Appendix A Datasets

To evaluate the performance of our framework, we collected three multimodal datasets, including GastricRes, GastricSur, and TCGA-STAD. In this section, we introduce the detail of each dataset.

A.1 GastricRes

This dataset was collected from four prominent medical hospitals in China: Yunnan Cancer Hospital (Kunming, China), Shanxi Cancer Hospital (Taiyuan, China), Sichuan Cancer Hospital (Chengdu, China), and the Sixth Affiliated Hospital of Sun Yat-sen University (Guangzhou, China). This dataset encompasses comprehensive information from 698 patients who were diagnosed with gastric cancer and underwent Neoadjuvant Chemotherapy (NACT) treatment. More details can be found in Figure A1. The dataset consists of three modalities: clinical records, whole slide images (WSI), and computed tomography images (CT). Tumor Regression Grade (TRG) is a widely used grading system employed to evaluate the extent of tumor regression and the response to NACT in patients with gastric cancer. TRG0 represents the absence of residual tumor cells observed under microscopic examination, indicating a pathologically complete response. TRG1 indicates the presence of only single cells or small clusters of cancer cells visible under the microscope. TRG2 denotes the presence of residual cancer cells surrounded by fibrosis. TRG3 indicates significant fibrosis predominating over cancer cells. In this study, we define TRG(0-2) as a good responder, indicating favorable treatment response, while TRG3 is classified as a non-responder, indicating limited response to therapy. There are 325 good responders and 373 non-responders in the GastricRes dataset.

A.2 GastricSur

This dataset was collected from two prominent medical hospitals in China: the First Affiliated Hospital of Kunming Medical University (Kunming, China) and Yunnan Cancer Hospital. It comprises comprehensive data from a cohort of 801 patients who were diagnosed with gastric cancer and subsequently underwent surgical resection. Detailed information regarding the dataset can be found in Figure A1. Similar to the GastricRes dataset, this collection encompasses three distinct modalities: clinical records, whole slide images (WSI), and computed tomography scans (CT). Throughout the follow-up period, there are 286 patients who died.

A.3 TCGA-STAD

This dataset was obtained from The Cancer Genome Atlas (TCGA) database. It contains data from 400 patients diagnosed with gastric cancer. Different from GastricRes and GastricSur datasets, the modalities involved in this dataset are: clinical records, WSI, and RNA-seq. The clinical records are downloaded from the LinkedOmics. The WSI are downloaded from the GDC Data Portal. The RNA-seq data is downloaded from the cBioPortal. Specifically, all patients in this dataset have clinical records, while 363 patients have WSI and 374 patients have RNA-seq.

Appendix B Method details

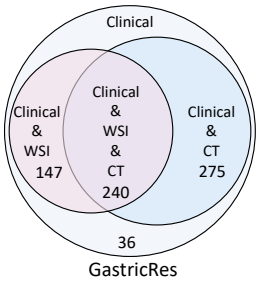
B.1 Loss functions

There are two tasks in this study: NACT response prediction and survival prediction. The former is a classification task, while the latter is a regression task. We employ different loss functions for these two tasks. In addition to the loss function defined for specific tasks, there is also a loss function defined for knowledge distillation.

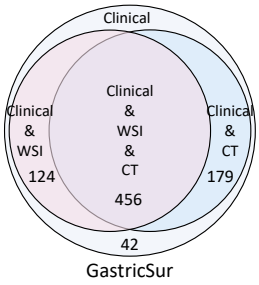
NACT response prediction. We employ the cross-entropy function to constrain the NACT response prediction task. The cross-entropy loss function is defined as follows:

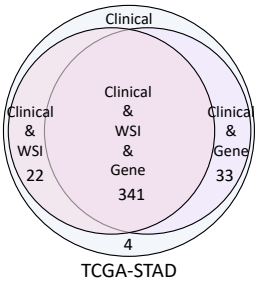
$$D_{\text{ce}}(\mathcal{Y}, \hat{\mathcal{Y}}) = -\sum_{i=1}^{C} \mathcal{Y}_i \log(\hat{\mathcal{Y}}_i) \tag{B1}$$

where $\mathcal{Y}$ is the ground truth label, $\hat{\mathcal{Y}}$ is the predicted label, and log is the natural logarithm. There are two types of knowledge distillation: feature-level distillation and response-level distillation. Feature-level distillation ensures that the student models generate representations similar to those of the teacher model, while response-level distillation focuses on ensuring that the student models produce predictions comparable to those



