Selective Cross-modal Feature Enhancement Through Collaborative Learning

Mohammed Belghadi [1,*], Younès Bennani [1], and Valérie Paradis [2]

[1] LIPN-CNRS, UMR 7030-LaMSN, Université Sorbonne Paris Nord, Paris, France
[2] INSERM, UMR 1149-CNRS EMR 8252, Université Paris Cité, Paris, France
Email: mohammed.belghadi@sorbonne-paris-nord.fr (M.B.); younes.bennani@sorbonne-paris-nord.fr (Y.B.); valerie.paradis@aphp.fr (V.P.)
*Corresponding author

***Abstract*—Multimodal survival prediction is based on integrating heterogeneous data sources—including radiology, pathology, genomics, and clinical attributes—that are often incomplete in real clinical workflows. Most existing approaches depend on late fusion, which limits cross-modal interaction and degrades under missing modalities. We introduce Selective Cross-Modal Learning (SCML), a lightweight pre-fusion refinement framework that enriches unimodal embeddings by selectively incorporating complementary cues from concurrently available modalities. SCML operates in three stages: (1) projection-space alignment to enable meaningful cross-modal similarity measurement with-out corrupting task-discriminative representations; (2) selective cross-modal attention guided by pairwise synergy analysis and global modality-importance weighting, ensuring that each modality is enhanced only by empirically supportive counterparts; and (3) a reconstruction constraint that preserves modality-specific structure and improves robustness to incomplete data. We evaluated SCML in a cohort of 962 patients with glioma using 15-fold Monte Carlo cross-validation, achieving a median C-index of 0.8147—surpassing strong baselines including Mean-Vector fusion, Pathomic Fusion, and Multimodal learning with Missing Data (MMD). SCML maintains its advantage under missing-pathology and missing-genomics conditions. The framework is encoder-agnostic, modality-agnostic, and readily applicable to other multimodal prediction tasks that involve partially missing and heterogeneous inputs.**

***Keywords*—multimodal learning, cross-modal attention, missing modalities, feature refinement, representation learning**

I. Introduction

Multimodal learning has emerged as a powerful approach for integrating heterogeneous information sources, enabling richer representations and more reliable predictions [1]. This is especially important in complex medical settings, where each modality captures a different aspect of patient biology. In healthcare, radiological imaging, histology, molecular profiling, and patient demographics provide complementary perspectives on disease characterization [2–5]. Leveraging such multimodal information is therefore essential for improving diagnostic and prognostic accuracy, as well as supporting treatment planning [6].

Glioblastoma (GBM), the most aggressive primary brain tumor, illustrates both the challenges and the opportunities of multimodal learning. Clinical decision-making relies on diverse examinations, from imaging scans to biopsies, each offering partial yet complementary insights [7]. However, existing deep learning methods often face three major limitations: (i) they rely heavily on late fusion, combining representations only at the final stage and therefore missing fine-grained cross-modal dependencies; (ii) they are sensitive to missing or incomplete modalities, which are frequent in real clinical workflows; and (iii) they lack mechanisms to promote stable and consistent representations across heterogeneous sources.

The central idea of this work is that intermediate enhancement of unimodal embeddings can allow informative cross-modal cues (e.g., subtle radio-genomic patterns) to be captured before fusion, leading to more robust and clinically meaningful representations than late fusion alone. To this end, a selective refinement framework is introduced to enhance unimodal features through cross-modal attention prior to fusion. A projection alignment layer enables heterogeneous modalities to interact in a shared space, while a reconstruction-based regularization helps preserve modality-specific characteristics and improves robustness under incomplete inputs. Although the framework is evaluated on glioma prognosis, its design principles are intended to remain applicable to other multimodal healthcare settings involving heterogeneous and partially missing data.

The main contributions of this paper are:

- Selective refinement module: an attention-based mechanism that dynamically enriches unimodal features with complementary cues from other modalities before fusion, going beyond conventional late-fusion pipelines.

Manuscript received January 9, 2026; revised February 12, 2026; accepted May 25, 2026; published September 4, 2026.

- Agnostic and generalizable framework: a modality-encoder-independent design, validated on glioma prognosis and readily applicable to a broad range of multimodal learning tasks involving heterogeneous encoders.
- HVS-guided cross-modal weighting: an interpretable mechanism that quantifies and balances contributions across modalities, limiting dominance by strong modalities while highlighting underrepresented sources such as demographics and omics.
- Projection-space InfoNCE alignment: a mutual-information objective applied in a projection space to promote cross-modal similarity without degrading the discriminative capacity of the task-oriented embeddings.

II. Literature Review

Initial studies explored individual modalities for glioma prognosis. Radiological imaging was analyzed using either hand-crafted features [8] or Convolutional Neural Networks (CNNs) to learn high-level representations [9]. In parallel, genomic studies identified molecular signatures associated with patient outcomes [10], while histopathology research used deep learning on H&E-stained slides for risk stratification [11]. More recent work has shifted toward multimodal integration. Mobadersany *et al.* [12] combined genomic biomarkers with pathology features, Chen *et al.* [13] introduced pathomic fusion for genomics and histopathology, and Braman *et al.* [14] further incorporated radiology and clinical variables. Cui *et al.* [15] extended this line of work by explicitly addressing missing modalities in multimodal prognosis modeling. Recent multimodal studies have further explored glioma survival prediction. M4Survive [16] proposed a foundation-model-based pipeline combining radiology and pathology through a Mamba adapter, but required complete paired data. Yuan *et al.* [17] integrated MRI, histology, cfDNA, and clinical data using squeeze-and-excitation encoders, but still relied on late fusion and assumed that all modalities were avail-able. More broadly, recent cross-modal representation learning methods have explored prompt-based interaction strategies in vision-language and biomedical vision-language models, including Hierarchical Cross-modal Prompt Learning, VaMP, and vMFCoOp [18–20]. Although these methods further demonstrate the value of cross-modal alignment, they operate on large pretrained multimodal backbones that assume a shared embedding space across modalities, target classification or retrieval objectives with complete paired inputs, and provide no missing-modality mechanism. By contrast, SCML addresses multimodal survival prediction from independently encoded heterogeneous medical embeddings, operates under arbitrary modality incompleteness, and introduces a selective enhancement strategy grounded in empirical synergy analysis—a setting and design that cannot be covered by prompt adaptation techniques for vision-language models.

