

HiLa: Hierarchical Vision-Language Collaboration for Cancer Survival Prediction

Jiaqi Cui¹, Lu Wen¹, Yuchen Fei¹, Bo Liu², Luping Zhou³, Dinggang Shen^{4,5,6}, Yan Wang¹✉

¹ School of Computer Science, Sichuan University, Chengdu, China
wangyanscu@hotmail.com

² Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China

³ School of Electrical and Computer Engineering, University of Sydney, Sydney, Australia

⁴ School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai, China

⁵ Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China

⁶ Shanghai Clinical Research and Trial Center, Shanghai, China

Abstract. Survival prediction using whole-slide images (WSIs) is crucial in cancer research. Despite notable success, existing approaches are limited by their reliance on sparse slide-level labels, which hinders the learning of discriminative representations from gigapixel WSIs. Recently, vision language (VL) models, which incorporate additional language supervision, have emerged as a promising solution. However, VL-based survival prediction remains largely unexplored due to two key challenges. First, current methods often rely on only one simple language prompt and basic cosine similarity, which fails to learn fine-grained associations between multi-faceted linguistic information and visual features within WSI, resulting in inadequate vision-language alignment. Second, these methods primarily exploit patch-level information, overlooking the intrinsic hierarchy of WSIs and their interactions, causing ineffective modeling of hierarchical interactions. To tackle these problems, we propose a novel **H**ierarchical vision-**L**anguage collaboration (HiLa) framework for improved survival prediction. Specifically, HiLa employs pretrained feature extractors to generate hierarchical visual features from WSIs at both patch and region levels. At each level, a series of language prompts describing various survival-related attributes are constructed and aligned with visual features via Optimal Prompt Learning (OPL). This approach enables the comprehensive learning of discriminative visual features corresponding to different survival-related attributes from prompts, thereby improving vision-language alignment. Furthermore, we introduce two modules, i.e., Cross-Level Propagation (CLP) and Mutual Contrastive Learning (MCL) to maximize hierarchical cooperation by promoting interactions and consistency between patch and region levels. Experiments on The Cancer Genome Atlas (TCGA) datasets demonstrate our state-of-the-art performance.

Keywords: Survival Prediction, Vision-Language Models, Multimodal Learning, Computational Pathology.

1 Introduction

Survival prediction is a fundamental topic in cancer research which aims to assess the relative risks of death in cancer patients [1]. Accurate survival prediction holds significant importance in supporting clinical decision-making and maximizing therapeutic benefits [2-4]. Digital whole slide images (WSIs), which capture microscopic changes in tumor cells and their microenvironment, have become indispensable in computational pathology (CPath) for survival prediction [5].

Given the ultra-high resolution and expensive pixel-level annotations of WSIs, multiple instance learning (MIL) paradigm has become the predominant approach for WSI analysis. Traditional MIL-based methods [6-14], which rely solely on slide-level annotation for supervision, typically follow a three-step process: 1) dividing a WSI (slide) into multiple small patches (instances); 2) extracting visual features from these instances using a pretrained encoder; and 3) aggregating the instances features for slide-level prediction. Despite achieving “clinical-grade” performance, these methods are constrained by their exclusive dependence on sparse slide-level labels, making it challenging to learn sufficient survival-related discriminative visual features from gigapixel WSIs. Recently, vision-language (VL) models have shown impressive performance across various tasks [15-20] by incorporating additional language supervision, providing a promising solution to the above-mentioned limitation of the MIL paradigm in CPath. For example, Zhang *et al.* [21] utilized disease-level biomedical description to enhance the interpretation of WSIs. Qu *et al.* [22] introduced pathology language prompts at for improved WSI classification. More recently, Liu *et al.* [23] presented the first vision-language framework for survival prediction, which leverages prognostic language priors to assign weights to different instances during aggregation.

