

<https://doi.org/10.1038/s41746-025-02257-y>

Multimodal deep learning for cancer prognosis prediction with clinical information prompts integration

Check for updates

Jiaxin Hou^{1,2}, Ranran Zhang¹, Yaoqin Xie¹, Chao Li^{3,4} & Wenjian Qin¹ ✉

Survival prediction is crucial for guiding cancer treatment and evaluating therapeutic efficacy. However, tumor heterogeneity presents challenges of accurate prognosis. Multimodal learning, which integrates data from imaging, genomics, and clinical records, offers a promising approach for this complex task. While recent studies mainly focus on imaging and genomic data, clinical information, which reflecting patients' overall health, remains underutilized due to its discrete, sparse, and low-dimensional characteristics. We propose SurvPGC, an integrated model combining pathology images, genomic data and clinical records for cancer prognosis. Clinical information is transformed into high-dimensional vectors using text templates and foundation models, enabling their integration through a cross-attention module. Validation on three datasets of The Cancer Genome Atlas demonstrated that the model effectively captures modality-specific features, with attention visualization revealing distinct focus areas across data types. This highlights the importance of incorporating diverse information sources for improved survival prediction.

Survival analysis models time-to-event data to predict the timing or probability of events occurring at specific points in time¹. Accurate survival prediction helps avoid medical over-treatment and provides a scientific basis for doctors to make decisions, aiding in treatment planning and outcome evaluation². However, survival prediction is challenging due to multiple influencing factors, especially in cancers with high heterogeneity³. Deep learning possesses powerful feature extraction and representation capabilities, enabling it to effectively model complex non-linear relationships. Consequently, it has been widely applied to address the challenge of tumor survival prediction, leading to the development of numerous models^{4–10}. Pathological images contain rich visual information and serve as the gold standard for cancer diagnosis. Therefore, they represent a critical data modality in these studies. However, the use of single-modal data limits the comprehensiveness of the analysis. To improve prediction accuracy and better meet clinical requirements, multimodal tumor survival prediction models have emerged as a prevailing trend.

Multimodal learning methods utilize comprehensive information from various data sources, making them suitable for addressing complex problems like survival prediction¹¹. In most multimodal prognosis studies^{12–14}, in addition to pathological images, genomic data is also often used, which provide insights into molecular mechanisms in tumor development. Nevertheless, readily available clinical information is often

overlooked or relegated to an auxiliary role¹¹, especially in deep learning-based models.

Clinical information can provide comprehensive context for survival predictions by enhancing the model's understanding of a patient's overall health¹⁵. However, clinical data are often discrete and low-dimensional, which limits their utility in deep learning models that typically excel with high-dimensional features¹⁶. Figure 1 shows three common approaches to integrating clinical information with other modalities. Approach a separately constructs models using clinical characteristics and data from other modalities, and then fuses the resulting risk scores, which limits the potential for inter-modal feature interaction. Approach b concatenates high-dimensional embeddings of other modalities with standardized clinical features, creating a dimensionality gap that diminishes the impact of clinical information^{17,18}. Approach c aligns modalities into the same dimension before fusion but struggles with encoding low-dimensional clinical information into high-dimensional features and neglects contextual information in clinical inputs¹⁹. Drawing from advancements in natural language processing (NLP)²⁰ and image-text integration²¹, this study developed text templates for clinical information and utilized a text foundation model to transform clinical text prompts into high-dimensional embeddings, facilitating deeper participation of clinical information in model training.

¹Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. ²University of Chinese Academy of Sciences, Beijing, China.

³Department of Applied Mathematics & Theoretical Physics, University of Cambridge, Cambridge, UK. ⁴Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK. ✉e-mail: wj.qin@siat.ac.cn

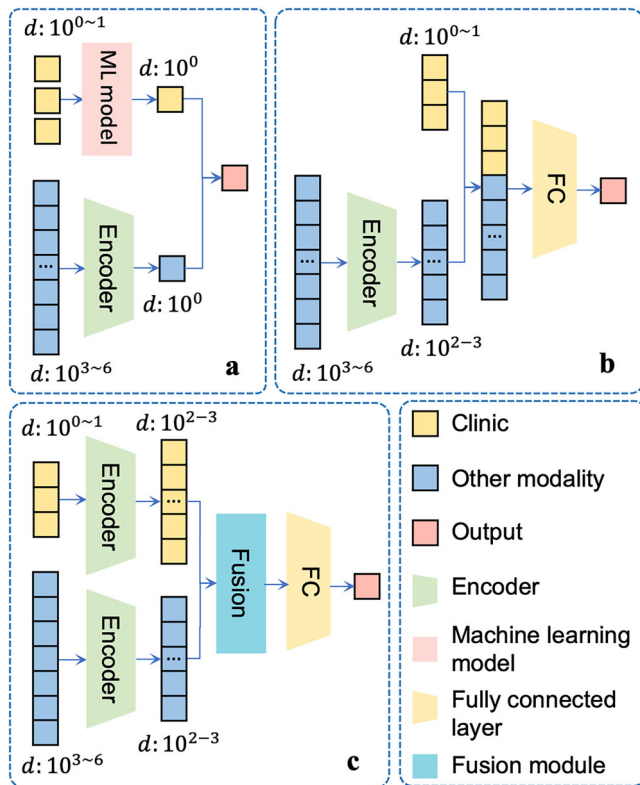


Fig. 1 | Three common approaches to multimodal learning that integrate clinical information with other modalities. **a** Constructing separate models for clinical characteristics and other modalities, followed by fusion of the output risk scores. **b** Concatenating clinical characteristics with high-dimensional embeddings of other modalities. **c** Encoding low-dimensional clinical data into high-dimensional embeddings and fusing with other modalities. *d*: dimension.

The strong representation ability of features and the appropriate fusion method of multimodal are two key points in multimodal learning research²². With the evolution of foundation models, the features they extract exhibit strong generalization capabilities, and numerous studies have validated the effectiveness of foundation model embeddings in summarizing data across various tasks^{23–25}. Consequently, this paper employed different foundation models to derive embeddings from data. Regarding multimodal fusion, particularly in the context of tumor survival prediction, these studies^{14,26–29} have all employed feature-level fusion to enable cross-modal information interaction, which represents a key advantage of this approach over early fusion (data-level fusion) and late fusion (decision-level fusion)³⁰. Among these, SurvPath²⁹ is a state-of-the-art (SOTA) method that integrates pathological images and genomic data for cancer survival prediction. SurvPath utilizes a cross-modal attention fusion strategy, which not only enhances information exchange between modalities but also links genomic data with pathway markers to specific regions in pathological images, yielding interpretable visualization results. SurvPath has demonstrated superior performance across five datasets from The Cancer Genome Atlas (TCGA), outperforming previous multimodal survival prediction approaches. However, SurvPath did not further explore patient characteristics that are more commonly available in clinical practice. Given that genomic data are more difficult to obtain than clinical features, this study aims to investigate whether incorporating clinical information can further enhance the performance of multimodal survival prediction models, whether clinical features can serve a role comparable to genomic data, and how clinical information differs from genomic data in terms of the pathological image regions it emphasizes. Since the clinical data in the TCGA dataset are low-dimensional tabular data, we propose a method to construct clinical prompts and tokenize them, enabling more effective participation in high-dimensional interactions.

In summary, the key contributions of this paper are: (1) Designing text templates for clinical information and tokenizing clinical text into high-dimensional embeddings to support survival prediction; (2) Enhancing survival prediction performance through the cross-attention fusion of multimodal embeddings from foundation models; (3) Validating the survival prediction model SurvPGC on liver hepatocellular carcinoma (LIHC), breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ) datasets from TCGA.

Results

This study proposed a workflow for cancer overall survival prediction based on multimodal data, including the preprocessing and embedding encoding of whole slide images (WSIs), transcriptomics and clinic information, the dual-path cross-attention fusion model SurvPGC, as shown in Fig. 2.

