Advanced IntelliEngineering

Vol. 1, No. 2, p. 52-59, 2026

Original article

CTF-UMTL: Cross-modal token fusion and uncertainty-aware multi-task learning framework for survival prediction

Wenwen Yue*

College of Information Engineering, Hubei University of Chinese Medicine, Wuhan 430065, P. R. China

Keywords:
Multimodal survival prediction
uncertainty-aware learning
multi-task learning

Cited as:
Yue, W. CTF-UMTL: Cross-modal token fusion and uncertainty-aware multi-task learning framework for survival prediction. Advanced IntelliEngineering, 2026, 1(2): 52-59.
https://doi.org/10.46690/aie.2026.02.02

Abstract:
Multimodal medical data provide complementary information for survival prediction, yet their effective integration remains challenging due to unaligned multi-modality feature representations, incomplete modality availability, and predictive uncertainty. We propose a cross-modal token fusion and uncertainty-aware multi-task learning framework for survival prediction, termed CTF-UMTL. Modality token encoding unifies multi-modality magnetic resonance imagin, whole-slide image, clinical-genomic representations through modality-aware tokenization. Robust multimodal transformer with gated fusion simulates missing modalities through random modality masking, and adaptively fuses available modalities while suppressing missing branches via availability-aware gating. Uncertainty-aware multi-task learning jointly acquires the OS logit, Cox risk score, and grade, while simultaneously outputting their means and log-variances. Experimental results on the real-world MMIST-Lung dataset demonstrate that CTF-UMTL achieves a BAcc of 0.7953, an AUC of 0.8614, an F1-score of 0.9436, a sensitivity of 0.8305, and a specificity of 0.7600, outperforming the best baseline, MCAT, by 1.92%, 1.94%, 7.23%, 2.04%, and 1.78%, respectively. Ablation studies further demonstrate the different contributions of the proposed components. Our framework provides a promising paradigm of reliable survival prediction, constructing a synergy model of unified modality tokenization, missing-modality-robust fusion, and uncertainty-aware multi-task learning.

1. Introduction

Multimodal medical data fusion is a cornerstone of survival prediction. Multimodal medical data, including magnetic resonance imaging (MRI), whole-slide images (WSIs), clinical records, and genomic profiles, provide information for survival prediction. Compared with conventional unimodal approaches, multimodal learning enables the integration of multimodal information and captures comprehensive patient representations for improved prognostic assessment. Consequently, various machine learning and deep learning methods have been developed to exploit cross-modal relationships and enhance cancer survival prediction[1].

Recently, Transformer-based multimodal learning has attracted considerable attention in survival prediction due to its ability to capture complex dependencies across multiple medical modalities. Several studies have explored cross-modal attention, token-based representations, and adaptive fusion strategies to improve the effectiveness of prognostic modeling. SurvPath introduces biological pathway tokens and histology patch tokens to model fine-grained interactions between transcriptomic and pathological information[2]. To address incomplete multimodal observations, MUST leverages modality-specific representation decomposition and diffusion-based reconstruction to improve the effectiveness of survival prediction under missing modalities[3]. Furthermore, MADSurv investigates uncertainty-aware multimodal survival analysis, highlighting the importance of accounting for predictive un-

Yandy Scientific Press

*Corresponding author.
E-mail address: wyue6405@stmail.hbucm.edu.cn (W. Yue).

3135-3118 © The Author(s) 2026.
Received July 16, 2026; revised August 8, 2026; accepted August 21, 2026; available online August 23, 2026.

certainty in prognostic modeling[4]. These studies demonstrate the potential of multimodal representation learning, robust multimodal fusion, and uncertainty estimation for survival prediction[5].

Multimodal transformers and cross-attention mechanisms partially alleviate multi-modality fusion but rely on large-scale data and do not explicitly handle missing modalities. Some studies introduce uncertainty estimation, but mostly in unimodal settings or simple mean-variance regression, without comprehensive integration with multimodal fusion[6, 7]. Therefore, unifying multiple modalities while achieving missing-modality robustness and prediction confidence remains an open problem. Overall, three challenges remain important: Unifying diverse medical modalities within a consistent token representation, maintaining reliable cross-modal fusion when modalities are unavailable, and integrating survival-related predictions with uncertainty quantification for confidence-aware risk assessment.

To address these limitations, we propose a CTF-UMTL framework, a cross-modal token fusion and uncertainty-aware multi-task learning framework for survival prediction. The contributions of the framework are as follows:

1) We propose a modality token encoding, which integrates imaging, WSI, and clinical–genomic data into a shared token vocabulary through modality-specific tokenizers and modality-type embeddings, addressing the difficulty of unifying multi-modality modalities.
2) A robust multimodal transformer was designed, which combines modality-aware attention and random modality masking for cross-modal representation learning and availability-aware gating for adaptive evidence aggregation under missing modalities.
3) An uncertainty-aware multi-task learning is proposed, which jointly models survival risk and grade while producing mean and variance estimates; stochastic inference further converts these estimates into a risk distribution for risk–uncertainty stratification to address the predictive uncertainty.

The subsequent sections are organized as follows: Section 2 presents related work on multimodal data fusion. Section 3 introduces the architecture and steps of the CTF-UMTL model. Section 4 describes the experimental design, comparative experiments, ablation experiments, and result analyses. Section 5 gives the conclusions and future work.