Because all modalities describe the same patient, they can in principle share complementary information. Yet most existing pipelines extract unimodal embeddings and fuse them only afterward, without allowing intermediate cross-modal inter-action. To address this limitation, the proposed framework introduces an intermediate stage of selective cross-modal attention, in which each modality is refined by integrating only the most useful signals from the others. Built on top of missing-modality-aware learning, this design remains robust under incomplete data while better exploiting shared patient context when multiple modalities are available.

III. Proposed Framework

This section defines the methodologies and procedures employed in the construction and training of our deep learning pipeline. The framework is structured into three phases. (i) Modality-specific feature extraction: unimodal embeddings are generated based on the characteristics of each modality (radiology, pathology, genomics, demographics), ensuring that their representations reside in a uniform latent space. (ii) Selective Cross-Modal Learning (SCML): each embedding is dynamically optimized by focusing exclusively on complementary information from alternative modalities. In contrast to traditional attention mechanisms, SCML presents two innovations: Heuristic Variable Selection (HVS) for balancing and interpreting modality contributions, and mutual-information alignment (InfoNCE) to maintain unimodal discriminative strength while promoting significant cross-modal synergy. (iii) Multimodal fusion: the augmented embeddings are subsequently consolidated by a mean-fusion approach, yielding a joint representation that remains stable even when some modalities are unavailable (see Fig. 1).

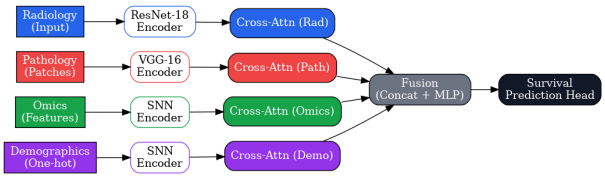

Fig. 1. Overview of the proposed multimodal deep learning framework, showing modality-specific feature extraction, cross-modal attention-based enhancement, fusion, and survival prediction.

A. Modality-Specific Feature Extraction

To establish a common feature space across heterogeneous inputs, we follow the modality-specific extraction strategies introduced in prior work [15], which demonstrated that robust survival models can be built from independently encoded representations with modality dropout. Our approach is con-structed to be modality-representation agnostic, irrespective of the encoder employed, the proposed Selective Cross-Modal Learning (SCML) block can enhance and augment embed-dings with further information. This approach facilitates direct comparison with [15], while emphasizing the enhanced benefit of intermediate cross-modal refinement.

Radiology: A pre-trained segmentation model nnUNet [21] was used to automatically identify and

segment each 3D MRI modality separately (Gd-T1w, T1w, T2w, and T2w-FLAIR) to obtain 2D slices containing the largest part of the tumor. Afterward, a 120×120 bounding box was applied to crop these 2D slices, which were stacked to form a multichannel input as illustrated in Fig. 2. This input was fed to a pre-trained ResNet-18 model to extract features. Since ResNet-18 is designed to work with 3-channel inputs, it was modified to accept 4 channels (one for each MRI modality), and the weights from the first channel were copied to the fourth channel for weight initialization. Furthermore, 361 handcrafted features derived from 3D tumor volumes by PyRadiomics [22] were injected into the ResNet-18 flatten layer to complement the deep features with the handcrafted ones. Then, all the features were projected into the 32-dimensional space using fully connected ResNet-18 layers. This approach has proven its efficiency in previous works [10, 15].

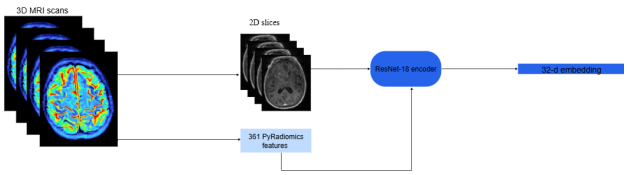

Fig. 2. Radiology branch: feature extraction from 3D MRI scans using a ResNet-18 encoder and handcrafted features.

Pathology: We used a pretrained VGG-16 model to extract features from digitized pathology slides—specifically Hematoxylin and Eosin (H&E)-stained images. Each image was 1024×1024 pixels, from which a 512×512 patch was cropped to isolate the tumor region. During testing, a sliding window of size 512×512 with a stride of 256 generated nine patches per image. The features extracted from these patches were then aggregated via averaging to form the final image-level representation as illustrated in Fig. 3. Following the best practices of previous studies [13–15], we fine-tuned only the final layers of the VGG-16 model, keeping the initial convolutional layers frozen with ImageNet-pretrained weights.

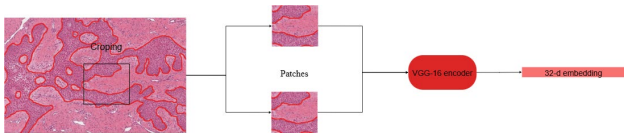

Fig. 3. Pathology branch: image patches extracted from whole-slide images are encoded with a VGG-16 network and projected into a 32-d embedding.

Genomics: Following prior work, the genomic data consisted of 80 DNA-based features—79 of which were the most informative Copy Number Variations (CNVs), along with a binary indicator of IDH1 mutation status, a critical factor in glioma prognosis that provides insight into tumor behavior and development as illustrated in Fig. 4. A Self-normalized Neural Network (SNN) [23] was utilized to efficiently encode these high-dimensional features while minimizing the risk of overfitting.

Demographics: Clinical variables included four attributes: age, race, ethnicity, and gender. These categorical variables were expanded into nine one-hot encoded features. A Self-normalizing Neural Network (SNN), as used for the genomic modality, was employed to encode these vectors into a com-pact representation suitable for integration with the other modalities as shown in as Fig. 5.

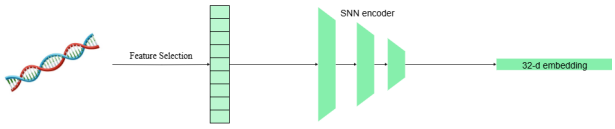

Fig. 4. Omics branch: genomic features processed by an SNN into a 32-D embedding.