Datasets and implementation details

We selected TCGA-LIHC (354 cases) and TCGA-BRCA (1035 cases) datasets to evaluate the model, adopted the dataset TCGA-COADREAD (298 cases) collected and organized in ref.²⁹ for additional model validation. The case selecting processes are shown in Supplementary Fig. 1 in the supplementary materials, outlining the inclusion and exclusion criteria used to ensure data quality and relevance. The censoring ratio of 3 datasets is presented in Supplementary Fig. 2. We focused on overall survival as the target task with the most complete record. We adopted 5-fold cross validation, and the result is the average of 5 folds. During model training, the loss generally begins to converge after 10 epochs. We therefore set the total number of epochs to 20 and selected the model with the best performance on the validation set during epochs 10–20. Other implementation details are provided in the supplementary materials Section 4.

Concordance index (C-index) comparison of unimodal and multimodal methods

We conducted comparison experiments with other unimodal and multimodal methods. The results are shown in Table 1.

Unimodal: Clinic: The Cox proportional hazards model is one of the most widely used multivariate survival prediction models³¹, which incorporates clinical characteristics as risk factors. We selected the Cox model as the baseline for clinic-based unimodal prediction. The multilayer perceptron (MLP)³² is a fundamental neural network architecture and was employed to evaluate the predictive capability of unimodal embeddings. Self-normalizing neural networks (SNN) integrate self-normalization operations into the layers of MLP, thereby enhancing the stability of model training³³. In prior studies, SNN has also been frequently applied to assess the survival prediction performance of unimodal features^{12,27}. **Genomic:** For genomic data, both MLP and SNN networks were utilized to validate the survival prediction capability of unimodal embeddings. **Pathology:** Due to the large size and weak labeling characteristics of WSIs, computations typically employ multi-instance learning methods. Similar to other modalities, MLP was used as the baseline. Prior to the final fully connected layer, patch-level outputs were aggregated by computing their mean value. Attention-based Multiple Instance Learning (ABMIL)³⁴ improves patch aggregation by learning to assign different weights to patches based on their importance. TransMIL³⁵ leverages the Transformer module to incorporate correlations among patch instances during the aggregation process, making it one of the representative methods for multi-instance learning of WSIs.

Multimodal: For the multimodal learning comparison experiments, we selected three intermediate fusion methods. PORPOISE²⁷ integrates feature vectors derived from genomic and pathological data through bilinear pooling, computing the outer product to comprehensively capture pairwise relationships across modalities. MCAT²⁸ employs an attention mechanism, using gene-encoded vectors to construct query (Q) vectors, thereby enabling gene-guided cross-modal attention fusion and identifying pathological regions associated with distinct gene functions. SurvPath²⁹, building upon MCAT, transforms unidirectional modal-guided fusion into bidirectional

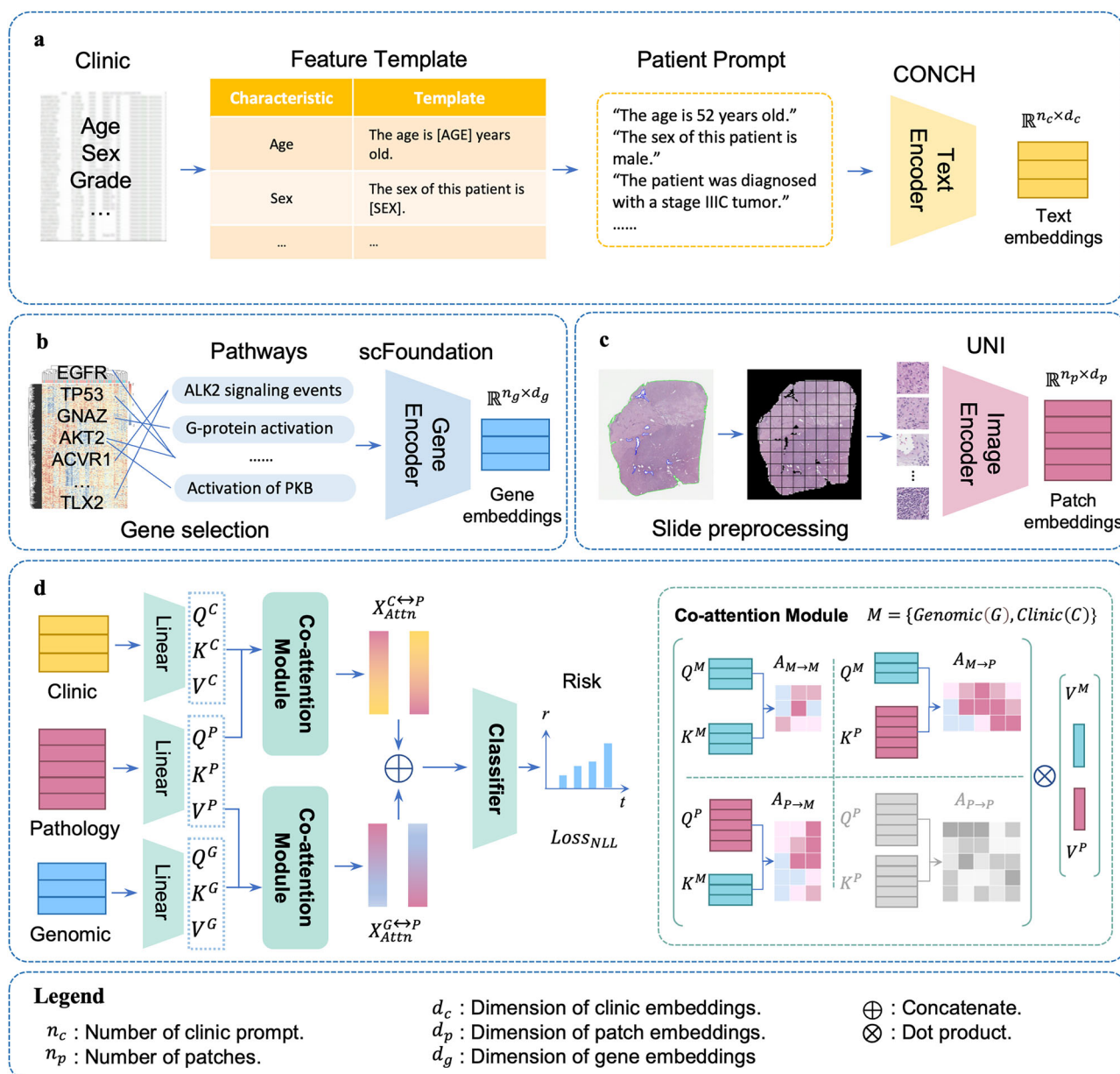


Fig. 2 | The overview of the proposed workflow. a Clinic information prompt generation and embedding encoding. **b** Transcriptome data preprocessing and embedding encoding. **c** Whole slide images preprocessing and embedding encoding. **d** The dual-path cross-attention fusion structure of SurvPGC. P: pathology, G: genomic, C: clinic.

fusion and refines the input gene pathways, thereby further improving the model's prognostic performance. SurvPath currently represents the SOTA method in this field.

From the perspective of unimodal models, the prognostic predictive capabilities of clinical information and genomic data vary across datasets. On TCGA-LIHC, SNN and MLP models based on genomic data achieved higher C-index values compared to those based on clinical features. In contrast, on TCGA-BRCA and TCGA-COADREAD, clinical models demonstrated superior performance. The predictive ability of pathological images also varied across datasets. From the perspective of multimodal models, SurvPGC, which integrates three modalities, outperformed the other models that incorporate only two modalities, with the most improvement observed on TCGA-BRCA. The C-index, a measure of a model's ability to rank patient risks, assesses the ratio of correctly predicted pairs to all pairs in the cohort, revealed that the performance of multimodal models was lower than that of unimodal models across the three datasets in some cases. If the predicted results for each pair of samples are exactly opposite to the actual situation, the value of the C-index is likely to be less

than 0.5. This also indicates that the model has learned in a completely wrong direction. The suboptimal fusion strategies may diminish the prognostic capability of the input data.

As SurvPGC is an extension of SurvPath with the incorporation of clinical information, we conducted bootstrap resampling and statistical difference testing on the C-index values derived from the risk outputs of the two models. Detailed procedures are described in Section 6 of the supplementary materials. In summary, across all five folds of cross-validation, the differences in C-index values between the two models were statistically significant.

Compare the prediction ability at different time points and patient stratification ability of multimodal methods

In addition to assessing the model's ability to rank patient risks, we evaluated the predictive accuracy of the multimodal model at clinically relevant time points (1 year, 3 years, and 5 years) using integrated area under the curve (iAUC). The iAUC values for the multimodal methods, along with the AUC at each individual time point, are summarized in Table 2.

