DRGFuse: Alignment-Guided Dual-Layer Reliability-Gated CT–WSI Fusion for Preoperative TRG0 vs TRG1–3 Prediction in Esophageal Cancer

Zhenbing Liu[1*], Hongzhi Liu[1*], Nuo Yang[2], Dong Zeng[3], Wenwen Min[4], Hongyu Li[1], Zaiyi Liu[5†], Changmiao Wang[6†], and Haoxiang Lu[1†]

[1] Guilin University of Electronic Technology, Guilin, China
[2] Hospital of Guangxi Medical University, Nanning, China
[3] Southern Medical University, Guangzhou, China
[4] Yunnan University, Kunming, China
[5] Department of Radiology, Nanfang Hospital, Southern Medical University, Guangzhou, China
[6] Shenzhen Research Institute of Big Data, Shenzhen, China
cmwangalbert@gmail.com, liuzaiyi@gdph.org.cn, hxlu1005@163.com
[†]Corresponding authors: Changmiao Wang, Zaiyi Liu, and Haoxiang Lu.

Abstract. Achieving tumor regression grade 0 (TRG0) after neoadjuvant therapy is critical for postoperative management in esophageal cancer. Recent studies indicate that preoperative CT, biopsy whole-slide images (WSIs), and their fusion can predict pCR or treatment response, suggesting that routine pre-treatment data carries efficacy-related signals. However, refining the endpoint to the decision-critical TRG0 vs TRG1–3 classification remains challenging: TRG is defined on post-treatment whole resection specimens, whereas preoperative CT and biopsy WSIs provide only partial and localized evidence, and false TRG0 predictions may lead to under-treatment or insufficient surveillance. This mismatch, compounded by multi-center shift, makes clinically required low-FPR operating point essential yet difficult. In this paper, we introduce DRGFuse, an alignment-guided CT–WSI framework for robust preoperative TRG0 vs TRG1–3 prediction. Specifically, DRGFuse leverages automatic segmentation to derive stable CT regions and encode tumor and peritumoral context, and employs a pathology foundation model to extract transferable histologic tokens from biopsy WSIs. Moreover, mismatch-aware reliability gating with alignment constraints suppresses irrelevant cross-modal interference and strengthens low-FPR discrimination. Extensive multicenter experiments demonstrate that DRGFuse consistently outperforms unimodal models and representative multimodal baselines, while exhibiting improved robustness under cross-center shift and missing-modality settings. Code is available at: https://github.com/Perasperaasastra/DRG-fuse

[*]Equal contribution.

Keywords: Esophageal Cancer · Tumor Regression Grade Zero · Computed Tomography · Whole Slide Imaging · Multimodal Fusion.

1 Introduction

Esophageal cancer remains a major global oncologic burden, and response stratification after neoadjuvant therapy is critical for subsequent treatment decisions [20]. In locally advanced disease, *TRG0* (no residual tumor) guides surgery and surveillance, and false TRG0 predictions can be clinically costly. However, TRG is defined on post-treatment resection specimens, whereas preoperative assessment relies on partial evidence from CT and biopsy WSIs, creating an information gap. Prior work can be broadly grouped into CT-side radiomics/deep models [2] and WSI-side weakly supervised multiple-instance learning with pathology foundation representations [13]. CT–WSI multimodal fusion (e.g., late fusion or static weighting) has also been explored, with some studies reporting external-center evaluation [16]. Yet most approaches still focus on pCR or other coarse endpoints, and fusion strategies remain limited; consequently, under deployment settings that require low false-positive rates, preoperative CT+biopsy WSI prediction of TRG0 versus TRG1–3 remains unreliable, making robust TRG0 prediction under low-FPR constraints an urgent clinical need.

To meet this need, a practical preoperative TRG0 predictor must address two key challenges. First, a central challenge in distinguishing TRG0 from TRG1–3 is the mismatch between supervision and preoperative evidence: TRG summarizes whole-specimen response, whereas pre-treatment CT and biopsy WSI provide only localized views. Borderline TRG0/1 cases are particularly vulnerable to missed residual disease, making low-FPR operation difficult to sustain across centers. Moreover, overall AUC does not reflect clinical priorities; training and evaluation should emphasize performance in the low-FPR region (e.g., Sp95/FPR=0.05) and adopt noise-robust learning strategies [25,9]. Second, effective multimodal fusion should exploit complementary information while avoiding harm when modalities disagree. In practice, sampling variability and cross-center differences can make CT-derived morphology/context and WSI-derived histologic evidence inconsistent, and mainstream fusion via concatenation or fixed weights may amplify spurious correlations and destabilize low-FPR performance [1]. Therefore, a suitable fusion mechanism should support local and global alignment, regulate cross-modal interaction, and provide a principled fallback.