2. Related work

This section presents multimodal medical data fusion, missing-modality robust learning, and uncertainty-aware multi-task learning for survival prediction.

2.1 Multimodal medical data fusion

Multimodal fusion is widely used for survival prediction, where MRI, WSI, and clinical-genomic data are integrated to provide comprehensive patient profiles. These methods typically aim to construct unified representations that encode both intra-modal features and cross-modal correlations. MCAT shifts the prediction paradigm from unimodal analysis to a global perspective, capturing image-genomic interactions via a cross-attention module[8]. SMuRF integrates pathological and radiological features through a self-attention pooling mechanism to improve prognostic accuracy[9]. However, these methods primarily focus on architectural innovation and lack explicit modeling of modality types, making it difficult to distinguish multi-modality tokens in a shared space, thereby failing to systematically unify the multiple modalities[10].

2.2 Missing-modality robust learning

Recent studies mitigate missing modalities in survival prediction primarily through feature reconstruction and adaptive fusion. Generative approaches, such as LD-CVAE[11, 12], reconstruct missing representations from available features but introduce modeling complexity and rely heavily on reconstruction fidelity. Alternatively, fusion-based frameworks employ intermediate fusion[13, 14] or adaptive gating[15–17] to handle incomplete observations, yet they require careful coordination between cross-modal interaction and modality availability. Consequently, learning informative cross-modal representations while explicitly controlling the contributions of available modalities remains a critical direction for robust survival prediction.

2.3 Uncertainty-aware and multi-task survival prediction

Recent advances in deep survival prediction have increasingly explored uncertainty estimation and multi-task learning. For instance, Wu et al. [18] jointly predicted molecular alterations, histological grade, and prognosis from MRI. Liu et al. [19] presented MM-DLS, which combined positron emission tomography/computed tomography and clinical features for simultaneous subtyping, staging, and risk assessment. More recently, NADMSurv integrated patient-level noise quantification and uncertainty-aware evidence fusion to enhance robustness[20]. These studies highlight the benefits of multi-task supervision and uncertainty modeling; A unified framework that jointly characterizes survival probability and relative risk, while explicitly decomposing predictive uncertainty for risk-uncertainty stratification, explores a significant perspective for comprehensive prognosis.

In summary, unifying heterogeneous medical modalities with distinct feature distributions and semantic characteristics into a coherent representation, preserving effective cross-modal interactions and predictive stability despite incomplete modality availability, and jointly modeling survival outcomes and clinical attributes while quantifying predictive uncertainty to support reliable risk assessment. To solve these challenges, we propose CTF-UMTL, which integrates unified modal tokenization, availability-aware cross-modal fusion, and uncertainty-aware multi-task learning within a framework.

3. Methodology

We propose a cross-modal token fusion and uncertainty-aware multi-task learning framework for survival prediction, termed CTF-UMTL. It integrates three modules: A modality