The results demonstrate that SurvPath outperforms other multimodal methods in the iAUC evaluation; however, it does not achieve the highest performance at every individual time point. Overall, the AUC of each model tends to be relatively high within shorter time intervals. As time progresses, both the proportion of censored patients and the effective sample size decrease (Supplementary Fig. 2), which introduces greater challenges for accurate risk prediction. The censoring rates are approximately 65% for TCGA-LIHC, 86% for TCGA-BRCA, and 81% for TCGA-COADREAD. Additionally, the cohort sizes for TCGA-LIHC and TCGA-COADREAD are relatively small. Consequently, the prediction task is more challenging for TCGA-COADREAD, and its iAUC and AUC performance is comparatively lower than that of the other two datasets. Compared with other multimodal methods, SurvPGC incorporates clinical information, which provides the model with a more comprehensive patient profile and leads to improved AUC values.

Figure 3 presents Kaplan-Meier (KM) curves comparing unimodal and multimodal methods on the TCGA-LIHC dataset. The Cox model based on clinical characteristics shows effective risk stratification. Both clinical information embeddings and pathological images demonstrate strong stratification capabilities, whereas genomic data exhibit relatively weaker

stratification performance. Among the multimodal models, PORPOISE and SurvPGC display robust stratification performance. When considering the C-index in Table 1 and the iAUC values in Table 2, PORPOISE shows lower ranking and prediction accuracy but maintains strong stratification ability. SurvPath demonstrates relatively strong ranking and prediction performance but weaker patient stratification. SurvPGC achieves consistently strong results across all evaluation metrics. The KM curves for unimodal and multimodal methods on TCGA-BRCA and TCGA-COADREAD are provided in Supplementary Figs. 3 and Fig. 4 of the supplementary materials.

Effect of clinical information to survival prediction

To evaluate the contribution of dual-path fusion embeddings to the final decisions in the SurvPGC, we employed integrated gradients (IG) to assess the influence of these embeddings before they entered the final classification module. The results, presented in Table 3, indicate that the fusion embeddings of genomic-pathology and clinic-pathology both made considerable contributions.

We visualized attention matrices $A_{c \rightarrow p}$ (clinic-to-pathology) and $A_{g \rightarrow p}$ (genomic-to-pathology) on pathological images to investigate the regions that genomic and clinical concerned. Figure 4 shows typical cases from the TCGA-LIHC, where red indicates high attention and blue indicates low attention. Cases were divided into high- and low-risk groups based on survival times, higher risk correlating with shorter survival times.

Additionally, we compared the top 50 patches selected according to cumulative attention values of $A_{c \rightarrow p}$ and $A_{g \rightarrow p}$, respectively. In TCGA-LIHC (Fig. 4), both clinical and genomic data primarily focus on tumor cell regions within the WSIs. Additionally, genomic data also directs attention to areas such as lymphocytes and necrosis, whereas these regions are rarely selected by $A_{c \rightarrow p}$. However, genomic data is more susceptible to noise and may inappropriately focus on irrelevant artifacts such as handwriting or scanning shadows. In TCGA-BRCA (Supplementary Fig. 5), the attention patterns of clinical and genomic data are similar to those observed in TCGA-LIHC: both modalities emphasize tumor regions, while genomic pathways also concentrate on patches containing tumor stroma and lymphocytes, and remain more prone to noise. In TCGA-COADREAD (Supplementary Fig. 6), the pattern is reversed: genomic data predominantly focuses on tumor cell regions, whereas clinical features highlight a broader range of areas, including tumor, lymphocytes, and stroma. Although the primary focus areas of clinical and genomic data differ across datasets, they are overall complementary.

In previous studies related to pathological images and survival prediction, Shirazi A et al. correlated specific pathological tissue regions with gene expression profiles associated with survival outcomes, indirectly establishing a link between tissue morphology and survival³⁶. Li H et al. demonstrated the association between disordered peritumoral collagen

Table 1 | Survival prediction C-index of unimodal and multimodal methods

| Models/Datasets | | TCGA-LIHC | TCGA-BRCA | TCGA-COADREAD |
|-----------------|----------------------|---------------|---------------|---------------|
| Unimodal | Cox (Clinic) | 0.626 ± 0.051 | 0.677 ± 0.042 | 0.505 ± 0.086 |
| | MLP (Clinic) | 0.599 ± 0.052 | 0.684 ± 0.101 | 0.640 ± 0.123 |
| | SNN (Clinic) | 0.608 ± 0.030 | 0.689 ± 0.072 | 0.615 ± 0.125 |
| | MLP (Genomic) | 0.677 ± 0.088 | 0.611 ± 0.075 | 0.585 ± 0.071 |
| | SNN (Genomic) | 0.668 ± 0.112 | 0.583 ± 0.063 | 0.590 ± 0.129 |
| | MLP (Pathology) | 0.611 ± 0.039 | 0.644 ± 0.040 | 0.499 ± 0.126 |
| | ABISL (Pathology) | 0.624 ± 0.044 | 0.580 ± 0.040 | 0.412 ± 0.100 |
| | TransMIL (Pathology) | 0.633 ± 0.028 | 0.639 ± 0.052 | 0.507 ± 0.077 |
| Multimodal | PORPOISE | 0.638 ± 0.028 | 0.625 ± 0.029 | 0.561 ± 0.085 |
| | MCAT | 0.644 ± 0.071 | 0.553 ± 0.039 | 0.508 ± 0.062 |
| | SurvPath | 0.652 ± 0.054 | 0.622 ± 0.038 | 0.644 ± 0.138 |
| | SurvPGC (ours) | 0.701 ± 0.054 | 0.701 ± 0.057 | 0.676 ± 0.087 |

Table 2 | Survival prediction iAUC of multimodal methods

| Models | | iAUC | AUC (1-year) | AUC (3-year) | AUC (5-year) |
|---------------|----------------|---------------|---------------|---------------|---------------|
| TCGA-LIHC | PORPOISE | 0.649 ± 0.054 | 0.648 ± 0.036 | 0.673 ± 0.104 | 0.617 ± 0.048 |
| | MCAT | 0.656 ± 0.072 | 0.682 ± 0.089 | 0.666 ± 0.094 | 0.625 ± 0.101 |
| | SurvPath | 0.676 ± 0.029 | 0.693 ± 0.051 | 0.671 ± 0.093 | 0.686 ± 0.074 |
| | SurvPGC (ours) | 0.681 ± 0.103 | 0.694 ± 0.112 | 0.715 ± 0.103 | 0.643 ± 0.119 |
| TCGA-BRCA | PORPOISE | 0.619 ± 0.053 | 0.514 ± 0.162 | 0.638 ± 0.077 | 0.628 ± 0.065 |
| | MCAT | 0.571 ± 0.032 | 0.416 ± 0.153 | 0.575 ± 0.040 | 0.602 ± 0.065 |
| | SurvPath | 0.617 ± 0.080 | 0.591 ± 0.196 | 0.589 ± 0.144 | 0.660 ± 0.049 |
| | SurvPGC (ours) | 0.685 ± 0.068 | 0.668 ± 0.132 | 0.700 ± 0.066 | 0.675 ± 0.064 |
| TCGA-COADREAD | PORPOISE | 0.541 ± 0.107 | 0.583 ± 0.132 | 0.558 ± 0.123 | 0.521 ± 0.162 |
| | MCAT | 0.439 ± 0.066 | 0.489 ± 0.073 | 0.476 ± 0.120 | 0.414 ± 0.144 |
| | SurvPath | 0.591 ± 0.150 | 0.720 ± 0.179 | 0.479 ± 0.123 | 0.613 ± 0.182 |
| | SurvPGC (ours) | 0.639 ± 0.109 | 0.657 ± 0.078 | 0.694 ± 0.151 | 0.599 ± 0.189 |