GastricRes	Center	Clinical Records				WSI				CT			
		TRG0	TRG1	TRG2	TRG3	TRG0	TRG1	TRG2	TRG3	TRG0	TRG1	TRG2	TRG3
	YN	23	16	74	99	15	14	46	61	20	16	67	91
	SX	19	17	26	236	9	9	6	65	15	17	24	222
	SC	1	7	31	21	1	3	23	12	0	3	24	16
	YS	14	15	82	17	13	14	80	16	0	0	0	0
	Total	57	55	213	373	38	40	155	154	35	36	115	329





GastricSur	Center	Clinical	WSI	CT
	KM	283	260	222
	YN	518	320	413
	Total	801	580	635

Center	Clinical	WSI	Genomics
TCGA-STAD	400	363	374

Fig. A1 Details of datasets used in this study. There are three datasets involved in this study: GastricRes dataset for response prediction, GastricSur dataset for survival analysis, and TCGA-STAD dataset for survival analysis. GastricRes comprises information from 698 patients diagnosed with gastric cancer who underwent NACT treatment, including three modalities: clinical information, WSI, and CT. GastricSur consists of data from 801 patients diagnosed with gastric cancer who underwent surgical resection, including three modalities: clinical information, WSI, and CT. TCGA-STAD contains data from 400 patients diagnosed with gastric cancer, including three modalities: clinical information, WSI, and transcriptomics profiles.

of the teacher model. We employ the Kullback-Leibler divergence to constrain the knowledge distillation process. For feature-level distillation, the Kullback-Leibler divergence is defined as follows:

$$D_{\mathrm{KL}}(\mathcal{F}_t \| \mathcal{F}_s) = \sum_{i=1}^{K} (\sigma(\frac{\mathcal{F}_t}{T})_i \log(\frac{\sigma(\frac{\mathcal{F}_t}{T})_i}{\sigma(\frac{\mathcal{F}_s}{T})_i})) \quad \text{(B2)}$$

where $\mathcal{F}_t$ and $\mathcal{F}_s$ are the feature representations of the teacher model and student model, respectively. T is the temperature parameter. σ is the softmax function. K is the dimension of features. For response-level distillation, the Kullback-Leibler divergence is defined as follows:

$$D_{\mathrm{KL}}(\mathcal{Y}_t \| \mathcal{Y}_s) = \sum_{i=1}^{C} (\sigma(\frac{\mathcal{Y}_t}{T})_i \log(\frac{\sigma(\frac{\mathcal{Y}_t}{T})_i}{\sigma(\frac{\mathcal{Y}_s}{T})_i})) \quad \text{(B3)}$$

where $\mathcal{Y}_t$ and $\mathcal{Y}_s$ are the predicted labels of the teacher model and student model, respectively. T is the temperature parameter. σ is the softmax function. C is the number of classes. In the proposed "more-to-less" knowledge distillation, the teacher model is trained using all available modalities, while the student models are trained using a subset of modalities. The knowledge distillation process can be formulated as follows:

$$\mathcal{L}_{cls} = \sum_{S \in \mathcal{P}^*(\mathcal{X})} D_{\mathrm{ce}}(\mathcal{Y}, \hat{\mathcal{Y}}_{\mathcal{S}}) \quad \text{(B4)}$$

$$\mathcal{L}_{fea} = \sum_{S \in \mathcal{P}^*(\mathcal{X})} \mathcal{D}_{KL}(\mathcal{F}_{\mathcal{X}} \| \mathcal{F}_{\mathcal{S}}) \quad \text{(B5)}$$

$$\mathcal{L}_{res} = \sum_{S \in \mathcal{P}^*(\mathcal{X})} \mathcal{D}_{KL}(\hat{\mathcal{Y}}_{\mathcal{X}} \| \hat{\mathcal{Y}}_{\mathcal{S}}) \quad \text{(B6)}$$

where $\mathcal{X}$ is the set of all available modalities. $\mathcal{P}^*(\mathcal{X})$ represents the power set of $\mathcal{X}$ excluding the empty set. $\mathcal{F}_{\mathcal{X}}$ and $\mathcal{F}_{\mathcal{S}}$ are the representations generated by the teacher model and the student model, respectively. $\hat{\mathcal{Y}}_{\mathcal{S}}$ denotes the prediction made by the student model. The total loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha \mathcal{L}_{fea} + \beta \mathcal{L}_{res} \quad \text{(B7)}$$

where α and β are the hyperparameters.