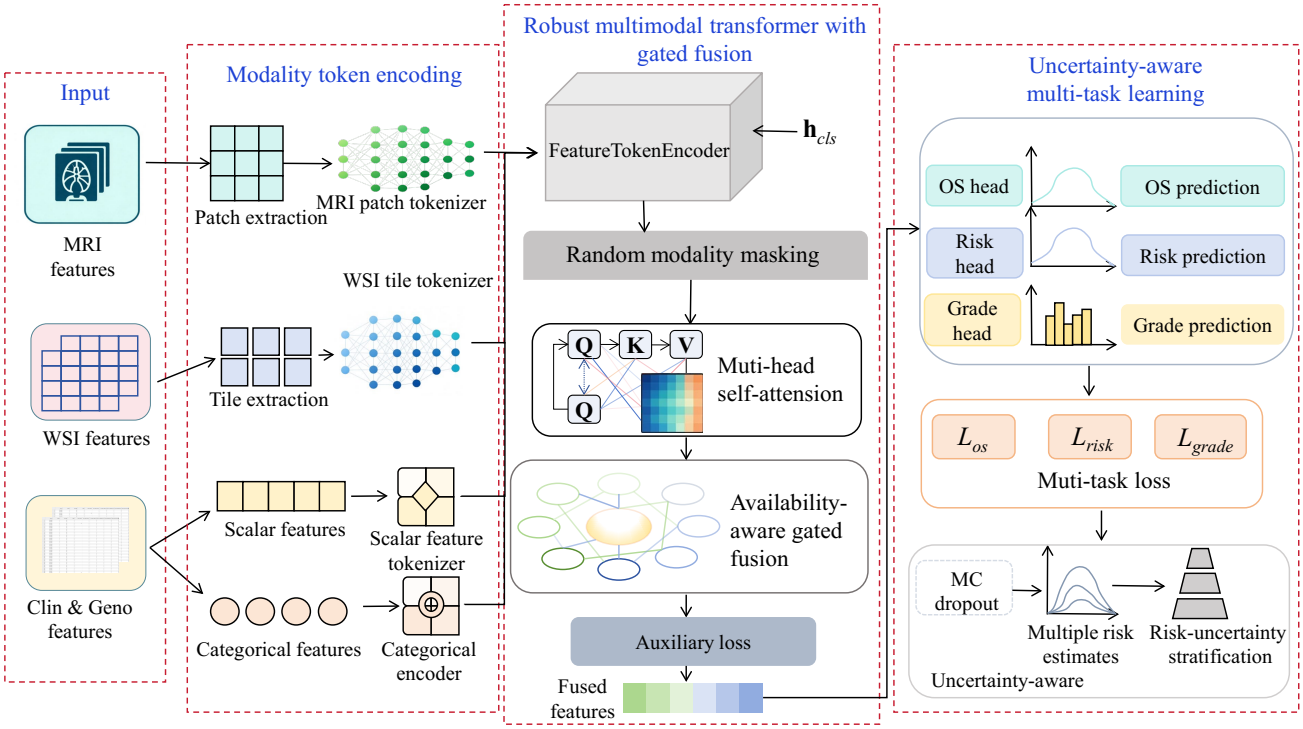

Fig. 1. The architecture of CTF-UMTL.

token encoding, a robust multimodal transformer with gated fusion, and an uncertainty-aware multi-task learning. Modality token encoding maps MRI, WSI, and clinical-genomic features into a unified token space, where MRI/WSI are encoded independently. The robust multimodal transformer with gated fusion applies multi-head self-attention to capture cross-modal dependencies and uses random modality masking to address modality absence. The uncertainty-aware multi-task learning mitigates prediction uncertainty by jointly modeling risk and grade with heteroscedastic loss and deriving risk distributions. The architecture of CTF-UMTL is shown in Fig. 1.

3.1 Modality token encoding

To integrate multi-modality data into a unified representation, we develop a modality token encoding module that transforms MRI, WSI, and clinical-genomic features into a shared token embedding space. Let $\mathrm{M} = \{\mathrm{MRI},\ \mathrm{WSI},\ \mathrm{Clin},\ \mathrm{Geno}\}$ denote the set of modalities, and let d be the common token embedding dimension.

For each modality $m \in \mathrm{M}$, a modality-specific tokenizer T_m maps its input features into a sequence of d-dimensional tokens. The related representation $\mathbf{H}_m$ is defined as:

$$\mathbf{H}_m = T_m(\mathbf{X}_m) \in \mathbb{R}^{N_m \times d} \tag{1}$$

where $\mathbf{X}_m$ denotes the input features of modality m, and N_m is the number of tokens generated for this modality.

For MRI and WSI, each image is represented by a set of patch-level feature vectors. Specifically, the i-th feature vector of modality $m \in \{\mathrm{MRI},\ \mathrm{WSI}\}$ is denoted as $\mathbf{x}_{m,i} \in \mathbb{R}^{D_m}$, where D_m is the modality-specific feature dimension. A learnable projection maps each feature vector into the shared token space:

$$\mathbf{h}_{m,i} = \mathbf{W}_m \mathbf{x}_{m,i} + \mathbf{b}_m \tag{2}$$

where $\mathbf{h}_{m,i}$ is the new feature vector after mapping, $\mathbf{W}_m \in \mathbb{R}^{d \times D_m}$ denotes weight matrix, and $\mathbf{b}_m \in \mathbb{R}^d$ denotes the bias vector. The projection parameters are not shared between MRI and WSI.

For clinical and genomic data, numerical and categorical features are encoded separately to preserve their individual characteristics. Given a numerical feature $\mathbf{x}_{m,k} \in \mathbb{R}$, its token representation $\mathbf{h}_{m,k}^{num}$ is obtained as:

$$\mathbf{h}_{m,k}^{num} = f_{m,k}(\mathbf{x}_{m,k}) + \mathbf{e}_{m,k} \tag{3}$$

where $f_{m,k} : \mathbb{R} \to \mathbb{R}^d$ is a learnable feature encoder and $\mathbf{e}_{m,k} \in \mathbb{R}^d$ is a learnable feature-identity embedding. The encoder is implemented using a multilayer perceptron.

For a categorical feature $\mathbf{x}_{m,k}$, the corresponding token $\mathbf{h}_{m,k}^{cat}$ is constructed as:

$$\mathbf{h}_{m,k}^{cat} = \mathbf{E}_{m,k}[\mathbf{x}_{m,k}] + \mathbf{e}_{m,k} \tag{4}$$

where $\mathbf{E}_{m,k}$ represents the learnable embeddings associated with the k-th categorical feature. The feature-identity embedding distinguishes different clinical or genomic variables. Numerical and categorical tokens are subsequently assembled into their respective modality-specific token sequences.

To explicitly encode modality type, a learnable modality-type embedding $\mathbf{a}_m \in \mathbb{R}^d$ is added to each token:

$$\widetilde{\mathbf{h}}_{m,i} = \mathbf{h}_{m,i} + \mathbf{a}_m \tag{5}$$

where $\widetilde{\mathbf{h}}_{m,i}$ represents the token feature vector. The modality-type embedding is shared among tokens belonging to the same

modality and jointly optimized with the network parameters.

Finally, a learnable classification token $\mathbf{h}_{cls} \in \mathbb{R}^d$ is prepended to the unified token sequence:

$$\mathbf{Z} = \left[\mathbf{h}_{cls}; \widetilde{\mathbf{H}}_{MRI}; \widetilde{\mathbf{H}}_{WSI}; \widetilde{\mathbf{H}}_{Clin}; \widetilde{\mathbf{H}}_{Geno}\right] \tag{6}$$

where $\widetilde{\mathbf{H}}_m$ ($m \in \mathrm{M}$) denotes the token sequence after adding modality-type embeddings.

