapidly increasing confidence curve in the insertion setting and a steeply decreasing curve in the deletion setting, indicating that highly ranked pixels have a strong influence on the model's prediction.

Figure 5 illustrates representative insertion and deletion curves for both mMGMT and uMGMT test patches. In these examples, insertion curves rise sharply as a small fraction of salient pixels is revealed, while deletion curves exhibit a pronounced confidence drop when the most salient regions are removed. Similar qualitative trends are observed across additional test samples, supporting that the SmoothGrad-CAM maps concentrate on image regions that are influential for the MGMT classification rather than diffuse background structure.

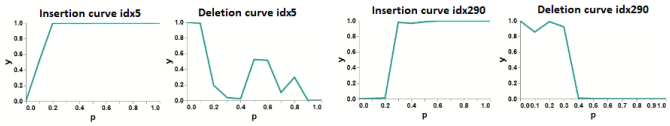

Fig. 5: Insertion and deletion faithfulness curves for two representative GBM test patches.

F. Ablation Analysis: Role of Class Conditioning in Textual Explanations

To assess the contribution of explicit MGMT class conditioning in the ViT-based captioning module, we conducted a targeted ablation in which the class embedding e_c was removed and captions were generated based solely on visual features. All other components of the framework, including the classifier, saliency maps, and training protocol, were kept unchanged.

Qualitatively, removing class conditioning resulted in more generic morphological descriptions that remained visually plausible but were less discriminative with respect to MGMT status. In particular, captions generated without class conditioning exhibited reduced specificity in describing cellular

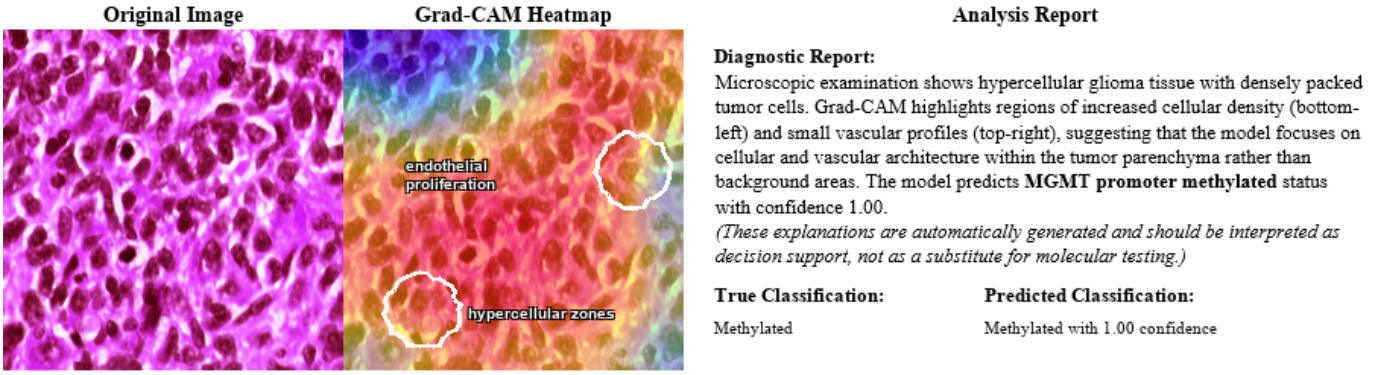

Fig. 4: Illustration of the proposed explainability pipeline: original histology image, SmoothGrad-CAM saliency map, and the corresponding generated textual explanation conditioned on MGMT prediction and spatially salient regions.

density and necrotic patterns, and more frequently overlapped across MGMT-methylated and unmethylated cases.

Quantitatively, we evaluate explanation quality using the caption–saliency alignment score A_cap, which measures the consistency between spatial references in the generated text and the corresponding saliency regions. While absolute values may vary across samples, we consistently observe higher alignment when both class conditioning and saliency grounding are enabled. In contrast, removing class conditioning or omitting saliency-based spatial cues leads to a noticeable decrease in A_cap, indicating weaker correspondence between textual explanations and visually discriminative regions. These results confirm that both components play a complementary role in ensuring that generated explanations remain faithful to the underlying prediction process.