Motivated by these considerations, we propose DRGFuse, an alignment-guided, dual reliability-gated CT–WSI fusion framework that improves stability in the low-FPR regime through controlled interaction and explicit fallback. In multi-center experiments, DRGFuse outperforms unimodal models and common fusion baselines in both overall discrimination and low-FPR metrics, while remaining robust under center shift and missing-modality settings. Our contributions include: (1) a preoperative CT+biopsy WSI framework for TRG0 versus TRG1–3 prediction targeting low false-positive decision making; (2) an alignment-guided dual reliability-gated fusion design that reduces harmful cross-

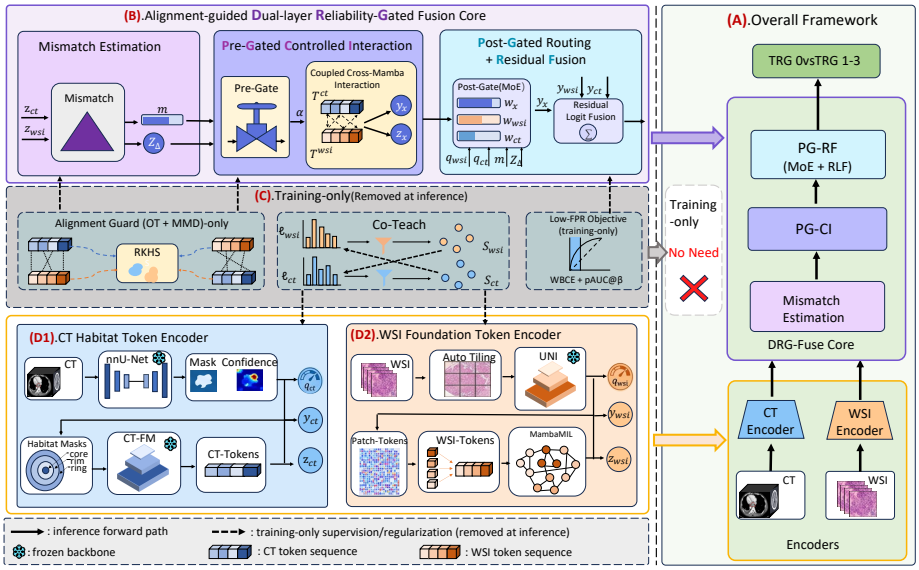

Fig. 1. DRGFuse overview. (A) Inference keeps only the CT/WSI encoders and the DRG-Fuse core. (B) Fusion follows a single reliability-driven chain: mismatch estimation → pre-gated controlled interaction → post-gated routing with residual logit fusion. (C) Training-only components (Align-Guard, Co-Teach, low-FPR objective) are disabled at inference. (D1/D2) CT and WSI encoders.

modal transfer under supervision mismatch and cross-center variation; (3) a deployment-oriented training and evaluation protocol emphasizing low-FPR performance with evidence of robust generalization.

2 Method

DRGFuse addresses the structural mismatch between *post-operative whole-specimen* TRG supervision and *pre-operative partial* evidence from 3D CT and biopsy WSI. This mismatch is compounded by multi-center distribution shifts, which can introduce label noise and spurious associations (Fig. 1). Given a CT volume X^{ct} and a biopsy WSI bag $X^{wsi} = \{x_i\}_{i=1}^{N_p}$, the CT and WSI branches produce token sequences (T^{ct}, T^{wsi}), global representations (z_{ct}, z_{wsi}), and unimodal logits (y_{ct}, y_{wsi}). After token-level interaction, the fusion core outputs a fused representation z_x and fused logit y_x. A second-stage gate predicts sample-adaptive weights $w = (w_x, w_{ct}, w_{wsi})$ with $w = \mathrm{softmax}(\cdot)$, where $w_x, w_{ct}, w_{wsi} \geq 0$ and $w_x + w_{ct} + w_{wsi} = 1$. Residual logit fusion (RLF) yields the final logit s and probability p:

$$\begin{aligned} s &= w_x\, y_x + w_{ct}\, y_{ct} + w_{wsi}\, y_{wsi}, \\ p &= \sigma(s), \end{aligned} \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid and p denotes the probability of TRG0 (positive class).