To adjust variable-length sequences within a mini-batch, zero-padding is applied to obtain $\mathbf{Z} \in \mathbb{R}^{B \times L_{\max} \times d}$, where B is the batch size and $L_{\max}$ is the maximum sequence length in the batch. A corresponding padding mask matrix $\mathbf{P}$, where $P_{b,i} \in \{0,1\}^{B \times L_{\max}}$ is constructed, with the fill position $P_{b,i} = 1$ indicating a padded position and $P_{b,i} = 0$ indicating a valid token. The unified token representations and padding masks are then passed to the next multimodal transformer.

3.2 Robust multimodal transformer with gated fusion

Random masking is applied to MRI and WSI to simulate incomplete multimodal observations. Let $r_m \in \{0,1\}$ indicate whether modality $m \in \mathrm{M}$ is available or not, and let a_m denote its availability after random masking. For each available modality $m \in \mathrm{M}_{nc} = \{\mathrm{MRI},\ \mathrm{WSI}\}$, a retention variable u_m is sampled from a Bernoulli distribution, and the availability is defined as:

$$u_m \sim \mathrm{Bernoulli}(1-p), \qquad a_m = \begin{cases} r_m u_m, & m \in \mathrm{M}_{nc} \\ r_m, & m \in \mathrm{M} \setminus \mathrm{M}_{nc} \end{cases} \tag{7}$$

where p is the masking probability. When at least one non-clinical modality is available, the sampled mask is constrained by $\sum_{m \in \mathrm{M}_{nc}} a_m \geq 1$, ensuring that at least one observed non-clinical modality is retained. During inference, random masking is disabled and $a_m = r_m$. The unified token sequence $\mathbf{Z}$ is subsequently processed by a Transformer encoder with N_T layers. To exclude padded and unavailable modality tokens, the padding mask matrix $\mathbf{P}$ is combined with modality availability to construct a token validity indicator v_i, where $v_i = 1$ for the global aggregation token and $v_i = a_{m(i)}(1 - P_{b,i})$ for other tokens, with $m(i)$ denoting the modality associated with position i. For attention head j at layer ℓ, the masked self-attention output $\mathbf{O}_j^{(\ell)}$ is formulated as:

$$\mathbf{O}_j^{(\ell)} = \mathrm{softmax}\left(\frac{\mathbf{Q}_j^{(\ell)} (\mathbf{K}_j^{(\ell)})^\top}{\sqrt{d_k}} + \mathbf{A}\right) \mathbf{V}_j^{(\ell)} \tag{8}$$

where $\mathbf{Q}_j^{(\ell)}$, $\mathbf{K}_j^{(\ell)}$, and $\mathbf{V}_j^{(\ell)}$ are linear projections of the preceding hidden representations, and d_k is the attention head dimension. The additive attention mask satisfies the attention mask matrix $\mathbf{A}$, where $A_{ij} = 0$ when $v_j = 1$ and $A_{ij} = -\infty$ otherwise, preventing padded or unavailable tokens from contributing as attention keys and values. After N_T Transformer layers, the global aggregation token generates a representation $\mathbf{h}_g \in \mathbb{R}^d$, while masked mean pooling over valid tokens of each modality generates its summary representation $\mathbf{s}_m \in \mathbb{R}^d$. For an unavailable modality, $\mathbf{s}_m$ is set to zero. To adaptively aggregate modality-specific evidence, an availability-aware gating mechanism computes the normalized modality weights as:

$$g_m = \frac{a_m \exp(\mathbf{w}_m^\top \mathbf{h}_g + b_m)}{\sum_{m' \in \mathrm{M}} a_{m'} \exp(\mathbf{w}_{m'}^\top \mathbf{h}_g + b_{m'})} \tag{9}$$

where g_m denotes normalized modality weight, m' represents all modalities, $\mathbf{w}_m \in \mathbb{R}^d$ and $b_m \in \mathbb{R}$ are learnable gating parameters. Assuming at least one modality is available, the gating weights sum to one, and unavailable modalities receive zero weight. The final fused representation is obtained by combining global contextual information with the weighted modality summaries:

$$\mathbf{f}_{fused} = \mathbf{h}_g + \sum_{m \in \mathrm{M}} g_m \mathbf{s}_m \tag{10}$$

To further regularize multimodal representation learning, we introduce cross-modal alignment loss and fusion consistency loss. Let $\mathbf{A}_b = \{m \in \mathrm{M} : a_{b,m} = 1\}$ denote the available modality set for patient b, and let $\mathbf{C}_b = \{(m,n) : m,n \in \mathbf{A}_b,\ m < n\}$ denote the set of available modality pairs under a fixed modality order. The alignment loss L_{align} encourages consistent representations across modalities from the same patient:

$$L_{align} = \frac{\sum_{b=1}^{B} \sum_{(m,n) \in \mathbf{C}_b} \left[1 - \cos(\mathbf{s}_{b,m}, \mathbf{s}_{b,n})\right]}{\max\left(1, \sum_{b=1}^{B} |\mathbf{C}_b|\right)} \tag{11}$$

where $\cos(\cdot)$ denotes cosine similarity. Additionally, the fusion consistency loss L_{cons} encourages the fused representation to preserve information shared with available modality representations:

$$L_{cons} = \frac{\sum_{b=1}^{B} \sum_{m \in \mathbf{A}_b} \left[1 - \cos(\mathbf{s}_{b,m}, \mathbf{f}_{fused,b})\right]}{\max\left(1, \sum_{b=1}^{B} |\mathbf{A}_b|\right)} \tag{12}$$

The overall loss L_{aux} is defined as:

$$L_{aux} = \lambda_{align} L_{align} + \lambda_{cons} L_{cons} \tag{13}$$

where λ_{align} and λ_{cons} are non-negative coefficients controlling the contributions of the two objectives.