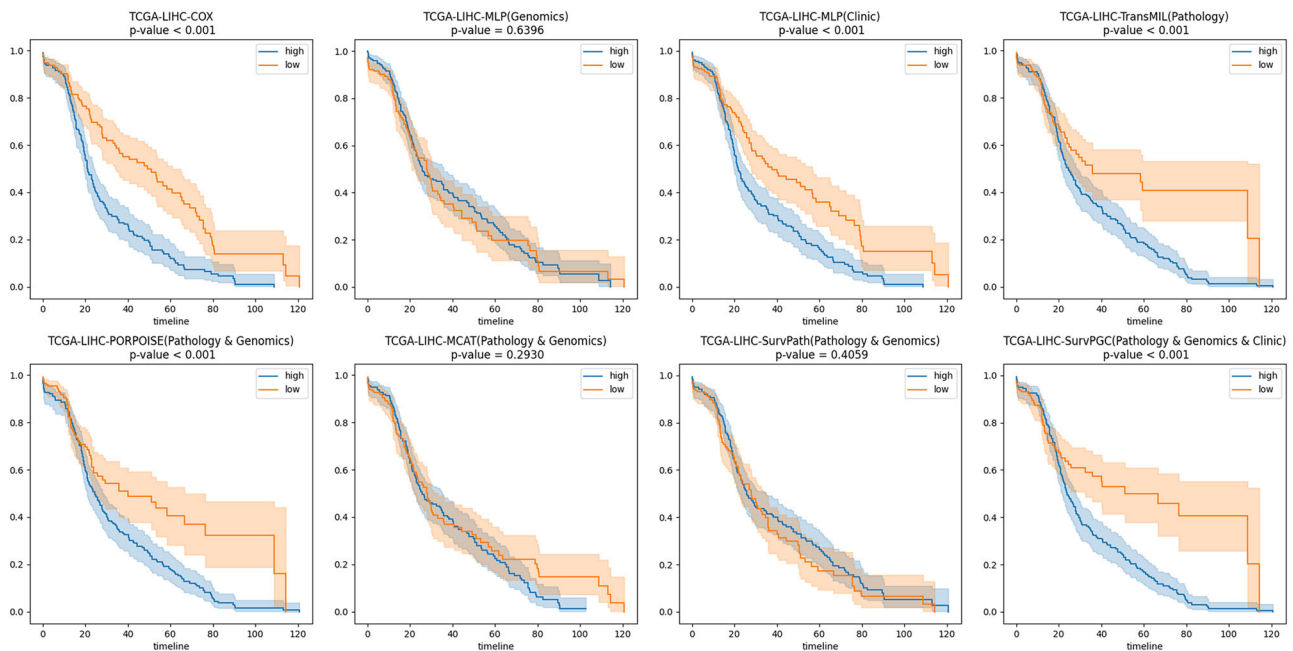


Fig. 3 | Kaplan–Meier curves of unimodal and multimodal methods on dataset TCGA-LIHC. The orange indicates low-risk group, and the blue indicates high-risk group. The p -value of log-rank test less than 0.05 was considered statistically significant.

architecture and the prognosis and metastasis of breast cancer³⁷. The tumor region reflects the degree of malignancy, while peritumoral stromal regions indicate the likelihood of tumor metastasis and invasion—factors that are closely associated with patient survival, including recurrence and metastasis. Therefore, the key information contained in these tissue regions, which clinical and genomic data highlight from different perspectives, has been effectively incorporated into the model's computation, contributing positively to survival prediction. More visualization examples are provided in the link of code resource page for reference (<https://github.com/Houjiixin123/SurvPGC>).

Discussion

Traditional survival analysis often uses the Cox proportional hazards model to estimate survival times and assess risk factors³⁸. With advancements in data analysis, researchers have increasingly incorporated NLP models to include unstructured text from electronic medical records in survival prediction³⁹. Structured clinical data, such as tabular statistics, can also be converted into text to add contextual information, improving model interpretability⁴⁰. This highlights the growing importance of utilizing clinical text effectively for predictive tasks.

In multimodal learning, incorporating clinical information is an effective way to further enhance predictive accuracy in some tasks. For example, our previous study integrated clinical data with image-based risk scores and transcriptomic features, resulting in a more accurate Cox model⁴¹. J. Yang et al. employed ResNet50 to extract features from pathological images and fused selected clinical characteristics with image features using multimodal compact bilinear pooling (MCB)⁴², an outer product-based method that achieves improved prediction accuracy for breast cancer recurrence and metastasis with relatively low computational complexity⁴³. However, these approaches often exhibit limited precision in handling clinical information, and discrepancies in feature dimensions between clinical and other modalities may constrain overall predictive performance.

As outlined in the Introduction, integrating clinical information into multimodal learning primarily involves two key challenges: encoding strategies and fusion methodologies. These are generally categorized into the three types illustrated in Fig. 1. To evaluate the impact of different encoding and fusion workflows on the prognostic predictive capability of clinical information, we conducted multimodal fusion (clinical information and

pathology images) experiments across three datasets. The results are summarized in Table 4, and detailed descriptions of Approaches a, b, and c are provided in Section 5 of the supplementary materials. It is evident that the proposed workflow, foundation model encoding combined with bidirectional cross-attention, achieves the highest C-index under conditions that enable feature-level interaction.

The impact of the diversity of templates used to generate clinical prompts on the performance of the model is also a topic worthy of discussion. Inspired by research on image-text contrastive learning²¹, we design text templates to transform clinical information into sentences that capture contextual semantics. Previous studies have explored text augmentation in contrastive learning, where the captions associated with images were rewritten using large language models. During the model training process, either the original or the rewritten text was randomly selected, demonstrating that increasing text diversity can enhance model performance⁴⁴. Therefore, for each characteristic, we designed description text templates centered on the key information and generated N_t synonymous templates using GPT-4o mini. One template was randomly chosen for each characteristic, producing six descriptive sentences per case. Templates are detailed in Supplementary Table 8. To determine the optimal number of synonymous sentences, we conducted experiments to compare the survival prediction performance of the MLP model with clinical embeddings as input when $N_t = 1, 3, 5, 10, 20$. The partial experimental results are presented in Table 5. In addition, we visualized the clinic embeddings generated based on N_t templates through the t-Distributed Stochastic Neighbor Embedding (t-SNE) method. The results indicate that when $N_t = 1$, the embeddings from patients with different survival durations exhibit the best discrimination and clustering performance. The partial visualization results are shown in Fig. 5. Considering both the C-index and the t-SNE visualization results, we ultimately selected $N_t = 1$ to generate clinical texts for multimodal survival prediction in the experiments. The complete experimental results are presented in the supplementary materials Section 1.

Foundation models, trained on vast datasets, demonstrate strong generalization and robustness, enabling one-shot or few-shot task performance. These models are widely applied in data preprocessing to enhance efficiency. In medicine, various foundation models handle different data modalities, such as scFoundation for gene sequencing data⁴⁵, UNI for whole-slide images⁴⁶, CONCH²⁴ and MUSK⁴⁷ for image-text pairs. Across diverse