2.1 CT and WSI Token Encoders

Frozen backbones. We keep nnU-Net, the CT visual encoder (CT-FM), and the UNI pathology encoder frozen (no back-propagation). We train only the lightweight token scoring/aggregation modules, reliability gates, fusion core, and prediction heads.

CT Habitat Token Encoder. As shown in Fig. 1(D1), the CT branch first applies nnU-Net to obtain a tumor ROI and a corresponding soft mask [11]. We then define three habitat masks (core/rim/ring) over the tumor and peri-tumoral context: core is the eroded tumor interior, rim is a thin boundary band, and ring is a dilated peritumoral band for surrounding context. We extract a voxel-level feature map F using a 3D visual encoder (CT-FM). For each habitat mask M_k, we perform masked average pooling over F within the mask to obtain a CT token t_k^{ct}. All habitat tokens form $T^{ct} = \{t_k^{ct}\}_{k=1}^{K_h}$, which are lightly aggregated to produce the global representation z_{ct} and a CT-only logit y_{ct} via a binary classifier. We further construct an ROI-quality vector q_{ct} (e.g., ROI coverage and mask confidence statistics) for the post-gate to assess CT reliability.

WSI Foundation Token Encoder. The WSI branch takes a pre-treatment biopsy WSI bag $X^{wsi} = \{x_i\}_{i=1}^{N_p}$ as input (Fig. 1(D2)). We first generate a tissue mask on a low-resolution thumbnail to remove background, tile patches at the target magnification, and filter low-quality patches using no-reference quality control, producing an effective patch set. Each patch is embedded using a frozen UNI encoder to obtain semantic tokens $u_i \in \mathbb{R}^d$ [3].

Biopsy WSIs can yield large N_p with a variable fraction of noisy patches. To obtain a stable, fixed-length representation, we score patch tokens with a lightweight network and retain the Top-K_p tokens to form T^{wsi}. We then apply MambaMIL for long-sequence modeling: a linear-time sequence module (Mamba) updates token context, followed by MIL pooling to produce a slide-level representation z_{wsi} and a WSI-only logit y_{wsi} [8]. To inform the second-stage gate, we also construct a WSI quality vector q_{wsi} that combines quality-control statistics (effective patch ratio, tissue coverage, and filtered blur/background ratios) with a reliability cue derived from the concentration of the attention distribution.

2.2 DRGFuse Core: Alignment-guided Dual-layer Reliability-gated Fusion

To reduce harmful transfer when CT and WSI evidence disagree, DRGFuse explicitly estimates cross-modal mismatch and uses it as a single control signal for both token interaction (pre-gate) and final decision weighting (post-gate) (Fig. 1(B)). Mismatch estimation maps the global representations (z_{ct}, z_{wsi}) to a discrepancy vector Z_Δ and a scalar mismatch score m:

$$\begin{aligned} Z_\Delta &= \phi\Big(\big[\, z_{ct}; z_{wsi}; (z_{ct} - z_{wsi}); (z_{ct} \odot z_{wsi}) \,\big]\Big), \\ m &= \sigma\big(\psi_m(Z_\Delta)\big), \end{aligned} \tag{2}$$

where $\phi(\cdot)$ is a discrepancy embedding network (MLP/linear layers) and $\psi_m(\cdot)$ is a scoring head that outputs the forward reliability score m. The pre-gate converts mismatch into an interaction coefficient

$$\alpha = 1 - m, \tag{3}$$

which enforces a monotonic rule: stronger mismatch leads to weaker cross-modal interaction.

Conditioned on α, we couple T^{ct} and T^{wsi} to produce fused tokens and predict (z_x, y_x). The interaction uses a lightweight design: cross-modal projections $\mathrm{Proj}_{w\to c}(\cdot)$ and $\mathrm{Proj}_{c\to w}(\cdot)$ align dimensions and inject information across modalities, and then $\mathrm{Mamba}(\cdot)$ updates token context to capture long-range dependencies. This produces updated tokens $\widetilde{T}^{ct}$ and $\widetilde{T}^{wsi}$. We then concatenate $[\widetilde{T}^{ct}; \widetilde{T}^{wsi}]$, apply $LN(\cdot)$ and $Agg(\cdot)$ to obtain z_x, and use a fusion head $f_x(\cdot)$ to produce y_x. Layer normalization and residual-style connections stabilize optimization and help preserve unimodal representations during interaction.