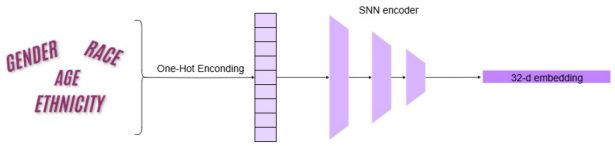

Fig. 5. Demographics branch: one-hot encoded attributes processed by an SNN into a 32-D embedding.

Algorithm 1. Feature Embedding for All Patients and Modalities

Require: Patients $\left\{\left\{m_i^{(p)}\right\}_{i=1}^{M}\right\}_{p=1}^{P}$; pretrained encoders $\{\mathcal{F}_i\}_{i=1}^{M}$; embedding dimension d

Ensure: Feature embeddings $\left\{\left\{\boldsymbol{z}_i^{(p)}\right\}_{i=1}^{M}\right\}_{p=1}^{P}$

for $p \leftarrow 1$ **to** P **do**
  for $i \leftarrow 1$ **to** M **do**
    if $m_i^{(p)} \neq \emptyset$ **then**
      $\boldsymbol{z}_i^{(p)} \leftarrow \mathcal{F}_i\left(m_i^{(p)}\right) \in \mathbb{R}^d$
    else
      $\boldsymbol{z}_i^{(p)} \leftarrow \mathbf{0}_d$ ▷zero vector for missing modality
    end if
  end for
end for

We denote by P the total number of patients, and by M the total number of modalities. Each patient $p \in 1, \ldots, P$ is represented as a set of modality-specific inputs $\left\{m_i^{(p)}\right\}_{i=1}^{M}$, where $m_i^{(p)}$ is the raw data of modality i for that patient. If a modality is missing, we use an availability mask $a_i^{(p)} \in \{0,1\}$ to indicate its presence $(a_i^{(p)} = 1)$ or absence $(a_i^{(p)} = 0)$.

B. Cross-Modal Feature Enhancement

1) Projection for cross-modal similarity

For each patient p, only the available modalities are projected into the similarity space. Each embedding $\boldsymbol{z}_i^{(p)} \in \mathbb{R}^d$ is mapped by a lightweight projection head $\mathcal{G}_i$:

$$\boldsymbol{q}_i^{(p)} = \begin{cases} \mathcal{G}_i\left(\boldsymbol{z}_i^{(p)}\right), & \text{if modality } i \text{ exists for patient } p, \\ \mathbf{0}_d, & \text{otherwise.} \end{cases}$$

Missing modalities are assigned zero vectors only as structural placeholders to maintain consistent tensor shapes, but they are excluded from all cross-modal computations. This guarantees that similarities and attention are computed only among the modalities that exist for each patient.

2) Cross-modal attention

For each patient p, attention weights are computed only between modalities that are simultaneously available. Let $\mathcal{A}_p$ denote the set of modalities present for patient p. The attention coefficient from modality i to modality j is defined as:

$$\alpha_{i,j}^{(p)} = \begin{cases} 0, & \text{if } a_j^{(p)} = 0, \\ \dfrac{\exp\left(\frac{\langle q_i^{(p)}, q_j^{(p)} \rangle}{\tau \|q_i^{(p)}\| \|q_j^{(p)}\|}\right)}{\sum_{u \in \mathcal{A}_p \setminus \{i\}} \exp\left(\frac{\langle q_i^{(p)}, q_u^{(p)} \rangle}{\tau \|q_i^{(p)}\| \|q_u^{(p)}\|}\right)}, & \text{otherwise.} \end{cases} \quad (1)$$

3) Context vector with HVS weighting

To quantify the relative contribution of each modality in prediction, we apply the Heuristic Variable Selection (HVS) principle [24, 25]. HVS decomposes the importance of a neuron into partial and relative components.

For a neuron n in the embedding $\boldsymbol{z}_i^{(p)} \in \mathbb{R}^d$, the relative contribution of n is given by:

$$S_n = \sum_{u \in \text{fan-out}(n)} \pi_{nu}\, \delta_u, \quad (2)$$

where $\text{fan-out}(n)$ denotes the set of units directly connected to n.

The partial contribution π_{nu} measures the relative share of neuron n among all inputs to unit u:

$$\pi_{nu} = \frac{|w_{nu}|}{\sum_{k \in \text{fan-in}(u)} |w_{ku}|}, \quad (3)$$

where w_{nu} is the connection weight from n to u, and $\text{fan-in}(u)$ is the set of units projecting into u.

The relative contribution δ_u is propagated backward through the network as

$$\delta_u = \begin{cases} 1, & \text{if } u \text{ belongs to the output layer (survival head),} \\ S_u, & \text{if } u \text{ is a hidden unit.} \end{cases} \quad (4)$$

Hence, S_n represents the overall HVS score of neuron n, which we denote as $HVS(n) = S_n$ for clarity. At the modality level, let $z_{i,n}^{(p)}$ denote the n-th neuron of the embedding $\boldsymbol{z}_i^{(p)}$. The raw importance of modality i for patient p is obtained by summing the HVS scores of its neurons:

$$H_i^{(p)} = \sum_{n=1}^{d} \text{HVS}\left(z_{i,n}^{(p)}\right). \quad (5)$$

The per-patient modality weights are first normalized across available modalities:

$$\beta_i^{(p)} = \frac{H_i^{(p)}}{\sum_{j \in \mathcal{A}_p} H_j^{(p)}}, \sum_{i \in \mathcal{A}_p} \beta_i^{(p)} = 1. \quad (6)$$

To obtain a global modality importance independent of individual patients, we compute the mean weight across all patients:

$$\beta_i = \frac{1}{P} \sum_{p=1}^{P} \beta_i^{(p)}. \quad (7)$$

Finally, during cross-modal enhancement, the context vector for each available modality is computed as:

$$c_i^{(p)} = \beta_i \sum_{j \in \mathcal{A}_p \setminus \{i\}} \alpha_{i,j}^{(p)}\, \boldsymbol{z}_j^{(p)}. \quad (8)$$

Unlike conventional attention layers, this formulation decouples the global importance of each modality, captured by the fixed coefficients β_i, from the patient-specific interactions modeled by $\alpha_{i,j}^{(p)}$. By masking missing modalities in both α and β, the framework prevents dominant modalities from over-whelming weaker but complementary ones, ensuring robust and interpretable cross-modal collaboration.