V. DISCUSSION

The results of this study demonstrate that weakly supervised, class-conditioned textual explanations can be effectively generated from histopathological images while remaining aligned with a high-performing MGMT classification model. Beyond achieving strong predictive performance, the proposed framework addresses a critical limitation of many computational pathology pipelines: the lack of interpretable, clinician-friendly explanations that accompany model predictions.

A. Interpretation of Classification Performance

The proposed MobileNetV2-based architecture, augmented with shared attention and NdLinear projections, achieves consistently higher validation and cross-validation performance than commonly used backbones such as ResNet-50 and EfficientNet-B0 under comparable training conditions. This suggests that lightweight architectures, when carefully designed and fully fine-tuned, can capture diagnostically relevant histomorphological patterns without the overhead of larger models. Importantly, the improved performance is achieved with fewer parameters, which may facilitate deployment in resource-constrained clinical environments.

B. Faithfulness and Role of Visual Explanations

SmoothGrad-CAM saliency maps exhibit strong faithfulness properties as evidenced by insertion and deletion analyses. The rapid confidence changes observed when salient regions are added or removed indicate that the highlighted areas are causally linked to the classifier's decision. This is particularly relevant in histopathology, where spatial coherence and localization of diagnostically meaningful regions are essential for clinical trust.

C. Textual Rationales as a Complement to Saliency Maps

The ViT-conditioned captioning module produces concise, class-aware textual rationales that remain grounded in both the predicted MGMT status and the spatial structure of saliency maps. Unlike free-form captioning systems, the proposed approach tightly couples language generation with classifier outputs and saliency-derived spatial cues. This coupling reduces the risk of generating plausible but unfaithful explanations, a known concern in post-hoc explanation methods.

From a clinical perspective, such textual rationales may offer a more natural and accessible interface for communicating model reasoning, particularly for multidisciplinary tumor boards where not all participants are experts in image-based XAI methods.

D. Clinical Interpretability of Textual Explanation

While a formal user study with neuropathologists lies beyond the scope of this work, we designed the explanation pipeline to maximize clinical plausibility. Pseudo-captions are built from a domain-informed phrase set S_c reflecting commonly recognized histomorphological patterns, and all generated rationales are explicitly grounded in SmoothGrad-CAM saliency maps and conditioned on the predicted MGMT class. This tight coupling between visual evidence, classifier output, and language generation mitigates the risk of producing hallucinated or purely post-hoc explanations. These design choices aim to support interpretability rather than to claim direct clinical decision support.

E. Limitations and Future Directions

Several limitations should be acknowledged. First, the study relies on a single-institution dataset, which may limit direct assessment of generalization across different staining protocols, scanners, and acquisition settings. However, several design choices aim to mitigate this limitation. In particular, stain normalization using the Macenko method and extensive data augmentation (including color jittering) are employed to reduce sensitivity to staining variability. Furthermore, the use of weakly supervised learning at the patch level encourages the model to focus on robust morphological patterns rather than dataset-specific artifacts.

Second, pseudo-captions are constructed from predefined phrase sets, which, while effective in ensuring faithfulness, may limit linguistic diversity and expressiveness. Incorporating semi-automated or self-refining caption vocabularies represents a promising direction for future work, potentially enabling richer yet still faithful textual explanations.

Finally, although perturbation-based metrics provide evidence of saliency faithfulness, quantitative evaluation of textual explanation faithfulness remains an open challenge. Developing standardized benchmarks—and conducting systematic, multi-expert evaluations of explanation usefulness in real clinical workflows—are important directions for future research.