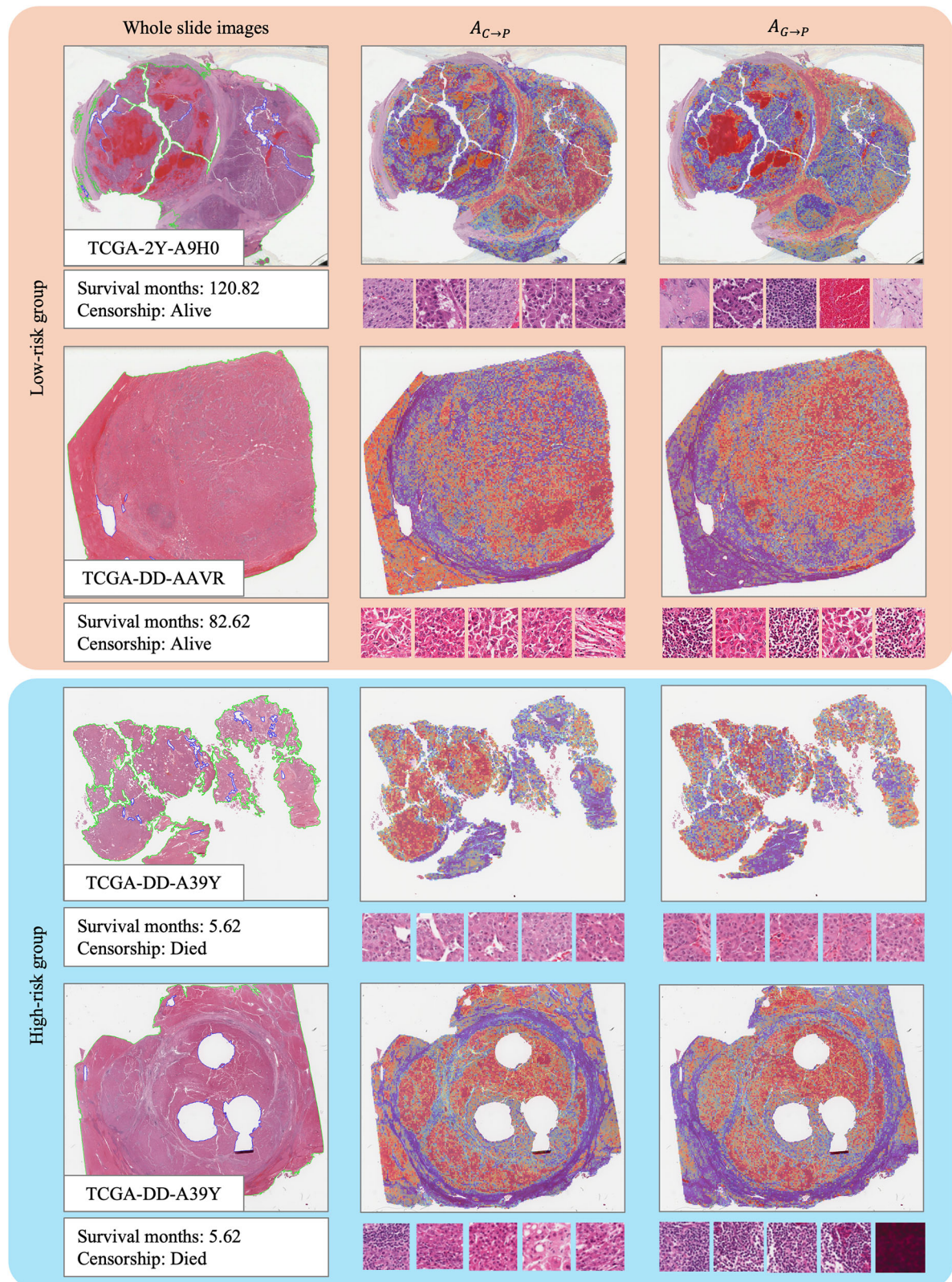


Fig. 4 | Interpretability visualization of LIHC patients. The left column displays thumbnail images of WSIs. The middle column shows heatmaps of the attention matrix $A_{C \rightarrow P}$, which highlights tissue regions of interest based on clinical characteristics. The right column presents heatmaps of the attention matrix $A_{G \rightarrow P}$, which

identifies tissue regions of interest based on transcriptome data. For each case, the top patches selected according to cumulative attention values of $A_{C \rightarrow P}$ and $A_{G \rightarrow P}$ are displayed beneath the corresponding heatmap.

task scenarios, these foundation models function as efficient and rapid data preprocessing tools, capturing key information from the data. Moreover, the appropriate application of foundation models can address the challenge of insufficient data encountered during the training of smaller, specialized deep learning models.

Modality fusion is a critical issue in multimodal learning. Based on the stages of fusion, it can be categorized into early fusion, late fusion, and intermediate fusion¹⁶. Early fusion is constrained by data forms, while late fusion lacks inter-modal interaction. Intermediate fusion unifies modalities into a shared high-dimensional space, utilizing multimodal advantages effectively. Most studies have conducted in-depth research on multimodal fusion methods⁴⁸, ranging from simple concatenation or element-based computational fusion⁴⁹, to bilinear pooling²⁷ and related improvements to reduce computational load¹², and then to fusion approaches based on attention mechanisms⁵⁰ and Transformers^{28,29,51}. The simulation capabilities of these methods for inter-modal connections have gradually increased. According to the characteristics of the data and the processing procedures, there are also studies that adopt a combination of multiple fusion methods^{7,52}.

The attention mechanism has become a focal point in multimodal research since the introduction of the Transformer architecture⁵³. It enables the model to assign adaptive weights to different parts of the input, thereby emphasizing information that is most relevant and critical to the current task. In multimodal settings, attention scores are computed based on inputs from different modalities, meaning that the computation explicitly considers cross-modal information. Attention-based feature fusion, in particular, excels at capturing contextual relationships across modalities and has become a widely adopted approach for feature-level fusion^{26,28,54} with demonstrated superior performance over other fusion strategies in various multimodal tasks⁴⁸.

Meanwhile, representation learning is inherently intertwined with modality fusion, especially in end-to-end deep learning frameworks^{55,56}. In this study, we designed coherent strategies for both representation learning and modality fusion. We proposes a workflow that uses text templates to encode contextual information and effectively incorporates it into the model. Then, we adopted an attention-based fusion strategy and analyzed the attention matrix during the fusion process, thereby establishing a connection between model computation and visual interpretation.

Additionally, this study has several limitations. First, the text templates and selection of clinical characteristics are manually set. An automated method for generating clinical text templates may facilitate broader application of the workflow. Second, the application of the foundation model was not extensively explored. Different embedding encoders can influence the

predictive performance of the model, so choosing a more suitable encoder for specific tasks remains an important consideration. While the foundation model emphasizes generalization across diverse data, enhancing the discriminative ability of embeddings for specific modalities and tasks is also crucial. Third, although SurvPGC achieves a relatively high C-index and demonstrates better risk ranking performance, there remains potential for improvement in the accuracy of predictions at specific time points. Future research should focus on enhancing the model’s ability to precisely estimate risks at distinct time intervals, particularly for longer survival durations.

This paper proposes a prognosis prediction workflow based on multimodal data, offering several key advantages: (1) the design of templates enables structured clinical data to capture richer contextual information, thereby enhancing the semantic representation of high-dimensional features; (2) utilizing foundation models for encoding high-dimensional features across different modalities simplifies the process while ensuring robust diagnostic discrimination; (3) the visualization and comparison of pathological tissue regions highlighted by both clinical information and transcriptome provide insights of prognosis. Our model has been validated using TCGA-LIHC, TCGA-BRCA and TCGA-COADREAD, demonstrating superior performance compared to previous methods. In future work, we aim to conduct more in-depth research on the extraction and selection of multi-modal features, extracting valuable information from limited data to better achieve target tasks and meet clinical demands for cancer computer-aided diagnosis.

Methods

Multimodal embeddings encoded by foundation models

Clinical information. Given a set of clinical characteristics of n_c values (text or numeric), denoted as $c \in \mathbb{R}^{n_c}$. Considering the completeness of the characteristics of the clinical records and their potential relationship with prognosis, six clinic characteristics (age, sex, tumor stage, diagnostic criteria version, and tumor subtype) were selected, thus $n_c = 6$. A pre-trained visual-language foundation model, CONCH²⁴, was used to encode text, resulting in clinical embedding $X_C \in \mathbb{R}^{n_c \times d_c}$, where d_c is the dimension of each clinic embedding. The process of this part is shown in Fig. 2a.