4) Enhanced representation

The enhanced embedding of modality i for patient p is obtained by concatenating the original and context representations:

$$\boldsymbol{z}_i^{+(p)} = \left[\boldsymbol{z}_i^{(p)};\, c_i^{(p)}\right]. \quad (9)$$

Algorithm 2. Collaborative Multimodal Feature Enhancement

Input: $\left\{\boldsymbol{z}_i^{(p)} \in \mathbb{R}^d\right\}_{i=1\ldots M,\, p=1\ldots P}$, fixed $\{\beta_i\}_{i=1\ldots M}$
Output: $\left\{\boldsymbol{z}_i^{+(p)}\right\}_{i=1\ldots M,\, p=1\ldots P}$
for $p = 1$ **to** P **do**
  $\mathcal{A}_p \leftarrow \left\{i \mid m_i^{(p)} \text{ exists}\right\}$
  for $i = 1$ **to** M **do**
    if $i \in \mathcal{A}_p$ **then**
      $\boldsymbol{q}_i^{(p)} \leftarrow \mathcal{G}_i\left(\boldsymbol{z}_i^{(p)}\right)$
    else
      $\boldsymbol{q}_i^{(p)} \leftarrow \mathbf{0}_d$
    end if
  end for
  for $i \in \mathcal{A}_p$ **do**
    for $j = 1$ **to** M **do**
      if $j \in \mathcal{A}_p \setminus \{i\}$ **then**
        $\alpha_{i,j}^{(p)} \leftarrow \dfrac{\exp\left(\frac{\langle q_i^{(p)}, q_j^{(p)} \rangle}{\tau \|q_i^{(p)}\| \|q_j^{(p)}\|}\right)}{\sum_{u \in \mathcal{A}_p \setminus \{i\}} \exp\left(\frac{\langle q_i^{(p)}, q_u^{(p)} \rangle}{\tau \|q_i^{(p)}\| \|q_u^{(p)}\|}\right)}$
      else
        $\alpha_{i,j}^{(p)} \leftarrow 0$
      end if
    end for
    $\boldsymbol{c}_i^{(p)} \leftarrow \beta_i \sum_{j \in \mathcal{A}_p \setminus \{i\}} \alpha_{i,j}^{(p)}\, \boldsymbol{z}_j^{(p)}$
    $\boldsymbol{z}_i^{+(p)} \leftarrow \left[\boldsymbol{z}_i^{(p)};\, \boldsymbol{c}_i^{(p)}\right]$
  end for
end for

C. Multimodal Fusion

The enhanced modality embeddings $\boldsymbol{z}_i^{+(p)}$ are combined into a single patient-level representation through a masked averaging scheme:

$$\boldsymbol{z}_{\text{fused}}^{(p)} = \frac{1}{\sum_{i=1}^{M} a_i^{(p)}} \sum_{i=1}^{M} a_i^{(p)} \boldsymbol{z}_i^{+(p)}, \tag{10}$$

where $a_i^{(p)} \in \{0,1\}$ indicates the presence of modality i. This operation fuses only available modalities, keeping the feature dimension fixed and ensuring consistent aggregation under missing data.

D. Loss Functions

1) Cox survival loss

Let $r^{(p)} = h_\phi\left(\boldsymbol{z}_{\text{fused}}^{(p)}\right)$ be the predicted risk score for patient p, with observed survival time $t^{(p)}$ and event indicator $e^{(p)} \in \{0,1\}$ ($e^{(p)} = 1$ if the event occurred, 0 if censored). The risk set for patient p is $\mathcal{R}^{(p)} = \{p' \mid t^{(p')} \geq t^{(p)}\}$. The negative Cox partial loglikelihood is

$$\mathcal{L}_{\text{cox}} = -\sum_{p:e^{(p)}=1} \left(r^{(p)} - \log \sum_{p' \in \mathcal{R}^{(p)}} \exp\left(r^{(p')}\right) \right).$$

This loss directly optimizes survival prediction from the fused representations $\boldsymbol{z}_{\text{fused}}^{(p)}$, ensuring the model captures event times while handling censored data. Importantly, it does not update the projection heads $\mathcal{G}_i$.

2) Mutual information loss

To encourage dependency between modalities, we apply a contrastive InfoNCE loss on projected embeddings $\boldsymbol{q}_i^{(p)} = \mathcal{G}_i\left(\boldsymbol{z}_i^{(p)}\right)$. For patient p and modalities (i, j), the positive pair is $\left(\boldsymbol{q}_i^{(p)}, \boldsymbol{q}_j^{(p)}\right)$, while negatives are $\left(\boldsymbol{q}_i^{(p)}, \boldsymbol{q}_j^{(p')}\right)$ for $p' \neq p$ within the minibatch $\mathcal{B}$. With cosine similarity $s(\cdot,\cdot)$ and temperature τ, the directional loss is

$$\ell_{\text{mi}}^{(p,i\to j)} = -\log \frac{\exp\left(s\left(\boldsymbol{q}_i^{(p)}, \boldsymbol{q}_j^{(p)}\right)\right)}{\sum_{p' \in \mathcal{B}} \exp\left(s\left(\boldsymbol{q}_i^{(p)}, \boldsymbol{q}_j^{(p')}\right)\right)}.$$

The batch-averaged, symmetric loss is

$$\mathcal{L}_{\text{mi}} = \frac{1}{|\mathcal{B}|} \sum_{p \in \mathcal{B}} \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \frac{\ell_{\text{mi}}^{(p,i\to j)} + \ell_{\text{mi}}^{(p,j\to i)}}{2},$$

where $\mathcal{P}$ is the set of available modality pairs for p, and $|\mathcal{B}|$ is the minibatch size. This loss aligns embeddings of different modalities for the same patient, while keeping different patients separated.