The post-gate (Fig. 1(B), Post-Gate) follows a mixture-of-experts style router [18]: it combines $(q_{ct}, q_{wsi}, m, Z_\Delta)$ to predict routing weights (w_x, w_{ct}, w_{wsi}), and the final (s, p) are computed by Eq. (1):

$$[\, w_x, w_{ct}, w_{wsi} \,] = \mathrm{softmax}\Big(\psi_g\big([\, q_{ct}; q_{wsi}; m; Z_\Delta \,]\big)\Big), \tag{4}$$

where $\psi_g(\cdot)$ is the router network. This sample-adaptive routing down-weights unreliable experts when modality quality is low or mismatch is high, and increases reliance on the fusion expert when cross-modal evidence is consistent, providing a structured fallback behavior.

2.3 Training-only Mechanisms and Deployability

Training-time components are summarized in Fig. 1(C). Because TRG labels come from post-operative whole-specimen assessment while the inputs reflect pre-operative, localized evidence, some level of weak-supervision noise is unavoidable. We use co-teaching to mitigate label noise by exchanging small-loss samples [9]. To further reduce cross-center drift and cross-modal semantic shift, we introduce Align-Guard (OT+MMD) as a soft alignment regularizer over token representations and global embeddings, denoted as $\mathcal{L}_{align}$ [5,6]. Finally, to match high-specificity deployment goals, we emphasize performance in the low-FPR regime by combining weighted BCE with a pAUC surrogate over FPR $\leq \beta$. The overall objective is:

$$\mathcal{L} = \mathcal{L}_{WBCE} + \lambda_p \mathcal{L}_{pAUC@\beta} + \lambda_\alpha \mathcal{L}_{align} + \lambda_c \mathcal{L}_{co}, \tag{5}$$

where $\mathcal{L}_{WBCE}$ up-weights the positive class (TRG0) to address class imbalance; $\mathcal{L}_{co}$ is the co-teaching loss computed on cross-selected small-loss samples; and $\lambda_p \mathcal{L}_{pAUC@\beta}$ uses a differentiable hardest-negative surrogate to focus learning on the clinically critical low-false-positive region. At inference, Align-Guard, co-teaching, and low-FPR-specific training objectives are removed; DRGFuse retains only the two encoders and the DRGFuse core forward path, ensuring consistent deployment cost.

3 Experiments

3.1 Experimental Setup

Dataset. We conducted a retrospective study on a multi-center private esophageal cancer cohort from collaborating hospitals. Inputs include pre-treatment 3D CT and pre-treatment endoscopic biopsy WSIs (H&E), and labels are TRG 0–3 assessed on post-operative whole resection specimens. We formulate a binary endpoint, TRG0 vs. TRG1–3, with TRG0 as the positive class. The cohort consists of Internal (n=525) and two held-out external centers External A/B (n=51/62). Cases must have traceable TRG, pre-treatment CT, and pre-treatment WSI; records with missing/invalid TRG are excluded. CT and WSI are paired at the patient level via in-hospital identifiers. Due to privacy constraints, the data are not publicly released.

Implementation details. We report AUC and deployment-oriented low-false-positive metrics: pAUCn (normalized pAUC over FPR $\in [0, 0.05]$) and TPR@Sp95 (sensitivity at 95% specificity). We perform patient-level stratified five-fold cross-validation on Internal: each fold is split into train/val/test; val is used for early stopping and model selection and to determine the Sp95 threshold. This threshold is then fixed for evaluation on the corresponding Internal test split and External A/B; external centers do not participate in training or any model/threshold selection. Preprocessing is frozen during development and applied consistently in training and inference. Training uses AdamW for up to 50 epochs with batch size 4–8, optimized by weighted BCE (pos_weight) combined with a low-FPR pAUC surrogate.

3.2 Experimental Results

Comparison methods. We compare four families of representative methods: radiomics baselines (PyRadiomics+XGBoost) [7], CT-only deep models [15,12], WSI-side weakly supervised MIL/foundation-representation methods (e.g., ABMIL/CLAM/TransMIL) [10,14], and strong multimodal baselines such as eSPARK and CTF [23,4]. All methods are evaluated under the same five-fold splits and the fixed-Sp95-threshold protocol; external centers do not participate in any selection.