Survival prediction. Due to the huge size of WSI, the model cannot be optimized with mini-batch manner. The alternative optimization strategy is to consider discrete time intervals and model each interval using an independent output. We leverage NLL (negative log-likelihood) survival loss [1] as the loss function of the survival prediction part. The loss functions for knowledge distillation are the same as those defined for the NACT response prediction task. The total loss function is defined as follows:

$$\mathcal{L}_{sur} = \sum_{S \in \mathcal{P}^*(\mathcal{X})} \mathcal{L}_{NLL}(T_S, l, c) \tag{B8}$$

$$\mathcal{L} = \mathcal{L}_{sur} + \alpha \mathcal{L}_{fea} + \beta \mathcal{L}_{res} \tag{B9}$$

where $\mathcal{L}_{NLL}$ is the NLL survival loss. $T_S = \{T_1, T_2, \cdots, T_i\}$ is the predicted hazard value at each discrete time interval. l is the discrete label. c is the censoring status. α and β are the hyperparameters.

B.2 Implementation details

The proposed framework was implemented using PyTorch 1.12.0 and Python 3.9.12. The complete source code can be accessed and reviewed at the following URL: https://github.com/FT-ZHOU-ZZZ/iMD4GC. To ensure reproducibility and minimize the impact of randomness on experimental results, the random seeds of both PyTorch and NumPy were set to a fixed value of 1, providing a consistent basis for comparison across experiments. In the clinical part, each record was fed into a 200-dimensional fully connected layer to obtain 200-dimension embedding. Similarly, in the pathology part, CTransPath [2] was employed to extract 768-dimensional patch features, which were then fed into a series of fully connected layers (768-256-200) to obtain compact and informative 200-dimensional embeddings. In the radiology part, the feature maps of penultimate layer from the ResNet-50 architecture were utilized. These feature maps were processed through a global average pooling layer to generate a concise 256-dimension feature vector. Subsequently, this feature vector was fed into a sequence of fully connected layers (256-256-200) to yield comprehensive and representative 200-dimensional embeddings. Regarding the genetic part, the expression level of each gene was fed into a 200-dimensional fully connected layer to obtain 200-dimension embedding. The dimension of all tokens was set to 200, which is equal to the dimension of gene embedding and expression embedding. It is widely recognized that augmenting the number of layers and parameters within a model generally amplifies its capacity, leading to improved performance. However, the availability of training samples in our dataset is restricted, thereby requiring a delicate balance between the number of network layers and the available training data to mitigate the risk of overfitting. In this study, we set the number of network layers to 2 for all datasets.

All methods were optimized with Adam optimizer. The initial learning rate is set to 0.0002. To further enhance the training process, a CosineAnnealingLR [3] scheduler was employed to dynamically adjust the learning rate, promoting convergence and preventing overfitting. Due to the different number of tokens, this framework cannot be trained with mini-batch manner. Therefore, the batch size is set to 1. The maximum number of epochs is set to 30. The weight decay is set to 0.00001. The temperature parameter in knowledge distillation is set to 4. More details can be found in the scripts of https://github.com/FT-ZHOU-ZZZ/iMD4GC. All methods involved in this study were trained and validated on a high-performance workstation with 8 NVIDIA RTX 3090 GPUs.

B.3 Ablation study

B.3.1 Impacts of feature fusion

As Ma *et al.* [4] pointed out, different fusion strategies do affect the robustness of Transformer models. Surprisingly, the optimal fusion strategy is dataset-dependent; there does not exist a universal strategy that works in general cases. To evaluate the impacts of feature fusion, we conduct a series of experiments to investigate the performance of different fusion strategies, including concatenation, multiplicative, tensor fusion [5], and low-rank tensor fusion [6], early fusion with transformer [7], and late fusion with transformer. The results are shown in Table B.1. It can be observed that the concatenation achieved the best performance across all datasets. The low-rank tensor fusion and early fusion with transformer also

Table B1 Impacts of feature fusoon strategies. This table presents the results of different feature fusion strategies on the GastricRes, GastricSur, and TCGA-STAD datasets. The best results are highlighted in **bold**, while the second-best results are underlined.