3) Reconstruction loss

To preserve modality-specific information, we reconstruct unimodal embeddings from $\boldsymbol{z}_{\text{fused}}^{(p)}$. For each modality i, a decoder $\mathcal{H}_i$ generates a reconstruction $\hat{\boldsymbol{z}}_i^{(p)} = \mathcal{H}_i\left(\boldsymbol{z}_{\text{fused}}^{(p)}\right)$. The reconstruction loss penalizes discrepancies for available modalities only:

$$\mathcal{L}_{\text{recon}} = \frac{1}{|\mathcal{B}|} \sum_{p \in \mathcal{B}} \frac{1}{\sum_{i=1}^{M} a_i^{(p)}} \sum_{i=1}^{M} a_i^{(p)} \left\| \boldsymbol{z}_i^{(p)} - \hat{\boldsymbol{z}}_i^{(p)} \right\|_2^{2.}$$

Here, $a_i^{(p)} \in \{0,1\}$ indicates availability of modality i for patient p, and $\sum_{i=1}^{M} a_i^{(p)}$ counts the number of modalities present. This loss ensures $\boldsymbol{z}_{\text{fused}}^{(p)}$ preserves unimodal characteristics, which improves robustness in missing-modality scenarios.

4) Overall loss

The final training objective combines the three components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cox}} + \lambda_{\text{mi}} \mathcal{L}_{\text{mi}} + \lambda_{\text{recon}} \mathcal{L}_{\text{recon}},$$

where λ_{mi} and λ_{recon} are scalar coefficients that balance the auxiliary objectives. This joint optimization enforces accurate survival prediction from the fused representation $\boldsymbol{z}_{\text{fused}}^{(p)}$, while simultaneously promoting cross-modal alignment through InfoNCE and preserving modality-specific information via reconstruction.

IV. Experiments

A. Datasets

We conducted our experiments using three well-established benchmark datasets: the BraTS dataset [26, 27], the Cancer Imaging Archive (TCIA) [28], and The Cancer Genome Atlas Program (TCGA) [29, 30].

Summarized in Table I, every patient had at least one available modality, and all patients had survival time and censor status. In line with clinical practice, pathology and genomic data are often more invasive to acquire than radiological scans or demographic information. To reflect this reality, we simulated missing-modality conditions by selectively dropping pathology and genomic modalities during evaluation.

TABLE I. Number of Patients by Data Type

Modalities	Number of patients
Pathology	769
Radiology	376
Genomics	697
Demographics	803
Patients with complete modalities	170
Patients with survival labels	962

For comparability, we followed Cui *et al.* [15] and used their original 15-fold Monte Carlo cross-validation splits.

B. Implementation Details

The Concordance index (C-index) was used as the primary evaluation metric, reported as the median across cross-validation folds. Following Cui *et al.* [15], we explicitly evaluated under missing-modality conditions, focusing on pathology and genomics, which are less frequently available in practice.

To improve robustness, we applied modality dropout during training with a dropout rate of 0.5, randomly masking modalities to simulate incomplete patient data. The unimodal encoders were trained with a batch size of 64, and the fusion module with a batch size of 8. Both were optimized with Adam at a learning rate of 0.0005. All other hyperparameters were preserved from Cui *et al.* [15] to isolate the impact of our proposed cross-modal feature enhancement block.

The weights for auxiliary objectives were empirically set to $\lambda_{\text{mi}} = \lambda_{\text{recon}} = 0.5$.

TABLE II. COMPARISON OF SCML AGAINST STATE-OF-THE-ART FRAMEWORKS

Method	C-index
Deep Orthogonal Fusion [14]	0.7624 ± 0.042
Pathomic Fusion [13]	0.7697 ± 0.047
MMD [15]	0.8053 ± 0.038
SCML (Ours)	**0.8147 ± 0.025**

V. RESULTS AND DISCUSSION

As shown in Table II, the proposed framework outperforms the evaluated multimodal baselines and achieves the strongest performance observed in our experiments on this dataset. Importantly, this improvement is obtained without modifying the baseline fusion mechanism or the unimodal encoders, underscoring that the performance gain comes from the SCML refinement strategy itself rather than from a stronger fusion architecture.

Table III reports a comprehensive evaluation of the proposed framework across different configurations. The best-performing setting—Selective Cross-Modal Learning (SCML) with attention, modality dropout, and reconstruction enabled—achieves the highest C-index in all evaluated scenarios, reaching 0.8147±0.028 under complete modalities. Because the unimodal encoders and the final mean-fusion strategy were kept identical to the strong baseline of Cui *et al.* [15], these gains can be attributed to the proposed SCML refinement module rather than to changes in encoder architecture or fusion design.

Both modality dropout and reconstruction also contribute to robustness under incomplete inputs. Modality dropout exposes the model to realistic missing-modality conditions during training, while the reconstruction objective encourages the enhanced representations to retain unimodal information. Their combination leads to consistent improvements across the missing-modality settings.

Table IV reports the C-index results obtained for all unimodal, pairwise, three-modality, and complete modality combinations. These results show that multimodal performance depends not only on the number of available modalities, but also on the quality of the interactions between modality representations. To further characterize these interactions, Table V reports pairwise synergy scores derived from the modality-combination results in Table IV. For each modality pair, synergy measures whether the fused pair improves over its strongest unimodal component, thereby providing a simple criterion to distinguish complementary from potentially interfering interactions. To better interpret whether two modalities contribute complementary or interfering information, we define the pairwise synergy score between modalities (i) and (j) as

$$\text{Synergy}(i,j) = \text{Perf}\left(z_i \oplus z_j\right) - \max\left\{\text{Perf}(z_i),\ \text{Perf}\left(z_j\right)\right\},$$

where $\text{Perf}(\cdot)$ denotes the validation C-index and $\oplus$ denotes fusion of the two modality representations. A positive value indicates complementary information, a value near zero indicates redundancy, and a negative value indicates potential interference or negative transfer.