Table 1. Comparison of single-modality baselines and our full method on all metrics. Bold denotes the best result.

Method	Internal			External A			External B		
	AUC ↑	pAUCn ↑	TPR@Sp95 ↑	AUC ↑	pAUCn ↑	TPR@Sp95 ↑	AUC ↑	pAUCn ↑	TPR@Sp95 ↑
CT-ONLY	0.671±0.030	0.395±0.051	0.300±0.062	0.633±0.054	0.332±0.080	0.252±0.079	0.620±0.051	0.320±0.068	0.242±0.072
WSI-ONLY	0.692±0.033	0.424±0.047	0.331±0.060	0.641±0.056	0.354±0.081	0.270±0.075	0.638±0.062	0.330±0.082	0.251±0.079
DRGFuse	**0.762**±0.018	**0.602**±0.042	**0.541**±0.050	**0.720**±0.039	**0.542**±0.069	**0.456**±0.072	**0.713**±0.044	**0.525**±0.072	**0.433**±0.068

As shown in Table 1, including complete-missing-modality CT-only/WSI-only boundary tests, DRGFuse achieves the best performance across all three

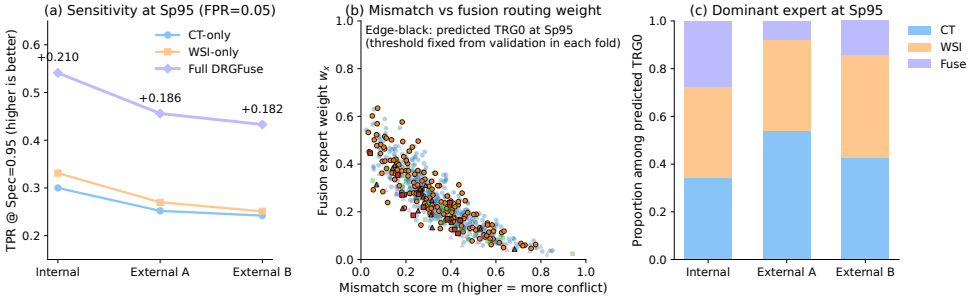

Fig. 2. Diagnostic analyses for low-FPR deployment. (a) TPR at 95% specificity (Sp95) across Internal and External A/B, with Δ indicating gains over the best unimodal model in each center. **(b)** The fusion-expert weight w_x decreases monotonically with increasing mismatch score m, including on samples predicted as TRG0 at the fixed Sp95 operating point. **(c)** Expert dominance (CT/WSI/Fuse) among samples predicted as TRG0 at Sp95, showing center-dependent routing and fallback behavior.

centers, with gains concentrated in low-FPR: Internal (AUC/pAUCn/TPR@Sp95) is 0.762/0.602/0.541, while External A/B are 0.720/0.542/0.456 and 0.713/0.525/0.433, respectively. Fig. 2**(a)** visualizes the Sp95 operating point: DRGFuse consistently attains higher TPR@Sp95 across centers, and "+Δ" highlights gains over the best unimodal model, emphasizing that improvements concentrate on low-FPR behavior rather than only overall AUC.

Table 2. Quantitative comparison with state-of-the-art methods on three datasets. Bold denotes the best result.