Despite notable progress, designing effective VL-based methods for survival prediction in CPath remains an open topic. The first challenge lies in the *insufficient vision-language alignment*. In clinic, pathologists rely on multiple cues (also regarded as language prompts) to assess patient survival from various perspectives (tumor differentiation, invasion, and lymph node metastasis, *etc.*). However, existing methods either condense this rich information to one oversimplified prompt (e.g., an H&E image of {cancer type}) [21, 24] or match all prompts with visual features using basic cosine similarity [23, 25]. This rudimentary alignment paradigm fails to capture fine-grained associations between multi-faceted pathological linguistic information and complex visual features within WSIs. The second challenge stems from the *ineffective modeling of hierarchical interactions*. WSIs, at a given magnification, exhibit a hierarchy across varying image resolutions, i.e., lower resolutions (e.g., 256×256 patch level) capture local cellular morphology, whereas higher resolutions (e.g., 4096×4096 region level) reveal global tissue organizations and tumor-immune interactions. However, current VL-based survival prediction methods primarily focus on patch-level analysis. This narrow focus has risk of overlooking critical global contexts, such as broader spatial organization of phenotypes in microenvironment, limiting prediction accuracy.

Motivated to address the above limitations, in this paper, we propose a novel **Hierarchical vision-Language collaboration (HiLa)** framework for cancer survival prediction. Specifically, our network employs pretrained feature extractors [26] to produce

hierarchical visual features from WSIs at both local patch and global region levels. At each level, rather than using conventional oversimplified prompts, we utilize a large language model (LLM) to generate multiple tailored language prompts enriched with diverse survival-related attributes. Then, inspired by [28, 29], we move beyond basic cosine similarity by designing Optimal Prompt Learning (OPL), which enhances optimal transport through a visual token selection mechanism to capture fine-grained, discriminative patterns in complex WSIs for improved survival assessment. On the other hand, to enhance interactions between patch level and region level, we design a Cross-Level Propagation (CLP) module that leverages patch-level knowledge to support region-level prediction, establishing a hierarchical and attentive connection between the two levels. Furthermore, we introduce a Mutual Contrastive Learning (MCL) module to enforce consistency between patch-level and region-level visual features for each patient, further boosting hierarchical interactions. Through the above hierarchical vision-language collaboration, our HiLa fully exploits the model’s representation capability, leading to improved survival prediction accuracy.

The contributions of this paper are summarized as follows. (1) We propose a novel hierarchical vision-language collaboration framework for survival prediction in CPath. (2) We introduce multiple language prompts and employ OPL to enhance the alignment between diverse survival-related attributes in the prompt set and WSI visual features, thereby improving the extraction of discriminative visual patterns. (3) We develop two modules, namely CLP and MCL, to ensure effective hierarchical interaction and consistency. (4) Experimental results on public cancer datasets from The Cancer Genome Atlas (TCGA) verify our superiority.

2 Methodology

2.1 Overview and Problem Formulation

The overall framework of the proposed method is illustrated in Fig. 1. Our network extracts visual features (tokens) from a WSI at both patch and region levels, capturing local cellular morphology and global tissue organization, respectively. Meanwhile, survival-related sentences are generated and encoded as a set of language prompts at both levels to guide the selection and aggregation of discriminative visual tokens via hierarchical vision-language collaboration. Details are provided in the following subsections.

Hierarchical Bags Construction. To capture the intrinsic hierarchy of WSIs, we employ pretrained feature extractors [26], which consist of a patch-level extractor $E_p(\cdot)$ and a region-level extractor $E_r(\cdot)$. Given the set of WSIs at 20× magnification, $E_p(\cdot)$ and $E_r(\cdot)$ generate hierarchical bags $H_p = \{H_p^i\}_{i=1}^{M_p} \in \mathbb{R}^{M_p \times d}$ for patches of 256×256 and $H_r = \{H_r^i\}_{i=1}^{M_r} \in \mathbb{R}^{M_r \times d}$ for regions of 4096×4096, respectively, where H_p^i and H_r^i denote visual tokens. M_p and M_r represent the total numbers of extracted patch-level and region-level visual tokens, respectively; d is the corresponding channel dimension.

Language Prompts Generation. We generate a set of language prompts that describe various perspectives on survival outcomes for comprehensive learning. To this end, we first construct a query for each cancer type: “What visual features of H&E-stained