VI. CONCLUSION

In this work, we presented a weakly supervised XAI framework for predicting MGMT promoter methylation status from H&E-stained glioma histology. The proposed approach integrates a lightweight convolutional classifier with a ViT-conditioned captioning module that generates class-aware textual rationales grounded in saliency maps. By combining strong predictive performance with interpretable visual and textual explanations, the framework moves beyond black-box molecular prediction toward more transparent and clinically meaningful decision support.

Our results demonstrate that faithful spatial attributions can be effectively leveraged to guide natural language explanations without requiring region-level annotations or expert-written reports. This suggests a promising pathway for scalable, interpretable computational pathology systems that align with clinical reasoning practices. Future work will focus on broader validation, richer linguistic modeling, and user-centered evaluation of explanation utility in real-world clinical settings.

REFERENCES

[1] Weller, M., van den Bent, M., Preusser, M., Le Rhun, E., Tonn, J. C., Minniti, G., ... & Wick, W. (2021). EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. Nature reviews Clinical oncology, 18(3), 170-186.

[2] Butler, M., Pongor, L., Su, Y. T., Xi, L., Raffeld, M., Quezado, M., ... & Wu, J. (2020). MGMT status as a clinical biomarker in glioblastoma. Trends in cancer, 6(5), 380-391.

[3] Davri, A., Birbas, E., Kanavos, T., Ntritsos, G., Giannakeas, N., Tzallas, A. T., & Batistatou, A. (2022). Deep learning on histopathological images for colorectal cancer diagnosis: a systematic review. Diagnostics, 12(4), 837.

[4] Mili, M., Abdallah, A. B., Otero, J. J., Kerkeni, A., & Bedoui, M. H. (2023, October). Revolutionizing Brain Cancer Diagnosis: Automated Prediction of MGMT Methylation Status using Histological Images. In 2023 International Conference on Cyberworlds (CW) (pp. 149-156). IEEE.

[5] Mili, M., Ben Abdeljelil, A., Manzanera, A., Ben Abdallah, A., Otero, J. J. and Bedoui, M. H. (2026). Quantitatively Audited Multi-View XAI for Medical Image Classification: Application to MGMT Promoter Methylation. In Proceedings of the 18th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART; ISBN 978-989-758-796-2; ISSN 2184-433X, SciTePress, pages 1924-1935. DOI: 10.5220/0014463000004052

[6] He, Y., Duan, L., Dong, G., Chen, F., & Li, W. (2024). Computational pathology-based weakly supervised prediction model for MGMT promoter methylation status in glioblastoma. Frontiers in Neurology, 15, 1345687.

[7] Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M., & Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering, 5(6), 555-570.

[8] Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE transactions on neural networks and learning systems, 32(11), 4793-4813.

[9] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

[10] Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825.

[11] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

[12] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[13] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[14] Jing, B., Xie, P., & Xing, E. (2018, July). On the automatic generation of medical imaging reports. In Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers) (pp. 2577-2586).

[15] Chen, Z., Song, Y., Chang, T. H., & Wan, X. (2020, November). Generating radiology reports via memory-driven transformer. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) (pp. 1439-1449).

[16] Singh, A., Krishna Raguru, J., Prasad, G., Chauhan, S., Tiwari, P. K., Zaguia, A., & Ullah, M. A. (2022). [Retracted] Medical Image Captioning Using Optimized Deep Learning Model. Computational Intelligence and Neuroscience, 2022(1), 9638438.

[17] Cevik, L., Landrove, M. V., Aslan, M. T., Khammad, V., Garagorry Guerra, F. J., Cabello-Izquierdo, Y., ... & Otero, J. J. (2022). Information theory approaches to improve glioma diagnostic workflows in surgical neuropathology. Brain Pathology, 32(5), e13050.

[18] Reneau, A., Hu, J. Y. C., Zhuang, Z., and Liu, T. C. NdLinear Is All You Need for Representation Learning. arXiv preprint arXiv:2503.17353, 2025. https://arxiv.org/abs/2503.17353