3.3 Uncertainty-aware multi-task learning

To jointly model patient prognosis and predictive uncertainty, we develop an uncertainty-aware multi-task learning module that integrates overall survival (OS) prediction, Cox proportional hazards modeling, and grade classification. Given the fused patient representation $\mathbf{f}_{fused} \in \mathbb{R}^d$, separate prediction heads estimate the mean and log-variance of the OS (μ_{os} and s_{os}) and Cox risk (μ_{risk} and s_{risk}) outputs:

$$\begin{aligned} \mu_{os} &= \mathbf{w}_{os}^\top \mathbf{f}_{fused} + b_{os} \\ s_{os} &= \mathbf{v}_{os}^\top \mathbf{f}_{fused} + c_{os} \\ \mu_{risk} &= \mathbf{w}_{risk}^\top \mathbf{f}_{fused} + b_{risk} \\ s_{risk} &= \mathbf{v}_{risk}^\top \mathbf{f}_{fused} + c_{risk} \end{aligned} \tag{14}$$

where the weight vectors $\mathbf{w}_{os}^{\top}$, $\mathbf{w}_{risk}^{\top}$, $\mathbf{v}_{os}^{\top}$, and $\mathbf{v}_{risk}^{\top}$ belong to $\mathbb{R}^d$, b_{os}, b_{risk}, c_{os}, and c_{risk} are the bias terms, and $s_{os} = \log \sigma_{os}^2$ and $s_{risk} = \log \sigma_{risk}^2$ denote the predicted log-variances. The OS predicts the probability of surviving beyond a predefined time horizon τ, whereas the Cox estimates a relative log-risk score associated with the hazard function. To model observation-dependent uncertainty, the OS logit is represented by a Gaussian distribution with mean μ_{os} and variance σ_{os}^2. Using the reparameterization trick, stochastic logits and their corresponding survival probabilities are obtained as:

$$z_{os}^{(k)} = \mu_{os} + \sigma_{os}\varepsilon^{(k)}, \qquad p_{os} \approx \frac{1}{K}\sum_{k=1}^{K} \text{Sigmoid}\left(z_{os}^{(k)}\right) \tag{15}$$

where random variable $\varepsilon^{(k)} \sim N(0,1)$ and K is the number of Monte Carlo samples used to approximate the predictive survival probability. Let $y_i^{os} \in \{0,1\}$ indicate whether patient i survives beyond τ, and let $q_i \in \{0,1\}$ indicate whether this outcome can be determined from the observed follow-up. The uncertainty-aware OS objective is defined as:

$$L_{os} = -\frac{1}{\max(1, \sum_i q_i)} \sum_i q_i \left[y_i^{os} \log p_{os,i} + (1 - y_i^{os}) \log(1 - p_{os,i})\right] \tag{16}$$

For time-to-event modeling, the Cox predicts a Gaussian distribution over the patient-specific log-risk score. A stochastic risk sample is generated as:

$$r_i^{(k)} = \mu_{risk,i} + \sigma_{risk,i}\varepsilon_i^{(k)}, \qquad \varepsilon_i^{(k)} \sim N(0,1) \tag{17}$$

The Cox partial likelihood is optimized by averaging over stochastic risk samples:

$$L_{cox} = -\frac{1}{K_c \max(1, \sum_i \delta_i)} \sum_{k=1}^{K_c} \sum_{i:\delta_i=1} \left[r_i^{(k)} - \log \sum_{j:t_j \geq t_i} \exp\left(r_j^{(k)}\right)\right] \tag{18}$$

where t_i is the observed follow-up time, $\delta_i \in \{0,1\}$ is the event indicator, K_c is the number of stochastic risk samples, and the risk sets are constructed using patients with observed survival information. For grade prediction, an additional classification head estimates the probability of the grade class:

$$p_{grade,i} = \text{Sigmoid}\left(\mathbf{w}_{grade}^{\top}\mathbf{f}_{fused,i} + b_{grade}\right) \tag{19}$$

where $\mathbf{w}_{grade} \in \mathbb{R}^d$ and b_{grade} are learnable parameters. The grade classification objective is defined as:

$$L_{grade} = -\frac{1}{\max(1, \sum_i z_i)} \sum_i z_i \Big[y_i^{grade} \log p_{grade,i} + (1 - y_i^{grade}) \log(1 - p_{grade,i})\Big] \tag{20}$$

where $y_i^{grade} \in \{0,1\}$ is the binary grade label and $z_i \in \{0,1\}$ indicates whether the grade annotation is available. The overall training objective jointly combines survival prediction, risk modeling, grade classification, and the auxiliary representation learning objectives:

$$L_{total} = L_{os} + \lambda_{cox}L_{cox} + \lambda_{grade}L_{grade} + \lambda_{align}L_{align} + \lambda_{cons}L_{cons} \tag{21}$$