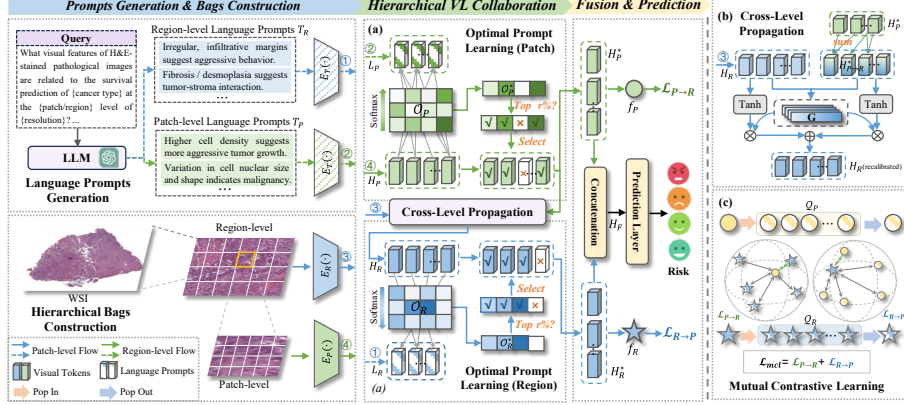


Fig. 1. Overview of the proposed HiLa. (a) The OPL module aligns language prompts with visual to-kens extracted from WSIs. (b) The CLP and (c) the MCL modules facilitate effective interactions between patch-level and region-level learning, enhancing hierarchical cooperation. Finally, visual tokens learned from both levels are concatenated to estimate risk scores.

pathological images are related to the survival prediction of {cancer type} at the {patch/region} level of {resolution}? Please list them point by point.” The query is input into LLM (i.e., GPT-4o) to generate multiple sentences $T_P = \{T_P^i\}_{i=1}^{N_P}$ and $T_R = \{T_R^i\}_{i=1}^{N_R}$ that detail different survival-related attributes observable in WSIs at the patch and region levels, respectively, where N_P and N_R are the number of sentences. The generated sentences T_P and T_R are then processed by the text encoder $E_T(\cdot)$ of PLIP, which is trained on substantial pathology images paired with expert-written texts, to produce the encoded language prompts, denoted as $L_P = \{L_P^i\}_{i=1}^{N_P \times d}$ and $L_R = \{L_R^i\}_{i=1}^{N_R \times d}$.

Survival Prediction Formulation. Survival prediction estimates the risk probability of an event before a specific time. Let $\mathcal{D} = \{P^n, c^n, t^n\}_{n=1}^N$ represent the data of N patients, where P^n is WSIs for the n -th patient, $c^n \in \{0, 1\}$ is the right censor status, and t^n is the overall survival time. Given P^n , we can acquire patch-level token bag H_P^n and region-level token bag H_R^n , and develop their associated sets of language prompts L_P^n and L_R^n . Then, discriminative visual tokens are extracted via hierarchical vision-language collaboration, yielding the final token set, denoted as H_F^n . The aim of our network is to use H_F^n to estimate the hazard function $\mathcal{F}_{hazard}^n(t|H_F^n)$, which is the probability of death occurring after a time point t . Instead of directly predicting the overall survival time t^n , survival function $\mathcal{F}_{surv}^n(t|H_F^n)$ estimates the risk by cumulating the negated hazard functions, defined as:

$$\mathcal{F}_{surv}^n(t|H_F^n) = \prod_{s=1}^t (1 - \mathcal{F}_{hazard}^n(s|H_F^n)), \quad (1)$$

2.2 Hierarchical Vision-Language Collaboration

In hierarchical vision-language collaboration, we adopt optimal prompt learning (OPL) to align diverse survival-related attributes in language prompts $L_{[P,G]}$ with visual tokens $H_{[P,G]}$ at both the patch and region levels. Moreover, a Cross-Level Propagation (CLP) module bridges the two levels, enabling patch-level knowledge to support region-level

predictions. During this process, a Mutual Contrastive Learning (MCL) module is incorporated to ensure the consistency between predictions at the two levels.

Optimal Prompt Learning. The OPL is designed to enhance the alignment between multiple language prompts and visual tokens, thereby enhancing the comprehensive learning of discriminative visual tokens corresponding to different survival-related attributes within gigapixel WSIs. As shown in Fig. 1(a), given the visual tokens $H_{[P,R]}$ and language prompts $L_{[P,R]}$ at patch and region levels (where the superscript n will be omitted when unnecessary for notational brevity), we first compute the optimal matching plan $\mathcal{O}_{[P,R]}$ with the optimal matching cost d_{OT} to align $L_{[P,R]}$ to $H_{[P,R]}$, defined as:

$$\begin{aligned} d_{OT}(u, v | \mathcal{C}_{[P,R]}) &= \min_{\mathcal{O}_{[P,R]}} \langle \mathcal{O}_{[P,R]}, \mathcal{C}_{[P,R]} \rangle, \\ \text{s.t. } \mathcal{O}_{[P,R]} 1_{M_{[P,R]}} &= u, \quad \mathcal{O}_{[P,R]} 1_{N_{[P,R]}} = v, \quad \mathcal{O}_{[P,R]} \in \mathbb{R}_+^{M_{[P,R]} \times N_{[P,R]}}, \end{aligned} \quad (2)$$

where $\mathcal{C}_{[P,R]} = 1 - \text{sim}(H_{[P,R]}, L_{[P,R]})$ is the cost matrix that quantifies the similarity between each token in $H_{[P,R]}$ and each prompt in $L_{[P,R]}$; u and v are marginal distributions. The obtained $\mathcal{O}_{[P,R]}$ assesses correspondence degree between language prompts and visual tokens, ensuring that each language prompt is optimally matched to the most relevant visual tokens. Then, $\mathcal{O}_{[P,R]}$ is converted to optimal matching probability $\mathcal{O}'_{[P,R]}$:

$$\mathcal{O}'_{[P,R]}[m, n] = \frac{\exp(1 - \mathcal{O}_{[P,R]}[m, n])}{\sum_{m'=1}^M \exp(1 - \mathcal{O}_{[P,R]}[m', n])} \in \mathbb{R}^{M_{[P,R]} \times N_{[P,R]}}, \quad (3)$$

where $m \in \{1, \dots, M_{[P,R]}\}$ and $n \in \{1, \dots, N_{[P,R]}\}$. Next, we sum $\mathcal{O}'_{[P,R]}$ along the second (language) dimension, thus obtaining the optimal alignment score $\mathcal{O}^*_{[P,R]} \in \mathbb{R}^{M_{[P,R]}}$. Each element in $\mathcal{O}^*_{[P,R]}$ measures the alignment between a visual token and all survival-related attributes within language prompts. Visual tokens with higher matching probabilities are considered more related to survival assessment. Accordingly, we reorganize the visual tokens $H_{[P,R]}$ by ranking $\mathcal{O}^*_{[P,R]}$ in descending order, and selecting the top $\tau\%$ visual tokens, yielding $H^*_{[P,R]} \in \mathbb{R}^{(M_{[P,R]} \times \tau\%) \times d}$ that contains rich discriminative information related to patient survival.

Cross-Level Propagation (CLP) Module. We develop the CLP module which establishes an attentive and hierarchical connection between the patch and region levels, enabling patch-level knowledge to effectively guide region-level learning, as displayed in Fig. 1(b). Instead of naive concatenation or summation, CLP uses a gating mechanism to adaptively modulate the influence of patch-level tokens to region-level predictions. As each region corresponds to multiple patch-level visual tokens, the above selected H^*_P are summed according to their belonged regions, yielding the selected region-level inputs, denoted as $H^*_{P \rightarrow R}$. We concatenate $H^*_{P \rightarrow R}$ with the unselected region-level inputs H_R to generate a gate matrix G via linear projection, followed by a sigmoid activation function. Then, G is employed to adjust the weights of Tanh-activated $H^*_{P \rightarrow R}$ and H_R . The above process can be formulated as follows:

$$\begin{aligned} G &= \text{Sigmoid}(\text{Concat}(H^*_{P \rightarrow R}; H_R); \theta_G), \\ H_R &= G * \text{Tanh}(H^*_{P \rightarrow R}; \theta_P) + (1 - G) * \text{Tanh}(H_R; \theta_R), \end{aligned} \quad (4)$$

where $\{\theta_G, \theta_P, \theta_R\}$ are learnable parameters. The obtained recalibrated H_R is then used as the input to the OPL module at the region level, enabling more precise predictions.

Mutual Contrastive Learning (MCL) Module. As illustrated in Fig. 1(c), we introduce a MCL module to maintain consistency between the selected patch-level and

region-level visual tokens for each patient, further improving hierarchical cooperation. Concretely, given $H_{[P,R]}^*$, we compute prototypes $f_{[P,R]} \in \mathbb{R}^{(M_{[P,R]} \times r\%) \times 1}$ by summing across their channel dimension. During the training, we initialize a patch-level memory queue Q_P with $B - 1$ patch prototypes, and a region-level memory queue Q_R with the same number of region prototypes. In subsequent iterations, $Q_{[P,R]}$ are updated by storing the current $f_{[P,R]}$ and popping out the oldest entries. The MCL module performs contrastive learning at both the patch and region levels. At the patch level, denote the current patch prototype f_P^n (for the n -th patient) as the anchor sample, while the corresponding region prototype f_R^n (for the same patient) acts as the positive sample; and the other region prototypes (from different patients) in Q_R are regarded as negative samples. Based on this, the patch-level contrastive loss $\mathcal{L}_{P \rightarrow R}$ for N patients is defined as:

$$\mathcal{L}_{P \rightarrow R} = - \sum_{n=1}^N \log \frac{\exp(f_P^n \cdot f_R^n)}{\exp(f_P^n \cdot f_R^n) + \sum_{k=1, k \neq n}^{B-1} \exp(f_P^n \cdot f_R^k)}, \quad (5)$$

Similarly, we can derive the region-level contrastive loss $\mathcal{L}_{R \rightarrow P}$. Together, these two losses constitute the MCL loss, which is expressed as:

$$\mathcal{L}_{mcl} = \mathcal{L}_{P \rightarrow R} + \mathcal{L}_{R \rightarrow P}, \quad (6)$$

2.3 Fusion and Prediction

To yield the final survival prediction, we concatenate the selected patch-level tokens H_P^* and region-level tokens H_R^* to form the final token set H_F^n . Then, H_F^n is passed through a linear-based prediction layer to estimate risk scores. The negative log-likelihood (NLL) loss [29] is applied to supervise this process, which is defined as follows:

$$\mathcal{L}_{sur} = \sum_{n=1}^N -c^n \log(\mathcal{F}_{surv}^n(t|H_F^n)) - (1 - c^n) \log(\mathcal{F}_{surv}^n(t - 1|H_F^n)) - (1 - c^n) \log(\mathcal{F}_{hazard}^n(t|H_F^n)), \quad (7)$$

where $\mathcal{F}_{surv}^n$ and $\mathcal{F}_{hazard}^n$ are the survival and hazard functions, as described in Sec. 2.1.

Finally, the total objective function of our framework is a weighted sum of the survival prediction loss $\mathcal{L}_{sur}$ and the mutual contrastive learning loss $\mathcal{L}_{mcl}$ defined in Eq. (6), which is formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{sur} + \lambda \mathcal{L}_{mcl}. \quad (8)$$

where λ is the weighting coefficient to balance these two terms.

3 Experiments

3.1 Experimental Settings

Datasets. We train and evaluate our method on public datasets from The Cancer Genome Atlas (TCGA), i.e., Breast Invasive Carcinoma (BRCA, $N=956$), Lung Adenocarcinoma (LUAD, $N=453$), and Uterine Corpus Endometrial Carcinoma (UCEC, $N=480$). Each dataset includes diagnostic WSIs and ground-truth survival labels.

Evaluation Metrics. We adopt the concordance index (CI), a standard metric for survival prediction that quantifies a model's ability to correctly rank predicted risk scores with respect to ground-truth survival labels. In addition, Kaplan-Meier analysis with the Log-rank test is employed to assess the stratification ability between high-risk and

low-risk patient groups. To mitigate the impact of data splits, we perform 5-fold cross-validation, and report the mean performance over five folds.

Implementation Details. Our network is implemented by Pytorch framework and trained end-to-end on a GeForce GTX 3090 GPUs for 20 epochs with a batch size of 1, using Adam optimizer. The learning rate is set to $2e-4$. Prompt features are shared across all WSIs within a dataset. The prompt dimension d is set to 512. The ratio $r\%$ in OPL is set to 0.6 and the length of memory queue B in MCL is set to 20. The hyper-parameter λ in Eq. (8) is set as 0.01.

3.2 Comparisons with State-of-the-Art Approaches

We compare our method against two groups of state-of-the-art (SOTA) approaches: (1) **Vision-only (V-only) methods**, including ABMIL [8], CLAM [10], TransMIL [9], DSMIL [30], DTFT-MIL [31], and RRT-MIL [7]; and (2) **Vision-Language (VL) methods**, involving CoOp [32], and VLSA [21]. Specifically, since CoOp is not originally designed for survival prediction, we adapt it by calculating the cosine similarity between visual tokens and language prompts. The top $r\%$ visual tokens, based on cosine similarity values, are selected for predicting risk scores via a prediction layer.