Method	Internal			External A			External B		
	AUC↑	pAUCn↑	TPR@Sp95↑	AUC↑	pAUCn↑	TPR@Sp95↑	AUC↑	pAUCn↑	TPR@Sp95↑
Rad+XGB	0.611±0.031	0.282±0.050	0.184±0.054	0.564±0.061	0.224±0.078	0.141±0.076	0.570±0.051	0.231±0.068	0.157±0.065
3D CNN+MG [24]	0.640±0.029	0.324±0.049	0.225±0.058	0.590±0.061	0.268±0.080	0.183±0.083	0.605±0.052	0.277±0.073	0.191±0.071
OmniCT [12]	0.661±0.021	0.353±0.052	0.257±0.065	0.618±0.051	0.290±0.083	0.215±0.088	0.620±0.049	0.303±0.071	0.220±0.070
CT-FM [15]	0.671±0.030	0.395±0.051	0.300±0.062	0.633±0.054	0.332±0.080	0.252±0.079	0.620±0.051	0.320±0.068	0.242±0.072
VR+VM3D [22]	0.690±0.035	0.393±0.044	0.312±0.052	0.661±0.042	0.327±0.082	0.249±0.081	0.658±0.043	0.324±0.077	0.244±0.065
ABMIL [10]	0.621±0.039	0.303±0.058	0.201±0.073	0.560±0.072	0.234±0.089	0.156±0.091	0.552±0.071	0.221±0.090	0.263±0.089
CLAM [14]	0.654±0.041	0.341±0.061	0.242±0.070	0.599±0.071	0.274±0.088	0.191±0.092	0.580±0.071	0.265±0.087	0.189±0.093
TransMIL [17]	0.660±0.028	0.350±0.051	0.253±0.062	0.605±0.059	0.284±0.077	0.201±0.081	0.598±0.061	0.274±0.082	0.198±0.080
PathFM [3]	0.692±0.033	0.424±0.047	0.331±0.060	0.641±0.056	0.354±0.081	0.270±0.075	0.638±0.062	0.330±0.082	0.251±0.079
ASMIL [21]	0.695±0.030	0.395±0.049	0.300±0.063	0.635±0.062	0.325±0.078	0.245±0.082	0.620±0.059	0.312±0.078	0.235±0.082
Late Fusion	0.701±0.028	0.462±0.053	0.391±0.066	0.645±0.052	0.362±0.084	0.293±0.077	0.641±0.052	0.352±0.083	0.288±0.080
GMU [1]	0.711±0.030	0.483±0.051	0.410±0.059	0.670±0.049	0.383±0.081	0.311±0.082	0.669±0.048	0.373±0.084	0.307±0.085
SMuRF [19]	0.720±0.029	0.503±0.055	0.436±0.060	0.683±0.050	0.401±0.082	0.335±0.085	0.690±0.055	0.391±0.081	0.320±0.082
eSPARK [23]	0.733±0.033	0.512±0.052	0.441±0.058	0.709±0.052	0.412±0.080	0.341±0.083	0.707±0.051	0.402±0.080	0.336±0.084
CTF [4]	0.742±0.025	0.526±0.055	0.454±0.061	0.705±0.048	0.423±0.082	0.346±0.080	0.708±0.050	0.412±0.082	0.344±0.082
DRGFuse	**0.762**±0.018	**0.602**±0.042	**0.541**±0.050	**0.720**±0.039	**0.542**±0.062	**0.456**±0.072	**0.713**±0.044	**0.525**±0.072	**0.433**±0.068

More importantly, as shown in Table 2, sustained gains in the low-FPR regime on external centers depend strongly on fusion strategy. Although CTF and eSPARK are competitive on Internal, their low-FPR metrics tend to saturate on External A/B, suggesting that under evidence disagreement and center shift, reliability-unaware fusion can amplify spurious correlations and limit high-specificity usability. In contrast, DRGFuse improves over CTF on External A/B

by +0.119/+0.113 in pAUCn and +0.110/+0.089 in TPR@Sp95, indicating that low-FPR external generalization hinges on mismatch-driven controlled interaction and conflict-aware fallback rather than a stronger backbone alone.

Table 3. Ablation study of key components in DRGFuse. CoTeach: mismatch-aware co-teaching; LowFPR: low-FPR objective (WBCE + pAUC surrogate, FPR≤0.05); Align: OT+MMD alignment regularizer (Align-Guard); SimpleBackbone: replace the DRG-Fuse core with a simple fusion backbone (concat/single interaction); ConventionalFusion: replace dual-gate reliability + conflict-aware fallback with feature concatenation + MLP classifier. Bold denotes the best result.

Method	Internal			External A			External B		
	AUC↑	pAUCn↑	TPR@Sp95↑	AUC↑	pAUCn↑	TPR@Sp95↑	AUC↑	pAUCn↑	TPR@Sp95↑
w/o CoTeach	0.751±0.023	0.541±0.046	0.464±0.058	0.712±0.042	0.422±0.081	0.311±0.082	0.708±0.044	0.405±0.080	0.297±0.079
w/o LowFPR	0.755±0.025	0.507±0.047	0.425±0.060	0.710±0.039	0.427±0.082	0.313±0.083	0.703±0.043	0.406±0.083	0.294±0.085
w/o Align	0.746±0.026	0.542±0.052	0.471±0.065	0.698±0.045	0.440±0.087	0.348±0.088	0.687±0.043	0.425±0.084	0.325±0.084
SimpleBackbone	0.739±0.025	0.551±0.055	0.480±0.062	0.687±0.040	0.476±0.082	0.373±0.084	0.692±0.041	0.456±0.082	0.351±0.085
ConventionalFusion	0.743±0.032	0.536±0.057	0.458±0.065	0.685±0.046	0.431±0.088	0.302±0.090	0.681±0.046	0.419±0.087	0.280±0.091
DRGFuse	**0.762**±0.018	**0.602**±0.042	**0.541**±0.050	**0.720**±0.039	**0.542**±0.069	**0.456**±0.072	**0.713**±0.044	**0.525**±0.072	**0.433**±0.068