where the λ coefficients control the contributions of the corresponding objectives. To further characterize predictive uncertainty at inference, Monte Carlo dropout is employed to obtain multiple stochastic predictions. Specifically, T time forward passes produce risk means $\mu_{risk}^{(t)}$ and variances $\sigma_{risk}^{2(t)}$. The expected log-risk score is estimated as:

$$\bar{r} = \frac{1}{T}\sum_{t=1}^{T} \mu_{risk}^{(t)} \tag{22}$$

The total predictive uncertainty is decomposed into aleatoric and epistemic components according to the law of total variance:

$$U_{total} = \underbrace{\frac{1}{T}\sum_{t=1}^{T} \sigma_{risk}^{2(t)}}_{U_{alea}} + \underbrace{\frac{1}{T}\sum_{t=1}^{T} \left(\mu_{risk}^{(t)} - \bar{r}\right)^2}_{U_{epis}} \tag{23}$$

where U_{alea} captures the predicted observation-dependent variability, while U_{epis} reflects variability across stochastic model predictions.

4. Experiments

4.1 Experimental settings

We validated our proposed framework on the MMIST-Lung dataset, a real-world, multi-center, multimodal lung cancer dataset comprising 1,365 patients from the cancer genome atlas (TCGA) and the clinical proteomic tumour analysis consortium (CPTAC). Specifically, 1,026 patients were extracted from TCGA-LUAD and TCGA-LUSC, while 339 patients were extracted from CPTAC-LUAD and CPTAC-LSCC, covering lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC/LSCC). The dataset integrates multimodality clinical, genomic, WSI, and MRI information. In this study, we utilized magnetic MRI, WSIs, and clinical-genomic features for multimodal survival prediction. For the 12-month overall survival (OS) prediction task, 1,169 patients (85.6%) were labeled as survivors and 196 (14.4%) as non-survivors.

All experiments were conducted on a system with 16.0 GB of memory and a 3.60 GHz Intel(R) Core(TM) i7-9700K CPU. The proposed CTF-UMTL framework was optimized using AdamW with an initial learning rate of 1×10^{-4}, a weight decay of 1×10^{-4}, and a mini-batch size of 4. The multimodal Transformer employed a unified token embedding dimension of 256, comprising 4 encoder layers with 8 attention heads per layer. A dropout rate of 0.2 was applied for regularization, while random modality masking was performed with a probability of 0.2 to improve robustness against incomplete observations. During inference, predictive uncertainty was estimated through 30 stochastic forward passes using Monte Carlo dropout. The auxiliary cross-modal alignment and fusion consistency losses were incorporated into the overall training objective, each weighted by a coefficient of 0.01.

4.2 Comparison experiments

To verify the effectiveness of our framework, we compared the proposed CTF-UMTL framework with three multi-modal baselines: MCAT, MMIF, and LUNAI-FCT, on five evaluation

Table 1. Performance comparison of different models.

Model	BAcc	AUC	F1-score	Sensitivity	Specificity
CTF-UMTL	**0.7953 ± 0.0252**	**0.8614 ± 0.0462**	**0.9436 ± 0.0130**	**0.8305 ± 0.0404**	**0.7600 ± 0.0365**
MCAT	0.7803 ± 0.0481	0.8450 ± 0.0512	0.8800 ± 0.0225	0.8139 ± 0.0436	0.7467 ± 0.1193
MMIF	0.6733 ± 0.0718	0.6805 ± 0.0906	0.8718 ± 0.0418	0.8267 ± 0.0832	0.5200 ± 0.1789
LUNAI-FCT	0.7933 ± 0.0440	0.8231 ± 0.0623	0.8871 ± 0.0422	0.8266 ± 0.0835	0.7600 ± 0.1535

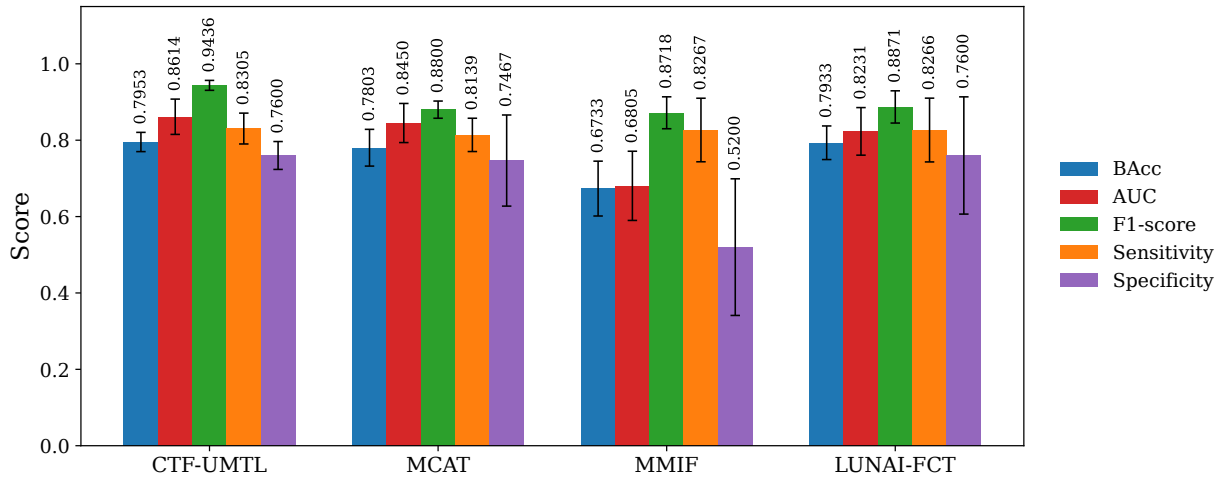

Fig. 2. Performance comparison with three baseline models across different evaluation metrics.

metrics: BAcc, AUC, F1-score, Sensitivity, and Specificity. All baselines were trained and evaluated under identical 5-fold cross-validation settings and using the same feature extraction strategy. The experimental results are reported in Table 1 and Fig. 2.