Transcriptome expression file. Transcriptome data, including n_g expression measurements for approximately 60,000, denoted as $g \in \mathbb{R}^{n_g}$, was analyzed. To reduce redundancy, we applied the method described in SurvPath²⁹, selecting 331 survival-related pathways from Reactome and Hallmarks repositories, identifying 4999 genes associated with these pathways, thus $n_g = 4999$. We then employed pre-trained scFoundation⁴⁵ to encode the selected data, resulting in genomic embedding for each sample

Table 3 | Modality attribution percentages calculated based on integrated gradients

| Modality | TCGA-LIHC | TCGA-BRCA | TCGA-COADREAD |
|-------------------|---------------|---------------|---------------|
| Genomic/Pathology | 0.554 ± 0.065 | 0.822 ± 0.024 | 0.339 ± 0.045 |
| Clinic/Pathology | 0.446 ± 0.065 | 0.178 ± 0.024 | 0.661 ± 0.045 |

Table 5 | Impact of clinical template variety on MLP performance (C-index)

| Number of templates | TCGA-LIHC | TCGA-BRCA | TCGA-COADREAD |
|---------------------|---------------|---------------|---------------|
| 1 | 0.599 ± 0.052 | 0.684 ± 0.101 | 0.640 ± 0.123 |
| 5 | 0.523 ± 0.047 | 0.613 ± 0.081 | 0.606 ± 0.085 |
| 20 | 0.546 ± 0.040 | 0.523 ± 0.072 | 0.575 ± 0.072 |

Table 4 | Comparison of clinic information integration methods

| Methods | Clinical information coding dimensions | Information interaction between modalities | TCGA-LIHC | TCGA-BRCA | TCGA-COADREAD |
|-------------------------------------|--|--|---------------|---------------|---------------|
| Approach a (Late fusion) | Low | No | 0.641 ± 0.030 | 0.674 ± 0.059 | 0.525 ± 0.083 |
| Approach b (Mixed fusion) | Low | Yes | 0.634 ± 0.039 | 0.632 ± 0.033 | 0.520 ± 0.104 |
| Approach c (Intermediate fusion) | High | Yes | 0.635 ± 0.029 | 0.662 ± 0.047 | 0.581 ± 0.085 |
| SurvPC (ours) (Intermediate fusion) | High | Yes | 0.650 ± 0.040 | 0.668 ± 0.054 | 0.670 ± 0.129 |

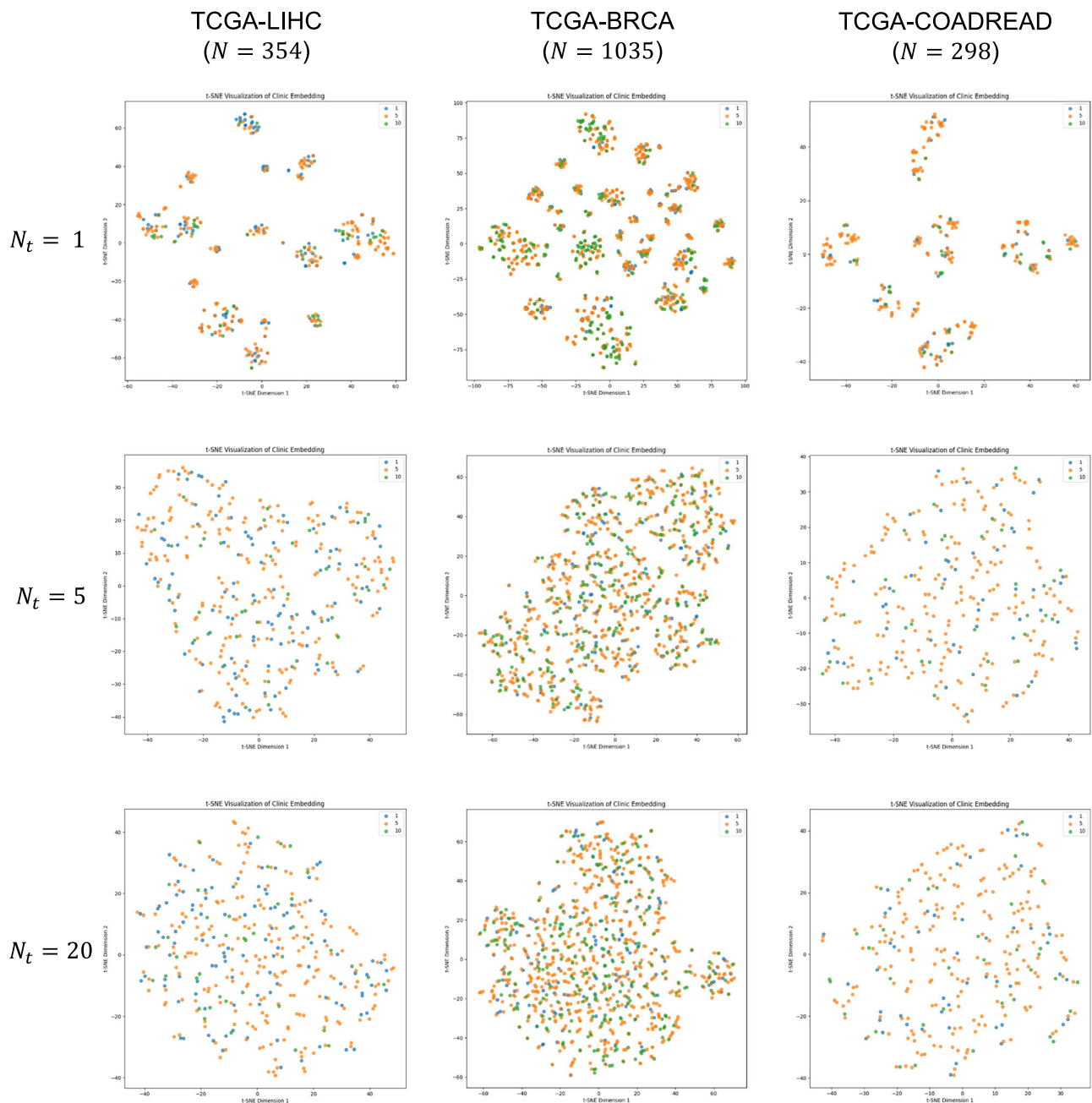


Fig. 5 | The t-SNE visualization results of the clinic embeddings generated with different quantity of synonymous templates. Blue points represent patients with a recorded survival time of less than 1 year; yellow points represent patients with survival times between 1 and 5 years; and green points represent patients with

survival times exceeding 5 years. N_t denotes the quantity of synonymous templates during the clinical prompt generation process. N denotes the sample size of each dataset.

represented as $X_G \in \mathbb{R}^{n_g \times d_g}$, where the output consists of $n_g = 4$ vectors (as per the default parameter settings of scFoundation) and d_g denotes the dimension of each genomic embedding. The process of this part is shown in Fig. 2b.

Pathology images. We selected high-quality paraffin-embedded hematoxylin and eosin-stained digital WSIs. CLAM⁵⁷ was used to segment tissue areas, from which 256×256 (pixelxpixel) non-overlapping patches were cut out at $20\times$ magnification. Depending on the size and quality of each slide, a set of n_p patches (with order of magnitude ranging from 1000 to 10,000) can be cut out, denoted as $P = \{p_1, p_2, \dots, p_{n_p}\}$. UNI⁴⁶, a pre-trained foundation model, extracted features to generate pathology image embeddings for each patch, denoted as $x_{p_i} \in \mathbb{R}^{d_p}$, where d_p denotes the

dimension of each patch. Thus, the embeddings of each sample contained n_p patches can be represented as $X_p \in \mathbb{R}^{n_p \times d_p}$. The process is shown in Fig. 2c.

Multimodal fusion based on cross-attention mechanism

First, we map the embeddings of each modality to the same dimension d by utilizing a learnable linear layer, respectively, obtaining $X_C \in \mathbb{R}^{n_c \times d}$, $X_G \in \mathbb{R}^{n_g \times d}$, $X_p \in \mathbb{R}^{n_p \times d}$. Subsequently, taking pathology and clinical information as an example, the embeddings of the two modalities were concatenated to obtain $X \in \mathbb{R}^{(n_c+n_p) \times d}$, and the Query (Q), Key (K) and Value (V) vectors are obtained by using the self-attention mechanism for linear projection (Eqs. (1)–(3)). The dimension of the learnable weight matrixes W are: $W_Q \in \mathbb{R}^{d \times d}$, $W_K \in \mathbb{R}^{d \times d}$, $W_V \in \mathbb{R}^{d \times d}$, where d

denotes the dimension of the embeddings after projection.

$$Q = XW_Q (Q \in \mathbb{R}^{(n_c+n_p) \times d'}) \quad (1)$$

$$K = XW_K (K \in \mathbb{R}^{(n_c+n_p) \times d'}) \quad (2)$$

$$V = XW_V (V \in \mathbb{R}^{(n_c+n_p) \times d'}) \quad (3)$$

By utilizing the embeddings from two modalities to compute the attention matrix A (referred to Eq. (4)), we can effectively filter out the information most relevant to the downstream task.