Ablation studies. To explain why gains are most prominent in low-FPR and more evident on external centers, we ablate key components (Table 3). Removing the low-FPR objective, mismatch-aware co-teaching, or OT+MMD alignment degrades pAUCn/TPR@Sp95, with larger drops on external centers. Replacing dual-gate reliability and fallback with a conventional fusion causes the largest degradation (External A/B TPR@Sp95 drops to 0.302/0.280, i.e., $-0.154/-0.153$), confirming conflict-aware fallback as the main driver of low-FPR robustness.

Fig. 2**(b)** and Fig. 2**(c)** further support this mechanism: as m increases, w_x decreases and routing among CT/WSI/Fuse shifts in a center-dependent manner at Sp95, consistent with adaptive fallback under shift. Therefore, the gains mainly come from the low-FPR objective and reliability-gated fallback, which together stabilize decisions under cross-center shift and modality disagreement.

4 Conclusion

We propose DRGFuse for robust preoperative TRG0 versus TRG1–3 classification from CT and biopsy WSI. It uses mismatch-aware dual-layer reliability gating to regulate cross-modal interaction and conflict-aware fallback. Multicenter internal and external evaluations show more consistent generalization than strong unimodal and representative multimodal baselines, with ablations and mechanism audits supporting the critical role of reliability gating. Limitations include information loss from biopsy sampling and dependence on upstream imaging/pathology quality. Future work will validate DRGFuse in larger prospective cohorts and improve missing-modality robustness, uncertainty calibration, and computational efficiency for safer deployment.

Ethics Approval. Approved by the Ethics Committee of Guangdong Provincial People's Hospital (No. KY2025-1117-01).

Acknowledgments. Supported by the Science Foundation of China (No. 82502451, No. 62462017, No. 82272075), Guangxi Key R&D Project (No. 24010086, No. AB24010167), Guangxi Science and Technology Program (No. FN2504240022), and Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011617).

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

References

1. Arévalo, J., Solorio, T., Montes-y Gómez, M., González, F.A.: Gated multimodal units for information fusion. In: arXiv (2017), arXiv:1702.01992
2. Chen, D., Wang, W., Li, Q., Rao, X., Shi, X., Lu, D., Diao, D., Zhai, J., Cai, K.: Preoperative CT-based radiomics for predicting response to neoadjuvant chemoimmunotherapy in esophageal squamous cell carcinoma. Radiology: Imaging Cancer **8**(1), e250128 (2026)
3. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F.K., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A.H., Shaban, M., Williams, M., Vaidya, A., Sahai, S., Oldenburg, L., Weishaupt, L.L., Wang, J.J., Williams, W., Le, L.P., Gerber, G., Mahmood, F.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**, 850–862 (2024)
4. Chen, Y., Huang, Y., Wang, F., Yeung, M., Jiang, Y., Wang, S., Yu, L.: Bridging radiology and pathology foundation models via concept-based multimodal co-adaptation. In: International Conference on Learning Representations (ICLR) (2026), poster
5. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 26, pp. 2292–2300 (2013)
6. Gretton, A., Borgwardt, K.M., Rasch, M., Schölkopf, B., Smola, A.: A kernel two-sample test. Journal of Machine Learning Research **13**, 723–773 (2012)
7. van Griethuysen, J.J.M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G.H., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.W.L.: Computational radiomics system to decode the radiographic phenotype. Cancer Research **77**(21), E104–E107 (2017)
8. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: International Conference on Learning Representations (ICLR) (2024), arXiv:2312.00752
9. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 8527–8537 (2018)
10. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. In: Proceedings of the 35th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 80, pp. 2127–2136. PMLR (2018)
11. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**, 203–211 (2021)
12. Lin, T., Qiu, Z., Zhang, W., Liu, J., Xie, Y., Gao, M., Fan, Z., Li, Z., Li, S., Xie, Z., Lu, P., Zhuang, Y., Xia, Y., Zhang, L., Ooi, B.C.: OmniCT: Towards a unified slice-volume LVLM for comprehensive CT analysis. arXiv (2026), arXiv:2602.16110