Method	GastricRes			GastricSur	TCGA-STAD
	AUC	ACC	F1-Score	C-Index	C-Index
ConcatWithLinear	**0.802**$_{\pm 0.050}$	**0.758**$_{\pm 0.043}$	**0.753**$_{\pm 0.027}$	**0.714**$_{\pm 0.008}$	**0.661**$_{\pm 0.058}$
Multiplicative	0.760$_{\pm 0.051}$	0.699$_{\pm 0.067}$	0.694$_{\pm 0.066}$	0.540$_{\pm 0.031}$	0.589$_{\pm 0.064}$
TensorFusion	0.759$_{\pm 0.029}$	0.588$_{\pm 0.051}$	0.462$_{\pm 0.129}$	0.691$_{\pm 0.018}$	0.656$_{\pm 0.048}$
LowRankTensorFusion	0.790$_{\pm 0.055}$	0.669$_{\pm 0.099}$	0.623$_{\pm 0.145}$	0.705$_{\pm 0.015}$	0.653$_{\pm 0.043}$
EarlyFusionTransformer	0.797$_{\pm 0.060}$	0.688$_{\pm 0.068}$	0.671$_{\pm 0.075}$	0.703$_{\pm 0.010}$	0.643$_{\pm 0.054}$
LateFusionTransformer	0.772$_{\pm 0.059}$	0.623$_{\pm 0.079}$	0.522$_{\pm 0.144}$	0.565$_{\pm 0.023}$	0.596$_{\pm 0.036}$

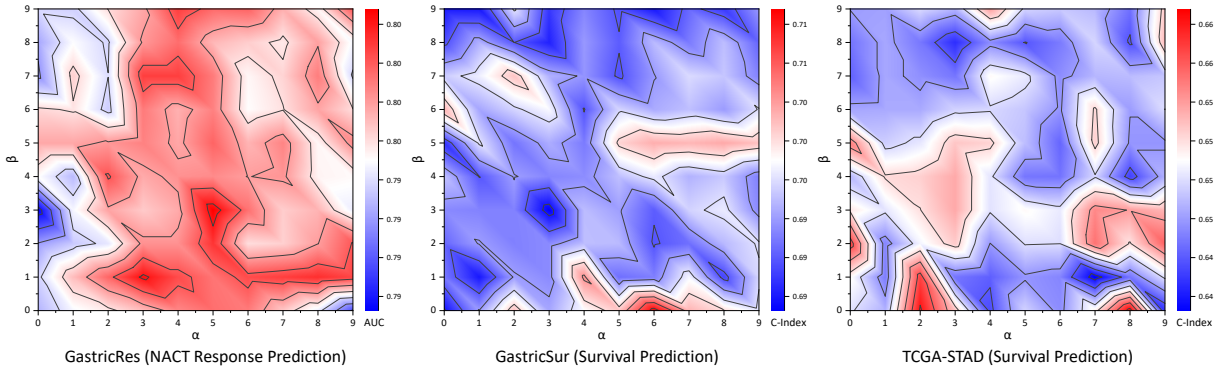


Fig. B2 Impacts of hyperparameters. These heatmap figures present the results of different hyperparameters on the GastricRes, GastricSur, and TCGA-STAD datasets. The x-axis represents α for feature-level distillation, while the y-axis represents α for response-level distillation.

achieved promising results and exhibited potential for further exploration.

B.3.2 Impacts of hyperparameters

There are two hyperparameters in our proposed framework: α and β. α is the weight of feature-level distillation, while β is the weight of response-level distillation. It is crucial to carefully select and balance these hyperparameters to enhance the model's performance based on the specific tasks and datasets. To investigate the impacts of these hyperparameters, we conduct a series of experiments to evaluate the performance of different hyperparameters. The results are shown in Figure B2. It can be observed that α and β have a significant impact on the performance of the proposed framework. The optimal hyperparameters are dataset-dependent. For the GastricRes dataset, the optimal hyperparameters are $\alpha = 5.0$ and $\beta = 3.0$. For the GastricSur dataset, the optimal hyperparameters are $\alpha = 6.0$ and $\beta = 0.0$. For the TCGA-STAD dataset, the optimal hyperparameters are $\alpha = 8.0$ and $\beta = 0.0$. That is, the feature-level distillation and response-level distillation are both important for the classification task, *i.e.*, the NACT response prediction task in the GastricRes dataset, while the response-level distillation is more important for regression task, *i.e.*, the survival analysis task in the GastricSur and TCGA-STAD datasets.

References

[1] Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision, pages 4015–4025, 2021.

[2] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81:102559, 2022.

[3] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2016.

[4] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18177–18186, 2022.

[5] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, 2017.

[6] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.