Table IV indicates that multimodal performance depends not only on the number of available modalities, but also on the quality of the interactions between modality representations. Rather than selecting the best global combination of enhanced modalities based only on the final C-index, the revised SCML strategy uses the pairwise synergy analysis reported in Table V to determine which modality pairs provide beneficial complementary information. Cross-modal attention is then applied only between modalities with positive synergy, so that a modality is enhanced only by representations that are empirically supportive of its refinement. This refinement is consistent with the intuition behind SCML and leads to an improvement of the C-index from 0.8113 to 0.8147, and reduces inter-fold variance from ±0.028 to ±0.025.

TABLE III. AVERAGE C-INDEX COMPARISON OF DIFFERENT METHODS AND CONFIGURATIONS

Fusion Method	Cross-modal Attention	Dropout	Recon	Complete Modalities	Pathology Missing	Gene + Pathology Missing
Mean Vector (MMD)	×	×	×	0.7744 ± 0.034	0.7616 ± 0.034	0.7357 ± 0.035
Mean Vector (CML)	✓	×	×	0.7728 ± 0.032	0.7726 ± 0.031	0.7075 ± 0.035
Mean Vector (SCML)	✓	×	×	0.7833 ± 0.028	0.7784 ± 0.031	0.7507 ± 0.031
Mean Vector (MMD)	×	✓	×	0.7858 ± 0.030	0.7700 ± 0.032	0.7415 ± 0.033
Mean Vector (CML)	✓	✓	×	0.7814 ± 0.027	0.7834 ± 0.028	0.7385 ± 0.032
Mean Vector (SCML)	✓	✓	×	0.7815 ± 0.025	0.7731 ± 0.029	0.7464 ± 0.025
Mean Vector (MMD)	×	×	✓	0.7805 ± 0.032	0.7727 ± 0.033	0.7297 ± 0.037
Mean Vector (CML)	✓	×	✓	0.7929 ± 0.028	0.7892 ± 0.026	0.7528 ± 0.036
Mean Vector (SCML)	✓	×	✓	0.7890 ± 0.029	0.7831 ± 0.027	0.7503 ± 0.035
Mean Vector (MMD)	×	✓	✓	0.7826 ± 0.031	0.7853 ± 0.032	0.7414 ± 0.034
Mean Vector (CML)	✓	✓	✓	0.7999 ± 0.025	0.7863 ± 0.027	0.7596 ± 0.030
Mean Vector (SCML)	✓	✓	✓	0.8113 ± 0.028	0.7922 ± 0.027	0.7612 ± 0.027

TABLE IV. C-INDEX RESULTS FOR SELECTIVE CROSS-MODAL LEARNING ACROSS MODALITY COMBINATIONS

Radiology	Pathology	Demographics	Genomics	C-index
✓	×	×	×	0.7747 ± 0.027
×	✓	×	×	0.7615 ± 0.025
×	×	✓	×	0.7809 ± 0.033
×	×	×	✓	0.7899 ± 0.032
✓	✓	×	×	0.7788 ± 0.027
✓	×	✓	×	0.7900 ± 0.026
✓	×	×	✓	0.7883 ± 0.029
×	✓	✓	×	0.7803 ± 0.027
×	✓	×	✓	0.7815 ± 0.028
×	×	✓	✓	0.7954 ± 0.026
✓	✓	✓	×	0.7810 ± 0.027
✓	✓	×	✓	0.7919 ± 0.025
✓	×	✓	✓	0.8113 ± 0.028
×	✓	✓	✓	0.7846 ± 0.028
✓	✓	✓	✓	0.7999 ± 0.025

TABLE V. PAIRWISE CROSS-MODAL SYNERGY SCORES DERIVED FROM THE MODALITY-COMBINATION RESULTS

Pair	C-idx(*i*)	C-idx(*j*)	C-idx(fused)	Synergy
Radio + Path	0.7747	0.7615	0.7788	+0.0041
Radio + Demo	0.7747	0.7809	0.7900	+0.0091
Radio + Gene	0.7747	0.7899	0.7883	–0.0016
Path + Demo	0.7615	0.7809	0.7803	–0.0006
Path + Gene	0.7615	0.7899	0.7815	–0.0084
Demo + Gene	0.7809	0.7899	0.7954	+0.0055

As shown in Table II, the proposed framework outperforms the evaluated multimodal baselines and achieves the strongest performance observed in our experiments on this dataset. Importantly, this improvement is obtained without modifying the baseline fusion mechanism or the unimodal encoders, underscoring that the performance gain comes from the SCML refinement strategy itself rather than from a stronger fusion architecture.

We further investigated modality contributions using the HVS measure. Fig. 6 demonstrates that SCML yields a more balanced distribution of modality importance, elevating underrepresented sources such as demographics and genomics. This supports our claim that HVS weighting makes the cross-modal refinement interpretable and prevents dominance by a single modality.

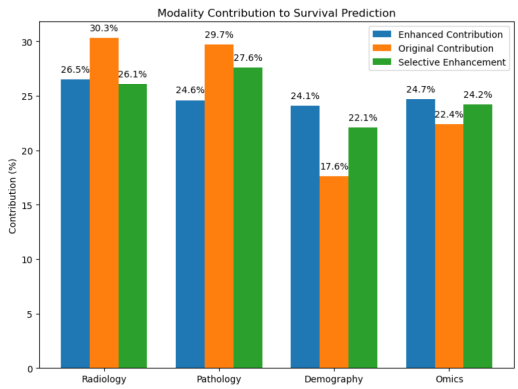

Fig. 6. Modality contribution percentages using HVS scores.

From Table III, each SCML component plays a distinct role. Modality dropout improves robustness under missing-modality conditions, while the reconstruction loss helps stabilize the fused representation and preserve unimodal characteristics. Cross-modal attention alone already improves upon the baselines, but the full SCML configuration (attention + dropout + reconstruction) yields the strongest and most consistent results, highlighting the complementary effect of these components.

The HVS analysis in Fig. 6 further illustrates how SCML reshapes modality contributions. Compared with the baselines, SCML yields a more balanced contribution profile and increases the relative importance of demographics and omics, which is consistent with their complementary prognostic value alongside radiology.

Finally, statistical validation using the Friedman test and the Nemenyi post-hoc test at $\alpha = 0.05$ confirmed significant differences among MMD (MR = 2.83), CML (MR = 1.83), and SCML (MR = 1.33), with $p < 0.001$ and a critical distance of 0.428, as shown in Figs. 7 and 8.