$$A = QK^T / \sqrt{d'} \quad (4)$$

According to G. Jaume et al.²⁹, bidirectional cross-attention fusion enhances the model's ability to utilize task-relevant information, improving predictive performance. We designed a dual-path cross-modality attention fusion model (Fig. 2d), the bidirectional cross-attention mechanism was applied separately to the fusion of clinical information and pathology, as well as the fusion of genomics and pathology. The attention matrixes for clinical-pathology fusion ($A_{C \leftrightarrow P}$) and genomic-pathology fusion ($A_{G \leftrightarrow P}$) were calculated. The fused vector X_{Attm} in each path can be calculated as Eq. (5), where σ is the softmax function:

$$X_{Attm} = \sigma(QK^T / \sqrt{d'})V = \sigma \begin{pmatrix} A_{M \rightarrow M} & A_{M \rightarrow P} \\ A_{P \rightarrow M} & A_{P \rightarrow P} \end{pmatrix} \begin{pmatrix} V_M \\ V_P \end{pmatrix} (M = \{Genomics(G), Clinic(C)\}) \quad (5)$$

In more detail, the aforementioned operation obtains X by concatenation and maps it to the Q, K, V vectors. Therefore, the corresponding Q, K, V vectors of specific modality can be directly chunked to calculate the attention matrix. To simplify the model and focus on cross-modal interactions, we omitted the within-modality attention matrix $A_{P \rightarrow P}$, as suggested in ref.²⁹. Taking the fusion of clinical information and pathology as an example, the bidirectional cross-attention process is as Eqs. (6)–(9).

$$A_{C \rightarrow C} = Q_c K_c^T / \sqrt{d'} (A_{C \rightarrow C} \in \mathbb{R}^{n_c \times n_c}) \quad (6)$$

$$A_{C \rightarrow P} = Q_c K_p^T / \sqrt{d'} (A_{C \rightarrow P} \in \mathbb{R}^{n_c \times n_p}) \quad (7)$$

$$A_{P \rightarrow C} = Q_p K_c^T / \sqrt{d'} (A_{P \rightarrow C} \in \mathbb{R}^{n_p \times n_c}) \quad (8)$$

$$\hat{A}_{C \rightarrow P} = A_{C \rightarrow C} \oplus A_{C \rightarrow P} (\hat{A}_{C \rightarrow P} \in \mathbb{R}^{n_c \times (n_c+n_p)}) \quad (9)$$

Multiply the bidirectional cross-attention matrixes $\hat{A}_{C \rightarrow P}, A_{P \rightarrow C}$ with the V vectors of each modality to obtain the fused vectors $X_{C \rightarrow P}, X_{P \rightarrow C}$ (Eqs. (10)–(11)). After passing through the forward propagation layer with normalization operations, the mean of the fused vector representation is calculated and obtained $\bar{X}_{C \rightarrow P} \in \mathbb{R}^{d'}$ and $\bar{X}_{P \rightarrow C} \in \mathbb{R}^{d'}$.

$$X_{C \rightarrow P} = \hat{A}_{C \rightarrow P} V (X_{C \rightarrow P} \in \mathbb{R}^{n_c \times d'}) \quad (10)$$

$$X_{P \rightarrow C} = A_{P \rightarrow C} V_C (X_{P \rightarrow C} \in \mathbb{R}^{n_p \times d'}) \quad (11)$$

The fusion process of genomics and pathology is similar. Eventually, 4 identical-dimensional fusion vectors with bidirectional attention fusion can be obtained: $\bar{X}_{C \rightarrow P}, \bar{X}_{P \rightarrow C}, \bar{X}_{G \rightarrow P}, \bar{X}_{P \rightarrow G}$. Concatenate the 4 fused vectors to obtain X_{Fusion} (Eq. (12)), as the input of the final model decision layer.

$$X_{Fusion} = X_{C \rightarrow P} \oplus X_{P \rightarrow C} \oplus X_{G \rightarrow P} \oplus X_{P \rightarrow G} (X_{Fusion} \in \mathbb{R}^{4d'}) \quad (12)$$

Loss function

We followed previous works^{12,27,58} and adopted the negative log-likelihood (NLL) loss to avoid the deviation caused by censor rate, which transforms survival prediction into a classification task by predicting the probability that survival time falls into different time intervals. As the previous work^{27–29}, t denoted the survival time, c denoted censorship status of patients, $c = 0$ represents the observed death at time point t of patient's follow-up records, while $c = 1$ represents that t is the last follow-up records and the time point of death is unknown. Based on the longest duration record, we defined i intervals, which be represents as $[0, t_1), [t_1, t_2), \dots, [t_{i-1}, t_i)$. The model's output \hat{y} is mapped through sigmoid to estimate the probability of death in time intervals, which represents as $\hat{y} = [y_1, y_2, \dots, y_i]$. Based on \hat{y} , the probability of survival in each interval can be computed referring to Eq. (13). Finally, by summing the negative of interval survival, a patient-level survival risk description is defined.

$$s = \prod_{i=1}^i (1 - y_i) (s \in \mathbb{R}^i) \quad (13)$$

To ensure that the model captures diverse regions of WSIs while avoiding redundancy, we introduced a cosine similarity loss function. This loss function reduces the similarity between the fused embeddings output by the two cross-attention modules, guiding the model to attend to complementary information from each modality. The loss function is as Eq. (14), where β is the weight of the $loss_{cosine}$.

$$loss = (1 - \beta) \times loss_{NLL} + \beta \times loss_{cosine} \quad (14)$$

Since the improvement in model performance brought by the similarity loss function is not stable, it was not included in the results of Table 1. The discussion about the similarity loss function is provided in Section 2 of supplement materials.

Evaluation matrixes

To evaluate the model's prognosis performance, we used multiple metrics: C-index evaluates agreement between predicted and actual results and iAUC evaluates performance at several time points (at 1-year, 3-year and 5-year). KM curve visually shows survival probabilities for risk groups, while log-rank tests assess survival differences, with $p < 0.05$ considered significant.

Data availability

The data analyzed in the study is from public datasets online, including TCGA (<http://portal.gdc.cancer.gov/projects/TCGA-LIHC> (accessed on 24 October 2024)), <http://portal.gdc.cancer.gov/projects/TCGA-BRCA> (accessed on 21 December 2024), <http://portal.gdc.cancer.gov/projects/TCGA-COAD> (accessed on 30 June 2025), <http://portal.gdc.cancer.gov/projects/TCGA-READ> (accessed on 30 June 2025)) and UCSC Xena (<https://xenabrowser.net/datapages/> (accessed on 26 December 2024)).

Code availability

Source codes are available at: <https://github.com/Houjiixin123/SurvPGC>.

Received: 22 May 2025; Accepted: 4 December 2025;

Published online: 27 December 2025

References

1. Stephen P. J. Survival analysis. *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK* **42**, 54–56 (2005).
2. Tran, K. A. et al. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Med.* **13**, 152 (2021).
3. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* **15**, 81–94 (2018).