The results show that CTF-UMTL achieved the best results in all evaluation metrics. Its BAcc, AUC, F1-score, Sensitivity, and Specificity were 0.7953, 0.8614, 0.9436, 0.8305, and 0.7600, respectively. Compared with the strongest baseline MCAT, CTF-UMTL improved approximately 1.92%, 1.94%, 7.23%, 2.04%, and 1.78% in these five metrics, respectively. Notably, the substantial improvement in AUC highlights our framework's distinct advantage in distinguishing between positive and negative samples. Furthermore, the highest F1-score and sensitivity values show that our method maintains high consistency and robustness across varying random seeds, outperforming other baselines in terms of training stability.

4.3 Ablation studies

To evaluate the different contributions of each module in our framework, we conducted ablation studies by sequentially removing specific modules from the full model. The results are shown in Table 2 and Fig. 3.

The ablation studies demonstrated the contributions of three modules to overall performance. Specifically, removing the multimodal transformer fusion module induced the most severe performance degradation, with BAcc dropping from 0.7953 to 0.5124 and Specificity plummeting from 0.7600 to 0.0267, indicating its vital role in capturing deep cross-modal interactions. Removing the modality token encoder led to a substantial decrease in BAcc to 0.6747 and Specificity to 0.3733, suggesting its indispensability for feature alignment and effective discrimination of multi-modality modalities. Furthermore, eliminating the uncertainty-aware multi-task head decreased the F1-score to 0.8187, AUC to 0.7771, and BAcc to 0.7484, indicating its significant gain in enhancing model generalization and predictive confidence.

5. Conclusions

We present a cross-modal token fusion and uncertainty-aware multi-task learning framework for survival prediction, named CTF-UMTL. The modality token encoding employs modality-specific tokenizers and modality-type embeddings to transform multi-modality MRI, WSI, clinical-genomic features into a unified representation space. A robust multimodal transformer with gated fusion integrates cross-modal self-attention, random modality masking, and availability-aware gated fusion to capture complementary information while adapting to incomplete modality observations. An uncertainty-aware multi-task learning module jointly models overall survival, Cox risk, and grade, incorporating probabilistic prediction and stochastic inference to quantify predictive uncertainty.

Experimental results on the MMIST-Lung dataset demonstrate that the proposed framework outperforms the baselines

Table 2. Ablation study results of CTF-UMTL.

Model	BAcc	AUC	F1-score	Sensitivity	Specificity
Full model	**0.7953 ± 0.0252**	**0.8614 ± 0.0462**	**0.9436 ± 0.0130**	**0.8305 ± 0.0404**	**0.7600 ± 0.0365**
w/o modality token encoder	0.6747 ± 0.0917	0.8632 ± 0.0557	0.8463 ± 0.0065	0.7760 ± 0.0330	0.3733 ± 0.2140
w/o multimodal transformer fusion	0.5124 ± 0.0278	0.7568 ± 0.0638	0.8361 ± 0.0017	0.7982 ± 0.0041	0.0267 ± 0.0596
w/o uncertainty-aware multi-task	0.7484 ± 0.0780	0.7771 ± 0.0406	0.8187 ± 0.0170	0.7969 ± 0.0308	0.6000 ± 0.1633

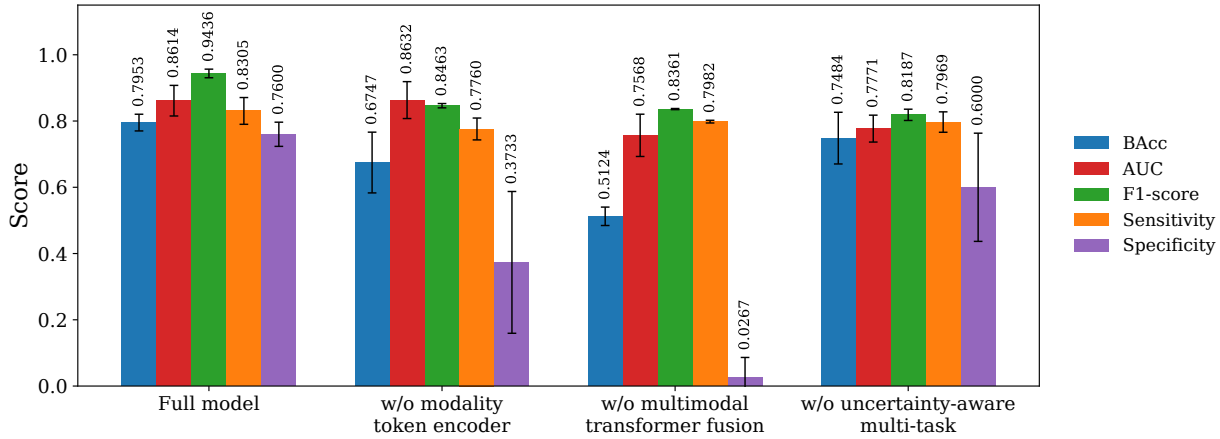

Fig. 3. Performance comparison of different ablation variants.