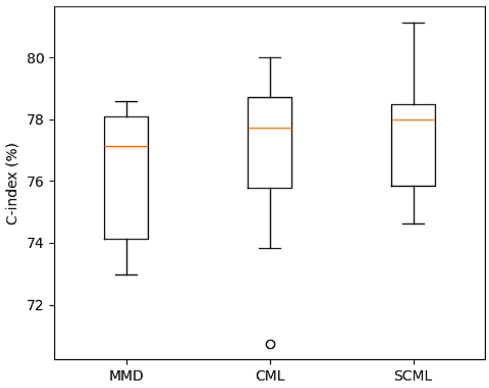

Fig. 7. Distribution of median C-index values for MMD, CML, and SCML.

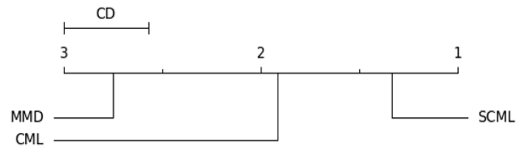

Fig. 8. Mean rank comparison from Friedman and Nemenyi tests.

The proposed framework introduces additional computational components. Nevertheless, the design improves robustness to missing data, provides interpretable cross-modal weighting, and achieves strong performance across the evaluated multimodal survival prediction settings.

Although masked averaging offers a simple and robust way to aggregate available modalities under incomplete data, it also imposes an important limitation: the final fusion stage treats all available enhanced embeddings uniformly and does not explicitly adapt to patient-specific modality reliability, uncertainty, or higher-order interactions among modalities. Consequently, an informative modality may be diluted when averaged with noisier or redundant representations, while clinically weak

modalities may still contribute equally once present. In SCML, this limitation is partially mitigated before fusion through selective cross-modal enhancement, HVS-guided modality weighting, and synergy-based attention, which aim to improve the quality of each modality representation prior to aggregation. Nevertheless, the final masked average remains a deliberately lightweight design choice, allowing us to isolate the contribution of the proposed refinement module and maintain direct comparability with mean-fusion baselines. Future work will explore learnable, uncertainty-aware, or patient-adaptive fusion mechanisms while preserving robustness to arbitrary missing-modality patterns.

VI. Conclusion

This paper introduced Selective Cross-Modal Learning (SCML), a lightweight pre-fusion enhancement framework for multimodal survival prediction with incomplete heterogeneous data. SCML refines modality-specific embeddings through projection-space alignment, selective cross-modal attention, HVS-guided weighting, and reconstruction-based regularization before final fusion. By enhancing unimodal representations prior to aggregation, the proposed approach addresses an important limitation of conventional late-fusion pipelines, where cross-modal interactions are only exploited after feature extraction.

Experiments on a glioma survival prediction cohort of 962 patients showed that SCML improves over strong multimodal baselines, achieving a median C-index of 0.8147±0.025. The ablation analysis further showed that cross-modal attention, modality dropout, and reconstruction provide complementary benefits, while the pairwise synergy analysis demonstrated that not all modality interactions are equally beneficial. These results support the central idea that selective, empirically guided cross-modal enhancement can improve robustness and reduce negative transfer under incomplete multimodal settings.

Despite these findings, the framework has some limitations. The final masked averaging fusion remains simple and may not fully capture patient-specific modality reliability, uncertainty, or higher-order interactions among modalities. In addition, the synergy-guided refinement was evaluated on a single glioma survival prediction setting and should be further validated on larger and more diverse cohorts. Future work will investigate learnable and uncertainty-aware fusion mechanisms, external clinical validation, and extensions to other multimodal health-care tasks. Overall, SCML provides a promising step toward robust, interpretable, and missing-modality-aware multimodal survival prediction.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Mohammed Belghadi conducted the experiments, implemented the proposed framework, analyzed the results, and wrote the initial draft of the manuscript. Younès Bennani supervised the research, contributed to the methodological design, and reviewed and revised the manuscript. Valérie Paradis provided clinical expertise, contributed to the interpretation and understanding of the biomedical data, and reviewed the manuscript from a medical perspective. All authors have read and approved the final version of the manuscript.

Funding

This research was funded as part of the SiRIC InsiTu project ("INSIghts Into Cancer: From Inflammation to Tumor"), by the National Cancer Institute, the Ministry of Health and Prevention, and Inserm; Grant Number: INCa-DGOS-Inserm-ITMo Cancer 18008.

References

[1] Q. Lin, Y. Zhu, X. Mei, L. Huang, J. Ma, K. He, Z. Peng, E. Cambria, and M. Feng, "Has multimodal learning delivered universal intelligence in healthcare? A comprehensive survey," *Inf. Fusion*, vol. 116, 102795, Nov. 2024. doi: 10.1016/j.inffus.2024.102795

[2] I. Ahmad and F. Alqurashi, "Early cancer detection using deep learning and medical imaging: A survey," *Crit. Rev. Oncol. Hematol.*, vol. 204, 104528, 2024. doi: 10.1016/j.critrevonc.2024.104528

[3] A. Tiwari, A. Ghose, M. Hasanova, S. S. Faria, S. Mohapatra, S. Adeleke, and S. S. Boussios, "The current landscape of artificial intelligence in computational histopathology for cancer diagnosis," *Discover Oncology*, vol. 16, 438, 2025. doi: 10.1007/s12672-025-02212-z

[4] K. H. Yu and M. Snyder, "Omics profiling in precision oncology," *Mol. Cell. Proteomics*, vol. 15, no. 8, pp. 2525–2536, Aug. 2016. doi: 10.1074/mcp.O116.059253

[5] D. Horgan, M. Van den Bulcke *et al.*, "Demographic analysis of cancer research priorities and treatment correlations," *Curr. Oncol.*, vol. 31, no. 4, pp. 1839–1864, Mar. 2024. doi: 10.3390/curroncol31040139