4. Lee, C., Zame, W. R., Yoon, J. & Der Schaar, M. Van. DeepHit: A deep learning approach to survival analysis with competing risks. *Proc. AAAI Conf. Artif. Intell.* **32**, 2314–2321 (2018).
5. Katzman, J. L. et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18**, 24 (2018).
6. Zhu, X., Yao, J. & Huang, J. Deep convolutional neural network for survival analysis with pathological images. *2016 IEEE Int. Conf. Bioinform. Biomed. (BIBM)* **2016**, 544–547 (2016).
7. Wang, Z., Li, R., Wang, M. & Li, A. GPDBN: deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction. *Bioinformatics* **37**, 2963–2970 (2021).
8. Mobadersany, P. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl. Acad. Sci. USA*. **115**, E2970–E2979 (2018).
9. Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N. & Huang, J. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Med. Image Anal.* **65**, 101789 (2020).
10. Zadeh Shirazi, A. et al. DeepSurvNet: deep survival convolutional network for brain cancer survival rate classification based on histopathological images. *Med. Biol. Eng. Comput.* **58**, 1031–1045 (2020).
11. Cui, C. et al. Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. *Prog. Biomed. Eng.* **5**, <https://doi.org/10.1088/2516-1091/acc2fe> (2023).
12. Chen, R. J. et al. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans. Med. Imaging* **41**, 757–770 (2022).
13. Cheng, J. et al. Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis. *Cancer Res.* **77**, e91–e100 (2017).
14. Lv, Z., Lin, Y., Yan, R., Wang, Y. & Zhang, F. TransSurv: transformer-based survival analysis model integrating histopathological images and genomic data for colorectal cancer. *IEEE/ACM Trans. Comput. Biol. Bioinforma* **20**, 3411–3420 (2023).
15. Duan, J., Xiong, J., Li, Y. & Ding, W. Deep learning based multimodal biomedical data fusion: an overview and comparative review. *Inf. Fusion* **112**, 102536 (2024).
16. Steyaert, S. et al. Multimodal data fusion for cancer biomarker discovery with deep learning. *Nat. Mach. Intell.* **5**, 351–362 (2023).
17. Pölsterl, S., Wolf, T. N., & Wachinger, C. Combining 3D Image and Tabular Data via the Dynamic Affine Feature Map Transform. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. Lecture Notes in Computer Science* **12905**, 688–698 (2021).
18. Yoo, Y. et al. Deep learning of brain lesion patterns and user-defined clinical and MRI features for predicting conversion to multiple sclerosis from clinically isolated syndrome. *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.* **7**, 250–259 (2019).
19. Holste, G. et al. End-to-end learning of fused image and non-image features for improved breast cancer classification from MRI. *Proc. IEEE Int. Conf. Comput. Vis.* **2021**, 3287–3296 (2021).
20. Thatoi, P., Choudhary, R., Shiwlani, A., Qureshi, H. A. & Kumar, S. Natural language processing (NLP) in the extraction of clinical information from electronic health records (EHRs) for. *Cancer Prognosis* **10**, 2676–2694 (2023).
21. Radford, A. et al. Learning transferable visual models from natural language supervision. *Proc. Mach. Learn. Res.* **139**, 8748–8763 (2021).
22. Baltrusaitis, T., Ahuja, C. & Morency, L. P. Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 423–443 (2019).
23. Lotfollahi, M. Toward learning a foundational representation of cells and genes. *Nat. Methods* **21**, 1416–1417 (2024).
24. Lu, M. Y., Chen, B. & Williamson, D. F. K. A visual-language foundation model for computational pathology. *Nat. Med.* **30**, 863–874 (2024).
25. Pai, S. et al. Foundation model for cancer imaging biomarkers. *Nat. Mach. Intell.* **6**, 354–367 (2024).
26. Li, Z., Jiang, Y., Lu, M., Li, R. & Xia, Y. Survival prediction via hierarchical multimodal co-attention transformer: a computational histology-radiology solution. *IEEE Trans. Med. Imaging* **42**, 2678–2689 (2023).
27. Chen, R. J. et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* **40**, 865–878.e6 (2022).
28. Chen, R. J. et al. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. *Proceedings of the IEEE/CVF International Conference on Computer Vision* **2021**, 3995–4005 (2021).
29. Jaume, G. et al. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* **2024**, 11579–11590 (2024).
30. Huang, S. C., Pareek, A., Seyyedi, S., Banerjee, I. & Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digit. Med.* **3**, 136 (2020).
31. Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**, 187–202 (1972).
32. Bishop, C. M. *Neural Networks for Pattern Recognition* (Oxford University Press, 1995).
33. Klambauer, G. et al. Self-normalizing neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* **2017**, 972–981 (2017).
34. Ilse, M., Tomczak, J. M. & Welling, M. Attention-based deep multiple instance learning. *Proceedings of the 35th International Conference on Machine Learning, PMLR* **80**, 2127–2136 (2018).
35. Shao, Z. et al. TransMIL: transformer based correlated multiple instance learning for whole slide image classification. *Adv. Neural Inf. Process. Syst.* **3**, 2136–2147 (2021).
36. Zadeh Shirazi, A. et al. A deep convolutional neural network for segmentation of whole-slide pathology images identifies novel tumour cell-perivascular niche interactions that are associated with poor survival in glioblastoma. *Br. J. Cancer* **125**, 337–350 (2021).
37. Li, H. et al. Collagen fiber orientation disorder from H&E images is prognostic for early stage breast cancer: clinical trial validation. *npj Breast Cancer* **7**, 104 (2021).
38. Moncada-Torres, A., van Maaren, M. C., Hendriks, M. P., Siesling, S. & Geleijnse, G. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Sci. Rep.* **11**, 6968 (2021).
39. Jee, J. et al. Automated real-world data integration improves cancer outcome prediction. *Nature* **636**, 728–736 (2024).
40. Zhou, K., Yang, J., Loy, C. C. & Liu, Z. Learning to prompt for vision-language models. *Int. J. Comput. Vis.* **130**, 2337–2348 (2022).
41. Hou, J., Jia, X., Xie, Y. & Qin, W. Integrative Histology-Genomic Analysis Predicts Hepatocellular Carcinoma Prognosis Using Deep Learning. *Genes* **13**, 1770 (2022).
42. Fukui, A. et al. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* **2016**, 457–468 (2016).
43. Yang, J. et al. Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. *Comput. Struct. Biotechnol. J.* **20**, 333–342 (2022).
44. Fan, L., Krishnan, D., Isola, P., Katabi, D. & Tian, Y. Improving CLIP training with language rewrites. *Adv. Neural Inform. Process. Systems* **36**, 35544–35575 (2023).

45. Hao, M. et al. Large-scale foundation model on single-cell transcriptomics. *Nat. Methods* **21**, 1481–1491 (2024).
46. Chen, R.J., Ding, T., Lu, M.Y. et al. Towards a general-purpose foundation model for computational pathology. *Nat Med* **30**, 850–862 (2024).
47. Xiang, J., Wang, X. & Zhang, X. A vision-language foundation model for precision oncology. *Nature* **638**, 769–778 (2025).
48. Schouten, D. et al. Navigating the landscape of multimodal AI in medicine: a scoping review on technical challenges and clinical applications. *Med. Image Anal.* **105**, 103621 (2025).
49. Joo, S. et al. Multimodal deep learning models for the prediction of pathologic response to neoadjuvant chemotherapy in breast cancer. *Sci. Rep.* **11**, 18800 (2021).
50. Vanguri, R. S. et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. *Nat. Cancer* **3**, 1151–1164 (2022).
51. Wang, Z., Yu, L., Ding, X., Liao, X. & Wang, L. Shared-specific feature learning with bottleneck fusion transformer for multi-modal whole slide image analysis. *IEEE Trans. Med. Imaging* **42**, 3374–3383 (2023).
52. Zheng, S. et al. Multi-modal graph learning for disease prediction. *IEEE Trans. Med. Imaging* **41**, 2207–2216 (2022).
53. Vaswani, A., et al. Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017)* **30**, 5998–6008 (2017).
54. Shao, Z. et al. HVTSurv: hierarchical vision transformer for patient-level survival prediction from whole slide image. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 2209–2217 (2023).
55. Wu, X., Shi, Y., Wang, M. & Li, A. CAMR: cross-aligned multimodal representation learning for cancer survival prediction. *Bioinformatics* **39**, btad025 (2023).
56. Zhou, H., Zhou, F. & Chen, H. Cohort-individual cooperative learning for multimodal cancer survival analysis. *IEEE Trans. Med. Imaging* **44**, 656–667 (2025).
57. Lu, M. Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
58. Zadeh, S. G. & Schmid, M. Bias in cross-entropy-based training of deep survival networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 3126–3137 (2021).

Acknowledgements

This study was supported in part by the National Natural Science Foundation of China (No. 62271475), Shenzhen-Hong Kong Joint Lab on Intelligence Computational Analysis for Tumor Imaging (E3G111), the Shenzhen Science and Technology Program of China (JCYJ20220818101804009), Shenzhen-Hong Kong-Macao Science and Technology Plan Project (Category C Project) under Shenzhen Municipal Science and Technology Innovation

Commission (SGDX20230821092359002), Guangdong Youth Talent program (2024TQ08A386), and the Youth Innovation Promotion Association CAS (2022365). This study acknowledges the public datasets supported by The Cancer Genome Atlas.

Author contributions

Conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing—original draft: J.H.; funding acquisition, resources: W.Q.; project administration: W.Q. and Y.X.; supervision: W.Q., C.L., and Y.X.; writing—review and editing: W.Q. and R.Z.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains

supplementary material available at

<https://doi.org/10.1038/s41746-025-02257-y>.

Correspondence and requests for materials should be addressed to Wenjian Qin.

Reprints and permissions information is available at

<http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025