C-index Comparison. As shown in Table 1, our HiLa method achieves the state-of-the-art performance with an average CI of 67.1%. Specifically, it surpasses V-only methods by 4.3%-8.2% in overall performance, demonstrating the benefits of integrating language information to enhance traditional MIL pipelines for survival prediction. Moreover, HiLa also delivers superior performance compared to other VL-based methods. Particularly, HiLa leads the second-best VLSA model by 3.7% on BRCA, 2.1% on LUAD, and 3.4% on UCEC, respectively, confirming its robust capability to exploit survival-related information via vision-language collaboration. Code and more results on additional datasets and analysis will be given in <https://github.com/gluucose/HiLa>.

Kaplan-Meier Analysis. To further validate the effectiveness of HiLa, we conduct the Kaplan-Meier (KM) analysis. As illustrated in Fig. 2, all patients are stratified into low-risk group (blue) and high-risk group (yellow) based on the median value of predicted risk scores. A Log-rank test is performed to assess the statistical significance of differences between the two risk groups via p-values. Compared to the second-best VLSA,

Table 1. C-index results for different methods on three TCGA datasets. “V” represents the utilization of WSIs and “L” signifies the use of language prompt(s).

Method	Modality		Datasets			Overall
	V	L	BRCA	LUAD	UCEC	
ABMIL [6]	✓		0.613±0.033	0.604±0.043	0.618±0.014	0.612
CLAM-SB [8]	✓		0.605±0.062	0.586±0.033	0.576±0.043	0.589
CLAM-MB [8]	✓		0.611±0.041	0.612±0.022	0.589±0.023	0.604
TransMIL [7]	✓		0.628±0.015	0.626±0.029	0.601±0.030	0.618
DSMIL [28]	✓		0.617±0.029	0.614±0.035	0.647±0.022	0.626
DTFT-MIL [29]	✓		0.612±0.032	0.611±0.038	0.641±0.032	0.621
RRT-MIL [5]	✓		0.641±0.006	0.624±0.029	0.618±0.030	0.628
CoOp [30]	✓	✓	0.608±0.057	0.626±0.055	0.664±0.023	0.633
VLSA [21]	✓	✓	0.622±0.021	0.622±0.050	0.678±0.043	0.641
HiLa (Ours)	✓	✓	0.659±0.044	0.643±0.055	0.712±0.047	0.671

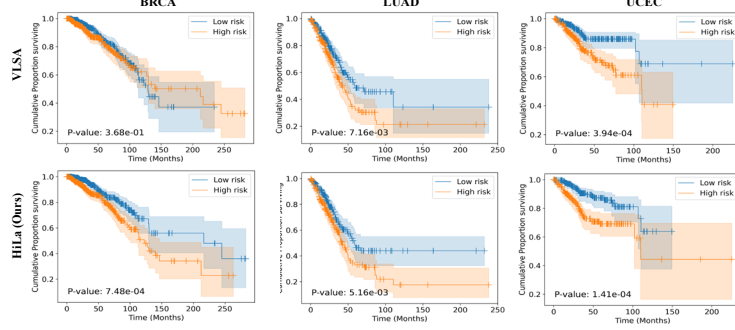


Fig. 2. Kaplan-Meier (KM) analysis of the best comparative method and our proposed framework on three datasets. The p-values for Log-rank test are provided in each sub-figure.

Table 2. C-index results of the ablation studies on three TCGA datasets.

ID	Method		Dataset			Overall
	Modules & Descriptions		BRCA	LUAD	UCEC	
(A)	PVLC	V-only Baseline	0.613±0.033	0.604±0.043	0.618±0.014	0.612
(B)		(A) + SLP (Cos)	0.615±0.046	0.611±0.035	0.637±0.039	0.620
(C)		(A) + MLP (Cos)	0.616±0.060	0.614±0.025	0.645±0.034	0.625
(D)		(A) + MLP (OPL)	0.623±0.017	0.617±0.057	0.679±0.046	0.640
(E)	RVLC	(D) + Region Tokens	0.644±0.028	0.628±0.058	0.693±0.042	0.655
(F)	CVLC	(E) + CPM	0.654±0.042	0.633±0.047	0.702±0.043	0.663
(G)		(F) + MCL (Ours)	0.659±0.044	0.643±0.055	0.712±0.047	0.671

our approach consistently yields p-values below 0.05, suggesting a statistically significant discrimination between the high-risk and low-risk patient groups.