[6] Y. Li, J. Jiang, X. Li, and M. Zhang, "Multi-modal data integration of dosiomics, radiomics, deep features, and clinical data for radiation-induced lung damage prediction in breast cancer patients," *J. Radiat. Res. Appl. Sci.*, vol. 18, no. 2, 101389, 2025. doi: 10.1016/j.jrras.2025.101389

[7] Glioblastoma Research Organization. (2023). What is the average glioblastoma survival rate? [Online]. Available: https://www.gbmresearch.org/blog/glioblastoma-survival-rate

[8] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, Feb. 2016. doi: 10.1148/radiol.2015151169

[9] L. Saba, M. Biswas *et al.*, "The present and future of deep learning in radiology," *Eur. J. Radiol.*, vol. 114, pp. 14–24, May 2019. doi: 10.1016/j.ejrad.2019.02.038

[10] W. S. El-Deiry, R. M. Goldberg *et al.*, "The current state of molecular testing in the treatment of patients with solid tumors, 2019," *CA Cancer J. Clin.*, vol. 69, no. 4, pp. 305–343, Jul. 2019. doi: 10.3322/caac.21560

[11] O. J. Skrede, S. de Raedt *et al.*, "Deep learning for prediction of colorectal cancer outcome: A discovery and validation study," *The Lancet*, vol. 395, no. 10221, pp. 350–360, Feb. 2020. doi: 10.1016/S0140-6736(19)32998-8

[12] P. Mobadersany, S. Yousefi, M. Amgad, D. Gutman, J. Barnholtz-Sloan, J. Vega, D. Brat, and L. Cooper, "Predicting cancer outcomes from histology and genomics using convolutional networks," *Proc. Natl. Acad. Sci. USA*, vol. 115, no. 13, pp. 1–6, Mar. 2018. doi: 10.1073/pnas.1717139115.

[13] R. Chen, M. Lu, J. Wang, D. Williamson, S. Rodig, N. Lindeman, and F. Mahmood, "Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," *IEEE Trans. Med. Imaging*, vol. 39, no. 9, pp. 2821–2831, Sep. 2020. doi: 10.1109/TMI.2020.2981028

[14] N. Braman, J. W. Gordon, E. Goossens, C. Willis, M. Stumpe, and J. Venkataraman, "Deep or+thogonal fusion: Multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, 2021, pp. 667–677. doi: 10.1007/978-3-030-87237-3_64

[15] C. Cui, H. Liu, Q. Liu, R. Deng, Z. Asad, Y. Wang, S. Zhao, H. Yang, B. A. Landman, and Y. Huo, "Survival prediction of brain cancer with incomplete radiology, pathology, genomics, and demographic data," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, 2022, pp. 1–10. doi: 10.1007/978-3-031-16443-9_1

[16] H. H. Lee, A. Santamaria-Pang, J. Merkov, M. Lungren, and I. Tarapov, "Multi-modal Mamba modeling for survival prediction (M4Survive): Adapting joint foundation model representations," arXiv preprint, arXiv:2503.12345, 2025.

[17] Y. Yifan, X. Zhang, Y. Wang, H. Li, Z. Qi, Z. Du, Y.-H. Chu, D. Feng, J. Hu, Q. Xie, J. Song, Y. Liu, and J. Cai, "Multimodal data integration using deep learning predicts overall survival of patients with glioma," *View*, vol. 5, no. 4, 20240015, Aug. 2024. doi: 10.1002/VIW.20240015

[18] H. Zheng, S. Yang, Z. He, J. Yang, and Z. Huang, "Hierarchical cross-modal prompt learning for vision-language models," arXiv preprint, arXiv:2507.14976, 2025.

[19] S. Cheng and K. Han, "Vamp: Variational multi-modal prompt learning for vision-language models," arXiv preprint, arXiv:2502.11838, 2025.

[20] M. Shao, S. Guo, X. Li, X. Miao, H. Duan, and Y. Long, "VMFCoop: Towards equilibrium on a unified hyperspherical manifold for prompting biomedical VLMs," arXiv preprint, arXiv:2511.09540, 2025.

[21] F. Isensee, J. Petersen, S. A. Kohl, P. F. Jäger, and K. H. Maier-Hein, "nnU-Net: Breaking the spell on successful medical image segmentation," arXiv preprint, arXiv:1904.08128, 2019.

[22] J. J. M. van Griethuysen, A. Fedorov *et al.*, "Computational radiomics system to decode the radiographic phenotype," *Cancer Res.*, vol. 77, no. 21, pp. e104–e107, 2017. doi: 10.1158/0008-5472.CAN-17-0339

[23] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 971–980.

[24] M. Yacoub and Y. Bennani, "HVS: A heuristic for variable selection in multilayer artificial neural network classifier," *Intell. Eng. Syst. Through Artif. Neural Netw.*, vol. 7, 1997, pp. 1–6.

[25] M. Yacoub and Y. Bennani, "Features selection and architecture optimisation in connectionist systems," *Int. J. Neural Syst.*, vol. 10, no. 5, pp. 379–395, 2000. doi: 10.1142/S0129065700000338

[26] S. Bakas, M. Reyes, A. Jakab *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge," arXiv preprint, arXiv:1811.02629, 2018.

[27] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.,* "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, 2014. doi: 10.1109/TMI.2014.2377694

[28] K. Clark, B. Vendt *et al.*, "The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imaging*, vol. 26, no. 6, pp. 1045–1057, 2013. doi: 10.1007/s10278-013-9622-7

[29] L. Scarpace, T. Mikkelsen, S. Cha, S. Rao, S. Tekchandani, D. Gutman, J. Saltz, and B. Erickson, "The Cancer Genome Atlas Glioblastoma Multiforme Collection (TCGA-GBM) (Version 5) [Data set]," *The Cancer Imaging Archive*, 2016. doi: 10.7937/K9/TCIA.2016.RNYFUYE9

[30] N. Pedano, A. Flanders, L. Scarpace, T. Mikkelsen, K. Eschbacher, B. Hermes, M. R. Rosenfeld, and B. Erickson, "The Cancer Genome Atlas Low Grade Glioma Collection (TCGA-LGG) (Version 3) [Data set]," *The Cancer Imaging Archive*, 2016. doi: 10.7937/K9/TCIA.2016.L4LTD3TK