13. Liu, Z., Zhao, L., Zhao, Z., Qin, W., Zhong, J., Lin, F., Lu, M., Guo, R., Guo, Q., Xu, H., Li, S., Zheng, H., Lu, H.: Automated tumor regression grade assessment and survival prediction in esophageal cancer via weakly supervised multiple instance learning. Frontiers in Medicine **13**, 1751768 (2026)
14. Lu, M.Y., Williamson, D.F.K., 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**, 555–570 (2021)
15. Pai, S., Hadzic, I., Bontempi, D., Bressem, K., Kann, B.H., Fedorov, A., Mak, R.H., Aerts, H.J.W.L.: Vision foundation models for Computed Tomography. arXiv (2025), arXiv:2501.09001
16. Qi, B., Jiang, Z., Shen, H., Li, J., Wang, Z., Fang, M., Wang, C., Jiang, Y., Yuan, J., Bermejo, I., Dekker, A., De Ruysscher, D., Wee, L., Zhang, W., Ji, Y., Zhang, Z.: Interpretable multimodal radiopathomics model predicting pathological complete response to neoadjuvant chemoimmunotherapy in esophageal squamous cell carcinoma. Journal for ImmunoTherapy of Cancer **13**, e013840 (2025)
17. Shao, Z., Bian, H., Chen, Y., Wang, J., Zhang, X., Ji, X.: TransMIL: Transformer-based correlated multiple instance learning. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 34, pp. 2136–2147. Curran Associates, Inc. (2021)
18. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: International Conference on Learning Representations (ICLR). pp. 1–19 (2017)
19. Song, B., Leroy, A., Yang, K., Dam, T., Wang, X., Maurya, H., Pathak, T., Lee, J., Stock, S., Li, X.T., Fu, P., Lu, C., Toro, P., Chute, D.J., Koyfman, S., Saba, N.F., Patel, M.R., Madabhushi, A.: Deep learning informed multimodal fusion of radiology and pathology to predict outcomes in hpv-associated oropharyngeal squamous cell carcinoma. eBioMedicine **114**, 105663 (2025)
20. Yang, H., Wang, F., Hallemeier, C.L., Lerut, T., Fu, J.: Oesophageal cancer. The Lancet **404**(10466), 1991–2005 (2024)
21. Ye, L., Hamidi, S.M., Chi, Z., Li, G., Pilanci, M., Ogawa, T., Haseyama, M., Plataniotis, K.N.: ASMIL: Attention-stabilized multiple instance learning for Whole-Slide imaging. In: International Conference on Learning Representations (ICLR). pp. 1–39 (2026), poster
22. Zhang, Z., Luo, T., Yan, M., et al.: Voxel-level radiomics and deep learning for predicting pathologic complete response in esophageal squamous cell carcinoma after neoadjuvant immunotherapy and chemotherapy. Journal for ImmunoTherapy of Cancer **13**, e011149 (2025)
23. Zhao, Z., Chen, D., Wei, X., Li, S., Zhang, X., Lin, W., Zheng, X., Zheng, K., Wu, S., Wen, X., Zhang, B., Zheng, Y., Chen, S., Xie, C., Li, S., Xie, D., Wang, R., Xing, W., Zhou, J., Cai, M.: A multimodal synergistic model for personalized neoadjuvant immunochemotherapy in esophageal cancer. Cell Reports Medicine **6**(12), 102479 (2025)
24. Zhou, Z., Sodha, V., Rahman Siddiquee, M.M., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J.: Models genesis: Generic autodidactic models for 3d medical image analysis. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Lecture Notes in Computer Science, vol. 11767, pp. 384–393. Springer (2019)
25. Zhu, D., Li, G., Wang, B., Wu, X., Yang, T.: When AUC meets DRO: Optimizing partial AUC for deep learning with non-convex convergence guarantee. In: Proceedings of the 39th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 162, pp. 27548–27573. PMLR (2022)