3.3 Ablation Studies

To verify the effectiveness of the key components in our HiLa framework, we conduct ablation studies with the following variants. First, ABMIL aggregates patch-level visual tokens, establishing the V-only baseline Model-A. Then, we introduce a single language prompt (SLP) and apply the cosine similarity (Cos) metric to select the top $r\%$ visual tokens, creating Model-B as the baseline for patch-level vision-language collaboration (PVLC). Next, we replace SLP with multiple language prompts (MLP) to form Model-C. Model-D is derived by substituting Cos with the proposed OPL. Building on Model-D, we involve region-level VLC (RVLC) by introducing the selected region-level visual tokens to produce Model-E. Advancing from Model-E, we inject cross-level VLC (CVLC) by progressively incorporating CPM and MCL, yielding Model-F and Model-G (Ours). The results are provided in Table 2. Particularly, the increasing CI values observed in Models A to D highlight the effectiveness of combining multiple language supervisions and leveraging OPL to exploit discriminative visual tokens associated with diverse survival-associated attributes comprehensively. Moreover, the model achieves superior performance by introducing RVLC (Model-E), indicating the potential of region-level contexts in providing crucial complementary information. In

addition, including CPM and MCL (into Models E and F) leads to substantial performance gains, confirming their critical role in fostering hierarchical cooperation.

4 Conclusion

In this paper, we have presented a hierarchical vision-language collaboration framework for survival prediction in CPath. To enhance vision-language alignment, we generate multiple survival-related language prompts using LLM. Then, we utilize OPL to encourage comprehensive exploration and identification of discriminative visual features from gigapixel WSIs. Moreover, we also design the CLP and MCL modules, to encourage effective hierarchical cooperation between visual features extracted at patch and region levels. Experimental results have shown our feasibility and superiority.

Acknowledgments. This work is supported by National Natural Science Foundation of China (NSFC 62371325, U23A20295, 62131015, 82441023, 82394432), Sichuan Science and Technology Program (2025NSFJQ0050, 2024ZDZX0018, MZGC20240058), and Key Lab of Internet Natural Language Processing of Sichuan Provincial Education Department (No. INLP202402).

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

References

1. Song, A.H., Jaume, G., Williamson, D.F., et al.: Artificial intelligence for digital and computational pathology. *Nat. Rev. Bioeng.*, 1(12), 930–949 (2023)
2. Chen, R.J., Ding, T., Lu, M.Y., et al.: Towards a general-purpose foundation model for computational pathology. *Nat. Med.*, 30(3), 850–862 (2024)
3. Cui, J., Xu, Y., Zheng, H., et al.: HMT: A Hybrid Multi-Modal Transformer With Multi-Task Learning for Survival Prediction in Head and Neck Cancer. *IEEE Trans. Radiat. Plasma Med. Sci.* (2025)
4. Cui, J., Zheng, H., Liu, Y., et al.: Ma²SP: Missing-aware Prompting with Modality-adaptive Integration for Incomplete Multi-modal Survival Prediction. *IEEE Signal Process. Lett.*, 31, 2455–2459 (2024)
5. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J. and Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. *Nat. Med.* 29(9), 2307–2316 (2023)
6. Chen, R.J., Lu, M.Y., Shaban, M., et al.: Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In: de Bruijne, M., et al. (eds.) *MICCAI 2021*, vol. 12908, pp. 339–349. Springer, Cham (2021)
7. Tang, W., Zhou, F., Huang, S., Zhu, X., Zhang, Y. and Liu, B: Feature Re-Embedding: Towards Foundation Model-Level Performance in Computational Pathology. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11343–11352 (2024)
8. Ilse, M., Tomczak, J. and Welling, M.: Attention-based deep multiple instance learning. In: *International Conference on Machine Learning*, pp. 2127–2136 (2018)

9. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J. and Ji, X.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In: *Advances in Neural Information Processing Systems*, vol. 34, pp. 2136–2147 (2021)
10. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M. and Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.*, 5(6), 555–570 (2021)
11. Liu, P., Ji, L., Ye, F. and Fu, B.: Advmil: Adversarial multiple instance learning for the survival analysis on whole-slide images. *Med. Image Anal.*, 91, 103020 (2024)
12. Jiang, S., Suriawinata, A.A. and Hassanpour, S.: MHAttnSurv: Multi-head attention for survival prediction using whole-slide pathology images. *Comput. Biol. Med.*, 158, 106883 (2023)
13. Xiang, J. and Zhang, J. Exploring low-rank property in multiple instance learning for whole slide image classification. In: *International Conference on Learning Representations* (2023)
14. Yu, J., Wang, X., Ma, T., Li, X. and Xu, Y.: Patch-Slide Discriminative Joint Learning for Weakly-Supervised Whole Slide Image Representation and Classification. In: Linguraru, M.G., et al. (eds.) *MICCAI 2024*, vol. 15003, pp. 713–722. Springer, Cham (2024)
15. Li, Z., Li, Y., Li, Q., et al.: Lvit: language meets vision transformer in medical image segmentation. *IEEE Trans. Med. Imaging*, 43(1), 96–107 (2023)
16. Cui, J., Zeng, X., Zeng, P., et al.: MCAD: Multi-modal Conditioned Adversarial Diffusion Model for High-Quality PET Image Reconstruction. In: Linguraru, M.G., et al. (eds.) *MICCAI 2024*, vol. 15009, pp. 467–477. Springer, Cham (2024)
17. Li, L., Lian, S., Luo, Z., Wang, B. and Li, S.: VCLIPSeg: Voxel-Wise CLIP-Enhanced Model for Semi-supervised Medical Image Segmentation. In: Linguraru, M.G., et al. (eds.) *MICCAI 2024*, vol. 15009, pp. 692–701. Springer, Cham (2024)
18. Zhang, X., Xu, M., Qiu, D., Yan, R., Lang, N. and Zhou, X.: Medclip: Adapting clip for few-shot medical image anomaly detection. In: Linguraru, M.G., et al. (eds.) *MICCAI 2024*, vol. 15001, pp. 458–468. Springer, Cham (2024)
19. Chikontwe, P., Kang, M., Luna, M., Nam, S. and Park, S.H.: Low-Shot Prompt Tuning for Multiple Instance Learning Based Histology Classification. In: Linguraru, M.G., et al. (eds.) *MICCAI 2024*, vol. 15004, pp. 285–295. Springer, Cham (2024)
20. Wang, P., Zhang, H. and Yuan, Y., MCPL: Multi-modal Collaborative Prompt Learning for Medical Vision-Language Model. *IEEE Trans. Med. Imaging*, 43(12), 4224–4235 (2024)
21. Zhang, Y., Gao, J., Zhou, M., et al.: Text-guided foundation model adaptation for pathological image classification. In: Greenspan, H., et al. (eds.) *MICCAI 2023*, vol. 13224, pp. 272–282. Springer, Cham (2023)
22. Qu, L., Fu, K., Wang, M. and Song, Z.: The rise of ai language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. *Advances in Neural Information Processing Systems*, 36 (2024)
23. Liu, P., Ji, L., Gou, J., Fu, B. and Ye, M.: Interpretable Vision-Language Survival Analysis with Ordinal Inductive Bias for Computational Pathology. *arXiv preprint arXiv:2409.09369* (2024)
24. Chikontwe, P., Kang, M., Luna, M., Nam, S. and Park, S.H.: Low-Shot Prompt Tuning for Multiple Instance Learning Based Histology Classification. In: Linguraru, M.G., et al. (eds.) *MICCAI 2024*, vol. 15004, pp. 285–295. Springer, Cham (2024)
25. Qu, L., Yang, D., Huang, D., et al.: Pathology-knowledge enhanced multi-instance prompt learning for few-shot whole slide image classification. In: Leonardis, A. et al. (eds.) *ECCV 2024*, vol. 15069, pp. 196–212. Springer, Cham (2024)

26. Chen, R.J., Chen, C., Li, Y., et al.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16144–16155 (2022)
27. Song, A.H., Chen, R.J., Jaume, G., Vaidya, A.J., Baras, A. and Mahmood, F., Multimodal Prototyping for cancer survival prediction. In: Forty-first International Conference on Machine Learning (2024)
28. Xu, Y. and Chen, H. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21241–21251 (2023)
29. Zadeh, S.G. and Schmid, M., Bias in cross-entropy-based training of deep survival networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(9), 3126–3137 (2020)
30. Li, B., Li, Y. and Eliceiri, K.W.: Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14318–14328 (2021)
31. Zhang, H., Meng, Y., Zhao, Y., et al.: Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18802–18812 (2022)
32. Zhou, K., Yang, J., Loy, C.C. and Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16816–16825 (2022)