in BAcc, AUC, F1-score, sensitivity, and specificity. Ablation experiments further confirm the significant contributions of the modality-type embeddings, random modality masking, and gated fusion modules. These findings highlight the importance of unified tokenization and uncertainty modeling in integrating multi-modality medical data. Overall, the framework provides a structured diagnostic solution, improving the accuracy of survival prediction and providing an effective approach for handling multi-modality medical data.

Future research will focus on evaluating the CTF-UMTL framework in larger-scale and diverse clinical data. At the same time, this framework will be extended to a wider range of medical prediction tasks.

Acknowledgements

We acknowledge the funding support from the Hubei Provincial International Science and Technology Cooperation Project (No. 2025EHA011).

Conflicts of interest

The author declares no competing interest.

Open Access This article is distributed under the terms and conditions of the Creative Commons Attribution (CC BY-NC-ND) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

References

[1] Teoh, J. R., Dong, J., Zuo, X., et al. Advancing healthcare through multimodal data fusion: A comprehensive review of techniques and applications. PeerJ Computer Science, 2024, 10: e2298.

[2] Jaume, G., Vaidya, A., Chen, R. J., et al. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. Paper Presented at the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 16-22 June, 2024.

[3] Kim, K., Hwang, D. MUST: Modality-specific representation-aware transformer for diffusion- enhanced survival prediction with missing modality. Paper Presented at the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Denver, Colorado, 2-6 June, 2026.

[4] Zhang, E., Raigir, V. S., Poellabauer, C., et al. MADSurv: An uncertainty-aware framework for multimodal cancer survival analysis. Paper Presented at the 16th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Philadelphia, USA, 12-15 October, 2025.

[5] Liu, J., Cen, X., Yi, C., et al. Challenges in AI-driven biomedical multimodal data fusion and analysis. Genomics, proteomics & bioinformatics, 2025, 23(1): qzaf011.

[6] Wu, C., Zhu, Y. Large language models for intelligent healthcare and medicine: A systematic survey. Advanced IntelliEngineering, 2026, 1(1): 21-44.

[7] Shaik, T., Tao, X., Li, L., et al. A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom. Information Fusion, 2024, 102: 102040.

[8] Mohamed, M., Koubeisi, M. W., Aljuaid, F., et al. Causalevidentialfuse: A unified framework for trustworthy multimodal clinical data fusion in healthcare diagnosis. Computer Methods and Programs in Biomedicine, 2026, 288: 109647.

[9] Al-Zoghby, A. M., Ebada, A. I., Saleh, A. S., et al. A comprehensive review of multimodal deep learning for enhanced medical diagnostics. Computers, Materials & Continua, 2025, 84(3): 4155-4193.

[10] Jiao, T., Guo, C., Feng, X., et al. A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications. Computers, Materials & Continua, 2024, 80(1): 1-35.

[11] Li, H., Wu, X. CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach. Information Fusion, 2024, 103: 102147.

[12] Zhou, J., Tang, J., Zuo, Y., et al. Robust multimodal survival prediction with conditional latent differentiation variational autoencoder. Paper Presented at the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 10-17 June, 2025.

[13] Lin, R., Hu, H. MissModal: Increasing robustness to missing modality in multimodal sentiment analysis. Transactions of the Association for Computational Linguistics, 2023, 11: 1686-1702.

[14] Ruffini, F., Caruso, C. M., Tacconi, C., et al. Handling missing modalities in multimodal survival prediction for non-small cell lung cancer. npj Digital Medicine, 2026, 9: 776.

[15] Li, S., Chen, C., Han, J. SimMLM: A simple framework for multi-modal learning with missing modality. Paper Presented at the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, USA, 19-20 October, 2025.

[16] Maheshwari, H., Liu, Y., Kira, Z. Missing modality robustness in semi-supervised multi-modal semantic segmentation. Paper Presented at the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, Hawaii, 4-8 January, 2024.

[17] Song, Y., Zhan, Z., Xie, F., et al. A cancer-type-aware framework for robust multimodal survival prediction under missing modalities. Briefings in Bioinformatics, 2026, 27(2): bbag124.

[18] Wu, X., Zhang, S., Zhang, Z., et al. Biologically interpretable multi-task deep learning pipeline predicts molecular alterations, grade, and prognosis in glioma patients. npj Precision Oncology, 2024, 8(1): 181.

[19] Liu, X., Dai, F., Dai, J., et al. Multimodal multitask deep learning for grading management system in non-small cell lung cancer. Nature Communications, 2026, 17: 7579.

[20] Ming, W., Ye, W., Liu, M., et al. Noise-aware importance-uncertainty disentangled multimodal learning for robust cancer survival prediction. Paper Presented at the International Conference on Medical Image Computing and Computer-Assisted Intervention, Strasbourg, France, 27 September-1 October, 2026.